European Journal of Computer Science and Information Technology

Vol.2, No.4, pp.8-20, December 2014

Published by European Centre for Research Training and Development UK (www.eajournals.org)

DNA COMPUTER CODE BASED ON EXPANDED GENETIC ALPHABET

Alfonso Jiménez-Sánchez
Dpt. Genetics, Universidad de Extremadura, E06080-Badajoz, Spain

ABSTRACT: *Due to the capacity of DNA to store large amounts of information in a very small physical volume, it has become an attractive molecule for storing information in future molecular computers. This information resides in the order of nucleotides (nt), and several proposals to correlate the 256 ASCII computer symbols with the same number of different groupings of nt have been described. Although a DNA molecule of any size can be synthesised, its use, however, has several limitations, the most important of which are related to stability and biosecurity. To circumvent these limitations, to increase the capacity to store information, and then diminish the probability of errors, I have considered the use of a DNA molecule made with the two standard nt-pairs and two non-standard (ns) synthetic nt-pairs. The use of this ns-DNA would generate 512 permutations for a triplet code and would, therefore, permit the encoding of the 256 computer symbols, together with a start and an end signal, coded by three nt. Printable symbols can be distributed into four groups (upper case, lower case letters, numbers and mathematical symbols). A common first nt was assigned to all triplets from the same group. The excess of 256 triplets was used to add a high redundancy in the third nt according with the frequency of use in English writings. The main advantages of this DNA encoding relative to the previously published are a lower size, a lower error probability, inability to contaminate any living cell, and an explicit non-biological origin.*

KEYWORDS: DNA computing; expanded genetic alphabet; molecular computing

INTRODUCTION

All living organisms store all of the information required to build their different cells and the entire organism in one or more DNA molecules. This high capacity for storing large amounts of information in a very small physical volumes has made this molecule very attractive for storing information in future molecular computers. DNA molecules house their information encoded in the order of its four nt, A, T, G and C. This encoded information has to be copied before a cell can divide to maintain a copy of the full information in each new-born cell. The coded information has to be translated into the building blocks to construct all of the cell and organism structures and into the functional blocks to assemble the operating machineries that will sustain all their functionalities. These two functionalities require DNA to have two main features. First, the complementarities of the two nt-pairs AT and GC, which together with an efficient DNA polymerase preserves the sequence and hence the information in the two copies of DNA obtained by the replication process. Second, the order of nt has to be read when its information is transferred into the structural and functional proteins, the building blocks of all living cells. In this transference, the information encoded in the DNA is translated into the sequence of amino acids (aa) that compose the proteins by using a coding system in which three nt, or a triplet, code for each aa. In this translation, the four nt yield 64

triplets that code for 20 aa and three stop signals. As all of the triplets code for an aa, most of the aa are coded for by two to six different triplets, producing a redundant genetic code. This redundancy has been biased by evolution to increase the stability of the genetic information by encoding the most frequent aa with a higher number of triplets that consequently lead to a lower mutation frequency.Any synthetic DNA strand designed for storing information will require the synthesis of its complementary strand by a DNA polymerase to produce the more stable ds-DNA form and to be copied a number of times for its safe preservation. Furthermore, these molecules will also eventually be read by a DNA polymerase to determine its nt sequence for decoding their information content. All these readings require DNA to be composed of pairs of complementary nt, with AT and GC being the two nt-pairs from biological origin (atdbio 2014).

The efforts aimed at the development of a semi-synthetic biology have succeeded with the expansion of the genetic alphabet (Malyshev et al. 2014) by including non-standard nucleotides (ns-nt) with interbase pairing based on ns-hydrogen-bonds (Yang et al. 2007, 2011; Sismour et al. 2004), steric complementarity (Hirao et al. 2006a, b), and hydrophobic interactions (Li et al. 2013; Malyshev et al. 2012, 2014) that are capable of *in vitro* replication (Hirao et al. 2006a, b; Li et al. 2013; Malyshev et al. 2012; Yamashige et al. 2011; Yang et al. 2011). Although they were not created for this purpose, these ns-nt might have an important role in determining the use of synthetic DNA molecules as the memory of future molecular computers.

The design of a DNA code for storing information in its nt sequence has been successful to date using the four natural nt, A, T, G, and C (Church et al. 2012; Goldman et al. 2013; Jiménez-Sánchez 2013). The use of these DNA molecules in molecular computers involves some advantages, as the use of all molecular biology technologies, and some disadvantages, such as the possibility of contamination of any biological system. Any DNA molecule containing ns-nt has the great advantage of being incapable of possible contamination of any living cell, and furthermore, it can be used with most of the *in vitro* DNA molecular techniques.

I herein suggest the use of a DNA molecule made with the two natural nt-pairs and the addition of two complementary ns-nt-pairs. This polynucleotide sequence read in groups of three will create up to 512 triplets, twice the number of the required computer symbols. This excess of triplets will be used to add redundancy to the most frequent symbols to lower error frequency. I suggest this redundancy should be concentrated at the third nt of each triplet. As the ns-nt are not found in any living cell, these synthetic DNA molecules will be incapable of contaminating any living organism and, consequently, will be exempt from the NIH Guidelines (Sebelius 2010), which will permit the synthesis of DNA of any length. Consequently, this encoding system will contain the highest storage capacity, shortest DNA molecule length, lowest error frequency, biological safety, and an evident non-biological origin.

RESULTS AND DISCUSSION

In a previous work, I proposed a DNA code by using the four natural nt to code for the required 256 computer symbols. As $4^4 = 256$, the use of four nt per symbol produces all of the required codes (Jiménez-Sánchez 2013). That code would create an error-proof memory

system with a length that was shorter than previous proposals (Church et al. 2012; Goldman et al. 2013). Herein I propose a more advantageous encoding system based on the expanded genetic alphabet.

According to life's way of coding, e.g., the genetic code, two properties should be satisfied for making a code: 1) all symbols have to be encoded by the same number of nt per code, and 2) the most frequent symbols should have a number of codes higher than the less frequent ones. To fulfill these premises, I could use an expanded ns-DNA made with one or more extra ns-nt-pairs. To obtain the encoding for more than 256 symbols, one can use 6 nt that when read in groups of four will produce 1296 codes or 8 nt to obtain an encoding containing 512 triplets. A further diminishing of this encoding to two nt per code would require a DNA made with the two natural nt-pairs plus 6 ns-nt-pairs.

Hence, herein I propose the use of a synthetic DNA molecule for memory storage made with 4 nt-pairs, the two natural AT and GC pairs and two ns-nt-pairs, which I will arbitrary name XY and VW. If every ASCII symbol were coded by a triplet, they will be an excess of 256 triplets that could be used to add redundancy to the third nt of the 96 printable symbols according to their frequency in English writings (Richardson et al. 2004).

To produce an enhanced error-proof encoding, I made the first nt of triplets from each of the four symbol groups to be fixed with the following assignments: A for all uppercase letters (ASCII codes 65 to 90), T for lowercase letters (ASCII 97-122), X for numbers (ASCII 21-30), and Y for other printable symbols (ASCII 32-47, 58-64, 91-96, and 123-127) (Table 1; Table 2). The remaining codes are distributed consecutively using G, C, and V as the first nt in every triplet. The second and third nt for every triplet were obtained by a three steps process: First, I distributed the frequency of their use in English writings, spanning from 0.07 to 12.02, into four discrete groups and 1 to 4 triplets were assigned to each corresponding group. Second, a base 8 number, from 00 to 77, was assigned to every triplet, always assuring for every symbol to have the same digit at the first position. Third, nt V, G, X, T, W, C, Y, and A were assigned to digits 0 to 7 correspondingly and each digit obtained in the second step assigned to its corresponding nt.

Finally, I assigned the unused W nt for the first nt of triplets WGX and WYA to specify, respectively, the start and end of every document. These signals could be used in any number of tandem repeats at the beginning and end of any document to ensure the correct reading frame (Table 2).

I compared the predictable error frequency of this encoding with that obtained by Church et al. (2012). Using a code of eight nt per Byte these authors found 10 errors out of 658,776 Bytes that renders a mutation frequency per nt of 1.8975×10^{-6}. The frequency of errors per Byte should be the addition of the probability of change in the meaning of the Byte due to a mutation in the first, second or third nt. With the herein proposed encoding, probability of the first nt to cause any error is null as they will be easily detected. Probability of the second nt does not change the above obtained mutation frequency per nt. The probability of an error in the third nt is affected by the frequency of appearance each letter in a text and by the redundancy that gives an average of 1.2697×10^{-6} (Table 3). Therefore, the frequency of errors per Byte should be 3.172×10^{-6} that will produce 2.086 mistakes in a text containing 658,776 letters which means about five times higher fidelity than in the above mentioned work.

CONCLUSIONS

Three crucial pillars for this expanded alphabet make it the best choice for future DNA memory: *i*) a number of nt per Byte that is the shortest among published DNA encodings, which permits the smallest memory size, *ii*) a fixed nt in the first nt of triplets from each symbol group, and a redundant third nt of triplets from all printable symbols, that make a high fidelity encoding, and *iii*) inability to contaminate any living cell. The lack of possible genetic contamination will make NIH Guidelines inapplicable to these DNA molecules (Sebelius 2010). This will permit the use of synthetic molecules of any length. Furthermore, the presence of ns-nt makes it obvious that these DNA molecules do not have a biological origin and implies designed encoded information.

In addition to the mentioned properties, including the short DNA length, low error probability, inability to lead to biocontamination, and the explicit non-biological designed encoded information, this encoding will enable the storage of $3.2486x10^{17}$ Bytes (approximately 324 PetaBytes) in 1 mg of a ds-DNA (more stable than ss-DNA molecules) that, at present, is by far the smallest component to hold this amount of information.

I believe that this encoding system could be widely accepted due to its high information capacity, low size, ability to synthesise of any DNA length, error resistance, biological safety, and clear synthetic origin. The advantageous properties make this encoding a forerunner in the attempt to create DNA-based computers.

REFERENCES

Atdbio(2014)*Transcription, translation and replication.* http://www.atdbio.com/content/14/Transcription-Translation-and-Replication

Church, G. M., Gao, Y. and Kosuri, S. (2012) *Next-generation digital information storage in DNA*. Science 337 1628.

Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B. and Birney, E. (2013) *Towards practical, high-capacity, low-maintenance information storage in synthesized DNA*. Nature doi: 101038/nature11875

Hirao, I., Kimoto, M., Mitsui, T., Fujiwara, T., Kawai, R., Sato, A., Harada, Y. and Yokoyama, S. (2006a) *An unnatural hydrophobic base pair system: Site-specific incorporation of nucleotide analogs into DNA and RNA*. Nature Methods 3 729-735.

Hirao, I., Kimoto, M., Mitsui, T., Fujiwara, T., Kawai, R., Sato, A., Harada, Y. and Yokoyama, S. (2006b) *An unnatural base pair system for in vitro replication and transcription*. Nucleic Acid Symposium Series 50 33-34. doi: 10.1093/nass/nrl017

Jiménez-Sánchez, A. (2013) *A proposal for a DNA-based computer code*. International Inventions Journal Biochemistry and Bioinformatics 1 1-4.

Li, L., Degardin, M., Lavergne, T., Malyshev, D. A., Dhami, K., Ordoukhanian, P. and Romesberg, F. E. (2014) *Natural-like Replication of an Unnatural Base Pair for the Expansion of the Genetic Alphabet and Biotechnology Applications*. Journal of American Chemical Society 136 826-829.

Malyshev, D. A., Dhami, K., Quacha, H. T., Lavergne, T., Ordoukhanianb, P., Torkamanic, A. and Romesberg, F. E. (2012) *Efficient and sequence-independent replication of*

DNA containing a third base pair establishes a functional six-letter genetic alphabet. Proceedings National Academy Science USA 109 12005-12010.

Malyshev, D. A., Dhami, K., Lavergne, T., Chen, T., Dai, N., Foster, J. M., Corrêa, I. R. and Romesberg, F. E. (2014) *A semi-synthetic organism with an expanded genetic alphabet*. Nature. doi: 10.1038/nature13314.

Richardson, M., Gabrosek, J., Reischman, D. and Curtiss, P. (2004) *Morse Code, Scrabble, and the Alphabet*. Journal of Statistics Education 12(3). http://www.amstat.org/publications/jse/v12n3/richardson.html

Sebelius, K. (2010) *Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA*. Federal Register 75(197):62820-62832. FR Doc No: 2010-25728.

Sismour, A. M., Lutz, S., Park, J. H., Lutz, M. J., Boyer, P. L., Hughes, S. H. and Benner, S. A. (2004) *PCR amplification of DNA containing non-standard base pairs by variants of reverse transcriptase from Human Immunodeficiency Virus-1*. Nucleic Acid Research 32 728-735.

Yamashige, R. and Kimoto, M. (2011) *Highly specific unnatural base pair systems as a third base pair for PCR amplification.* Nucleic Acid Research 40 2793−2806.

Yang, Z., Chen, F., Alvarado, J. B. and Benner, S. A. (2011) *Amplification, Mutation, and Sequencing of a Six-Letter Synthetic Genetic System*. Journal American Chemical Society 133 15105-15112.

Yang, Z., Sismour, A. M., Sheng, P., Puskar, N. L. and Benner, S. A. (2007) *Enzymatic incorporation of a third nucleobase pair*. Nucleic Acid Research 35 4238-4249

Table 1. Codes for upper and lower case letters (in italics) and numbers[1].

<table>
<tr><th></th><th colspan="8">second nt</th><th></th></tr>
<tr><th>1st nt</th><th>V</th><th>G</th><th>X</th><th>T</th><th>W</th><th>C</th><th>Y</th><th>A</th><th>3rd nt</th></tr>
<tr><td rowspan="8">A
(uppercase)
T
(lowercase)</td><td rowspan="3">A</td><td rowspan="3">D</td><td rowspan="2">F</td><td>J</td><td rowspan="3">N</td><td>P</td><td rowspan="3">T</td><td rowspan="2">W</td><td>V</td></tr>
<tr><td>K</td><td>Q</td><td>G</td></tr>
<tr><td>G</td><td rowspan="3">L</td><td rowspan="3">R</td><td>X</td></tr>
<tr><td rowspan="2">H</td><td>X</td><td>T</td></tr>
<tr><td>B</td><td rowspan="3">E</td><td rowspan="3">O</td><td rowspan="3">U</td><td rowspan="2">Y</td><td>W</td></tr>
<tr><td rowspan="2">C</td><td rowspan="2">I</td><td rowspan="2">M</td><td rowspan="2">S</td><td>C</td></tr>
<tr><td rowspan="2">Z</td><td>Y</td></tr>
<tr><td>V</td><td>A</td></tr>
<tr><td rowspan="2">X</td><td>0</td><td>2</td><td>4</td><td>6</td><td>8</td><td colspan="3" rowspan="2"></td><td>VGXT</td></tr>
<tr><td>1</td><td>3</td><td>5</td><td>7</td><td>9</td><td>WCYA</td></tr>
</table>

[1] A triplet code for a letter or number can be obtained by selecting the first nt from the left side of the table (A, T or X whether it belongs to an uppercase letter, lowercase letter or a number, respectively); the second nt is obtained from the heading of its column; and the third nt is obtained by selecting any of the elements from the right side of the table within the limits of its box. Full encoding of the 256 ASCII symbols is provided in Online Resource 1.

Table 2. Code for the 256 ASCII symbols in triplets made from a DNA synthesised with the four standard biological nt, A, T, G, and C, and four ns-synthetic-nt, X, Y, V, and W.

	ASCII	letter freq.	nº codes	numbers in base 8[1]				triplets 1st				
NULL	0		1	00	-	-	-	G	GVV	-	-	-
	1		1	01	-	-	-	G	GVG	-	-	-
	2		1	02	-	-	-	G	GVX	-	-	-
	3		1	03	-	-	-	G	GVT	-	-	-
	4		1	04	-	-	-	G	GVW	-	-	-
	5		1	05	-	-	-	G	GVC	-	-	-
	6		1	06	-	-	-	G	GVY	-	-	-
	7		1	07	-	-	-	G	GVA	-	-	-
	8		1	10	-	-	-	G	GGV	-	-	-
	9		1	11	-	-	-	G	GGG	-	-	-
	10		1	12	-	-	-	G	GGX	-	-	-
	11		1	13	-	-	-	G	GGT	-	-	-
	12		1	14	-	-	-	G	GGW	-	-	-
	13		1	15	-	-	-	G	GGC	-	-	-
	14		1	16	-	-	-	G	GGY	-	-	-
	15		1	17	-	-	-	G	GGA	-	-	-
	16		1	20	-	-	-	G	GXV	-	-	-
	17		1	21	-	-	-	G	GXG	-	-	-
	18		1	22	-	-	-	G	GXX	-	-	-
	19		1	23	-	-	-	G	GXT	-	-	-
	20		1	24	-	-	-	G	GXW	-	-	-
	21		1	25	-	-	-	G	GXC	-	-	-
	22		1	26	-	-	-	G	GXY	-	-	-
	23		1	27	-	-	-	G	GXA	-	-	-
	24		1	30	-	-	-	G	GTV	-	-	-
	25		1	31	-	-	-	G	GTG	-	-	-
	26		1	32	-	-	-	G	GTX	-	-	-
	27		1	33	-	-	-	G	GTT	-	-	-
	28		1	34	-	-	-	G	GTW	-	-	-
	29		1	35	-	-	-	G	GTC	-	-	-
	30		1	36	-	-	-	G	GTY	-	-	-
	31		1	37	-	-	-	G	GTA	-	-	-
space	32		4	00	01	02	03	Y	YVV	YVG	YVX	YVT
!	33		1	04	-	-	-	Y	YVW	-	-	-
"	34		1	05	-	-	-	Y	YVC	-	-	-
#	35		1	06	-	-	-	Y	YVY	-	-	-
$	36		2	07	10	-	-	Y	YVA	YGV	-	-
%	37		2	11	12	-	-	Y	YGG	YGX	-	-

&	38		1	13	-	-	-	Y	YGT	-	-	-
'	39		2	14	15	-	-	Y	YGW	YGC	-	-
(	40		2	16	17	-	-	Y	YGY	YGA	-	-
)	41		2	20	21	-	-	Y	YXV	YXG	-	-
*	42		3	22	23	24	-	Y	YXX	YXT	YXW	-
+	43		3	25	26	27	-	Y	YXC	YXY	YXA	-
,	44		3	30	31	32	-	Y	YTV	YTG	YTX	-
-	45		3	33	34	35	-	Y	YTT	YTW	YTC	-
.	46		3	36	37	40	-	Y	YTY	YTA	YWV	-
/	47		3	41	42	43	-	Y	YWG	YWX	YWT	-
0	48		4	00	01	02	03	X	XVV	XVG	XVX	XVT
1	49		4	04	05	06	07	X	XVW	XVC	XVY	XVA
2	50		4	10	11	12	13	X	XGV	XGG	XGX	XGT
3	51		4	14	15	16	17	X	XGW	XGC	XGY	XGA
4	52		4	20	21	22	23	X	XXV	XXG	XXX	XXT
5	53		4	24	25	26	27	X	XXW	XXC	XXY	XXA
6	54		4	30	31	32	33	X	XTV	XTG	XTX	XTT
7	55		4	34	35	36	37	X	XTW	XTC	XTY	XTA
8	56		4	40	41	42	43	X	XWV	XWG	XWX	XWT
9	57		4	44	45	46	47	X	XWW	XWC	XWY	XWA
:	58		2	44	45	-	-	Y	YWW	YWC	-	-
;	59		2	47	50	-	-	Y	YWA	YCV	-	-
<	60		2	51	52	-	-	Y	YCG	YCX	-	-
=	61		3	53	54	55	-	Y	YCT	YCW	YCC	-
>	62		2	56	57	-	-	Y	YCY	YCA	-	-
?	63		3	60	61	62	-	Y	YYV	YYG	YYX	-
@	64		2	63	64	-	-	Y	YYT	YYW	-	-
A	65	0.0841	4	00	01	02	03	A	AVV	AVG	AVX	AVT
B	66	0.0133	1	04	-	-	-	A	AVW	-	-	-
C	67	0.032	3	05	06	07	-	A	AVC	AVY	AVA	-
D	68	0.038	4	10	11	12	13	A	AGV	AGG	AGX	AGT
E	69	0.1216	4	14	15	16	17	A	AGW	AGC	AGY	AGA
F	70	0.0197	2	20	21	-	-	A	AXV	AXG	-	-
G	71	0.0217	1	22	-	-	-	A	AXX	-	-	-
H	72	0.0459	2	23	24	-	-	A	AXT	AXW	-	-
I	73	0.0737	3	25	26	27	-	A	AXC	AXY	AXA	-
J	74	0.0025	1	30	-	-	-	A	ATV	-	-	-
K	75	0.0084	1	31	-	-	-	A	ATG	-	-	-
L	76	0.0423	3	32	33	34	-	A	ATX	ATT	ATW	-
M	77	0.0239	3	35	36	37	-	A	ATC	ATY	ATA	-
N	78	0.0712	4	40	41	42	43	A	AWV	AWG	AWX	AWT
O	79	0.0728	4	44	45	46	47	A	AWW	AWC	AWY	AWA
P	80	0.0227	1	50	-	-	-	A	ACV	-	-	-

Q	81	0.0011	1	51	-	-	-	A	ACG	-	-	-
R	82	0.0609	3	52	53	54	-	A	ACX	ACT	ACW	-
S	83	0.0701	3	55	56	57	-	A	ACC	ACY	ACA	-
T	84	0.0913	4	60	61	62	63	A	AYV	AYG	AYX	AYT
U	85	0.0288	3	64	65	66	-	A	AYW	AYC	AYY	-
V	86	0.0093	1	67	-	-	-	A	AYA	-	-	-
W	87	0.0201	3	70	71	72	-	A	AAV	AAG	AAX	-
X	88	0.0027	1	73	-	-	-	A	AAT	-	-	-
Y	89	0.0204	2	74	75	-	-	A	AAW	AAC	-	-
Z	90	0.0013	2	76	77	-	-	A	AAY	AAA	-	-
[	91		1	65	-	-	-	Y	YYC	-	-	-
\	92		1	66	-	-	-	Y	YYY	-	-	-
]	93		1	67	-	-	-	Y	YYA	-	-	-
^	94		1	70	-	-	-	Y	YAV	-	-	-
_	95		1	71	-	-	-	Y	YAG	-	-	-
`	96		1	72	-	-	-	Y	YAX	-	-	-
a	97	0.0841	4	00	01	02	03	T	TVV	TVG	TVX	TVT
b	98	0.0133	1	04	-	-	-	T	TVW	-	-	-
c	99	0.032	3	05	06	07	-	T	TVC	TVY	TVA	-
d	100	0.038	4	10	11	12	13	T	TGV	TGG	TGX	TGT
e	101	0.1216	4	14	15	16	17	T	TGW	TGC	TGY	TGA
f	102	0.0197	2	20	21	-	-	T	TXV	TXG	-	-
g	103	0.0217	1	22	-	-	-	T	TXX	-	-	-
h	104	0.0459	2	23	24	-	-	T	TXT	TXW	-	-
i	105	0.0737	3	25	26	27	-	T	TXC	TXY	TXA	-
j	106	0.0025	1	30	-	-	-	T	TTV	-	-	-
k	107	0.0084	1	31	-	-	-	T	TTG	-	-	-
l	108	0.0423	3	32	33	34	-	T	TTX	TTT	TTW	-
m	109	0.0239	3	35	36	37	-	T	TTC	TTY	TTA	-
n	110	0.0712	4	40	41	42	43	T	TWV	TWG	TWX	TWT
o	111	0.0728	4	44	45	46	47	T	TWW	TWC	TWY	TWA
p	112	0.0227	1	50	-	-	-	T	TCV	-	-	-
q	113	0.0011	1	51	-	-	-	T	TCG	-	-	-
r	114	0.0609	3	52	53	54	-	T	TCX	TCT	TCW	-
s	115	0.0701	3	55	56	57	-	T	TCC	TCY	TCA	-
t	116	0.0913	4	60	61	62	63	T	TYV	TYG	TYX	TYT
u	117	0.0288	3	64	65	66	-	T	TYW	TYC	TYY	-
v	118	0.0093	1	67	-	-	-	T	TYA	-	-	-
w	119	0.0201	3	70	71	72	-	T	TAV	TAG	TAX	-
x	120	0.0027	1	73	-	-	-	T	TAT	-	-	-
y	121	0.0204	2	74	75	-	-	T	TAW	TAC	-	-
z	122	0.0013	2	76	77	-	-	T	TAY	TAA	-	-
{	123		1	73	-	-	-	Y	YAT	-	-	-

}	124	1	74	-	-	-	Y	YAW	-	-	-
~	125	1	75	-	-	-	Y	YAC	-	-	-
	126	1	76	-	-	-	Y	YAY	-	-	-
	127	1	77	-	-	-	Y	YAA	-	-	-
	128		40	-	-	-	G	GWV	-	-	-
	129		41	-	-	-	G	GWG	-	-	-
	130		42	-	-	-	G	GWX	-	-	-
	131		43	-	-	-	G	GWT	-	-	-
	132		44	-	-	-	G	GWW	-	-	-
	133		45	-	-	-	G	GWC	-	-	-
	134		46	-	-	-	G	GWY	-	-	-
	135		47	-	-	-	G	GWA	-	-	-
	136		50	-	-	-	G	GCV	-	-	-
	137		51	-	-	-	G	GCG	-	-	-
	138		52	-	-	-	G	GCX	-	-	-
	139		53	-	-	-	G	GCT	-	-	-
	140		54	-	-	-	G	GCW	-	-	-
	141		55	-	-	-	G	GCC	-	-	-
	142		56	-	-	-	G	GCY	-	-	-
	143		57	-	-	-	G	GCA	-	-	-
	144		60	-	-	-	G	GYV	-	-	-
	145		61	-	-	-	G	GYG	-	-	-
	146		62	-	-	-	G	GYX	-	-	-
	147		63	-	-	-	G	GYT	-	-	-
	148		64	-	-	-	G	GYW	-	-	-
	149		65	-	-	-	G	GYC	-	-	-
	150		66	-	-	-	G	GYY	-	-	-
	151		67	-	-	-	G	GYA	-	-	-
	152		70	-	-	-	G	GAV	-	-	-
	153		71	-	-	-	G	GAG	-	-	-
	154		72	-	-	-	G	GAX	-	-	-
	155		73	-	-	-	G	GAT	-	-	-
	156		74	-	-	-	G	GAW	-	-	-
	157		75	-	-	-	G	GAC	-	-	-
	158		76	-	-	-	G	GAY	-	-	-
	159		77	-	-	-	G	GAA	-	-	-
	160		00	-	-	-	C	CVV	-	-	-
	161		01	-	-	-	C	CVG	-	-	-
	162		02	-	-	-	C	CVX	-	-	-
	163		03	-	-	-	C	CVT	-	-	-
	164		04	-	-	-	C	CVW	-	-	-
	165		05	-	-	-	C	CVC	-	-	-
	166		06	-	-	-	C	CVY	-	-	-

167	07	-	-	-	C	CVA	-	-	-
168	10	-	-	-	C	CGV	-	-	-
169	11	-	-	-	C	CGG	-	-	-
170	12	-	-	-	C	CGX	-	-	-
171	13	-	-	-	C	CGT	-	-	-
172	14	-	-	-	C	CGW	-	-	-
173	15	-	-	-	C	CGC	-	-	-
174	16	-	-	-	C	CGY	-	-	-
175	17	-	-	-	C	CGA	-	-	-
176	20	-	-	-	C	CXV	-	-	-
177	21	-	-	-	C	CXG	-	-	-
178	22	-	-	-	C	CXX	-	-	-
179	23	-	-	-	C	CXT	-	-	-
180	24	-	-	-	C	CXW	-	-	-
181	25	-	-	-	C	CXC	-	-	-
182	26	-	-	-	C	CXY	-	-	-
183	27	-	-	-	C	CXA	-	-	-
184	30	-	-	-	C	CTV	-	-	-
185	31	-	-	-	C	CTG	-	-	-
186	32	-	-	-	C	CTX	-	-	-
187	33	-	-	-	C	CTT	-	-	-
188	34	-	-	-	C	CTW	-	-	-
189	35	-	-	-	C	CTC	-	-	-
190	36	-	-	-	C	CTY	-	-	-
191	37	-	-	-	C	CTA	-	-	-
192	40	-	-	-	C	CWV	-	-	-
193	41	-	-	-	C	CWG	-	-	-
194	42	-	-	-	C	CWX	-	-	-
195	43	-	-	-	C	CWT	-	-	-
196	44	-	-	-	C	CWW	-	-	-
197	45	-	-	-	C	CWC	-	-	-
198	46	-	-	-	C	CWY	-	-	-
199	47	-	-	-	C	CWA	-	-	-
200	50	-	-	-	C	CCV	-	-	-
201	51	-	-	-	C	CCG	-	-	-
202	52	-	-	-	C	CCX	-	-	-
203	53	-	-	-	C	CCT	-	-	-
204	54	-	-	-	C	CCW	-	-	-
205	55	-	-	-	C	CCC	-	-	-
206	56	-	-	-	C	CCY	-	-	-
207	57	-	-	-	C	CCA	-	-	-
208	60	-	-	-	C	CYV	-	-	-
209	61	-	-	-	C	CYG	-	-	-

210	62	-	-	-	C	CYX	-	-	-
211	63	-	-	-	C	CYT	-	-	-
212	64	-	-	-	C	CYW	-	-	-
213	65	-	-	-	C	CYC	-	-	-
214	66	-	-	-	C	CYY	-	-	-
215	67	-	-	-	C	CYA	-	-	-
216	70	-	-	-	C	CAV	-	-	-
217	71	-	-	-	C	CAG	-	-	-
218	72	-	-	-	C	CAX	-	-	-
219	73	-	-	-	C	CAT	-	-	-
220	74	-	-	-	C	CAW	-	-	-
221	75	-	-	-	C	CAC	-	-	-
222	76	-	-	-	C	CAY	-	-	-
223	77	-	-	-	C	CAA	-	-	-
224	00	-	-	-	V	VVV	-	-	-
225	01	-	-	-	V	VVG	-	-	-
226	02	-	-	-	V	VVX	-	-	-
227	03	-	-	-	V	VVT	-	-	-
228	04	-	-	-	V	VVW	-	-	-
229	05	-	-	-	V	VVC	-	-	-
230	06	-	-	-	V	VVY	-	-	-
231	07	-	-	-	V	VVA	-	-	-
232	10	-	-	-	V	VGV	-	-	-
233	11	-	-	-	V	VGG	-	-	-
234	12	-	-	-	V	VGX	-	-	-
235	13	-	-	-	V	VGT	-	-	-
236	14	-	-	-	V	VGW	-	-	-
237	15	-	-	-	V	VGC	-	-	-
238	16	-	-	-	V	VGY	-	-	-
239	17	-	-	-	V	VGA	-	-	-
240	20	-	-	-	V	VXV	-	-	-
241	21	-	-	-	V	VXG	-	-	-
242	22	-	-	-	V	VXX	-	-	-
243	23	-	-	-	V	VXT	-	-	-
244	24	-	-	-	V	VXW	-	-	-
245	25	-	-	-	V	VXC	-	-	-
246	26	-	-	-	V	VXY	-	-	-
247	27	-	-	-	V	VXA	-	-	-
248	30	-	-	-	V	VTV	-	-	-
249	31	-	-	-	V	VTG	-	-	-
250	32	-	-	-	V	VTX	-	-	-
251	33	-	-	-	V	VTT	-	-	-
252	34	-	-	-	V	VTW	-	-	-

253	35	-	-	-	V	VTC	-	-	-
254	36	-	-	-	V	VTY	-	-	-
255	37	-	-	-	V	VTA	-	-	-

Document start: WGX
Document end: WYA

[1] Nucleotides in second and third positions of triplets were obtained by assigning nt V, G, X, T, W, C, Y, A to digits from 0 to 7, correspondingly.

Table 3. Prediction of the mutation frequency of the third nt in triplets affected by different frequency and redundancy.

Letter[a]	FO[b]	R[c]	AMF[d]
A	0,0812	4	8,8042E-08
B	0,0149	1	2,8272E-08
C	0,0271	3	3,6729E-08
D	0,0432	3	5,8550E-08
E	0,1202	4	1,3033E-07
F	0,0230	2	3,7407E-08
G	0,0203	2	3,3016E-08
H	0,0592	3	8,0235E-08
I	0,0731	4	7,9260E-08
J	0,0010	1	1,8975E-09
K	0,0069	1	1,3092E-08
L	0,0398	3	5,3942E-08
M	0,0261	3	3,5374E-08
N	0,0695	4	7,5356E-08
O	0,0768	4	8,3271E-08
P	0,0182	1	3,4534E-08
Q	0,0011	1	2,0872E-09
R	0,0602	3	8,1591E-08
S	0,0628	3	8,5115E-08
T	0,0910	4	9,8668E-08
U	0,0288	3	3,9033E-08
V	0,0111	1	2,1062E-08
W	0,0209	2	3,3992E-08
X	0,0017	1	3,2257E-09
Y	0,0211	2	3,4317E-08
Z	0,0007	1	1,3282E-09
			1,2697E-06

[a] The frequency of use of letters is mainly due to lowercase letters. For this work I am assuming for uppercase letters the same frequency as for lowercase letters.
[b] FO: Frequency of occurrence of each letter in English writings (Richardson et al. 2004).
[c] R: number of different triplets encoding each letter.
[d] AMF: Mutation frequency affected by different FO and redundancy was obtained by using the algorithm:

$$AMF = \text{mut.-freq.} \times FO \times (8–R)/7$$

being mut.-freq. that obtained from Church et al. (2012).