

Beyond The Genome 2012 Informatics Challenge

James Taylor

Michael Schatz

David Dooling

The Prize



Computing



doodleaday.wordpress.com

The Challenge

Reads were generated by taking a portion of an organism's reference sequence and inserting a “DNA-encoded” famous quote into the sequence. Your challenge is to identify the inserted sequence, decode the quote, and identify its speaker. The first person to send the correct quote and its speaker to btg2012info@gmail.com wins.

The Response

The winner will be announce on
Twitter @ddgenome



The Data

<http://goo.gl/3Zwkk>

<http://genome.wustl.edu/pub/user/ddooling/BeyondTheGenome2012InformaticsChallenge.tar.gz>

<ftp://genome.wustl.edu/pub/user/ddooling/BeyondTheGenome2012InformaticsChallenge.tar.gz>



To the **bold**

You are free to proceed

To the brave

We will present more information
about the challenge

To the bewildered

We will pause, then discuss
possible approaches

DNA Encoding

The screenshot shows the Science magazine website interface. At the top, there is a search bar with 'Science Magazine' selected and a search button. Below the search bar is a navigation menu with links for AAAS.ORG, FEEDBACK, HELP, LIBRARIANS, WASHINGTON UNIV SCHL OF MED, ALERTS, ACCESS RIGHTS, MY ACCOUNT, and SIGN IN. A secondary navigation bar includes NEWS, SCIENCE JOURNALS, CAREERS, BLOGS & COMMUNITIES, MULTIMEDIA, COLLECTIONS, and JOIN / SUBSCRIBE. The main header features the Science logo and the tagline 'The World's Leading Journal of Original Scientific Research, Global News, and Commentary.' Below this is a sub-navigation bar with links for Science Home, Current Issue, Previous Issues, Science Express, Science Products, My Science, and About the Journal. The article page shows the breadcrumb 'Home > Science Magazine > Science Express > Church et al.' and the article title 'Next-Generation Digital Information Storage in DNA' by George M. Church^{1,2}, Yuan Gao³, and Sriram Kosuri^{1,2,*}. The abstract text is: 'Digital information is accumulating at an astounding rate, straining our ability to store and archive it. DNA is among the most dense and stable information media known. The development of new technologies in both DNA synthesis and sequencing make DNA an increasingly feasible digital storage medium. Here, we develop a strategy to encode arbitrary digital information in DNA, write a 5.27-megabit book using DNA microchips, and read the book using next-generation DNA sequencing.' On the right side, there are two advertisements: 'Get all of Science' with a 'Join Now!' button and 'WEBINAR: Clinical Validation of Cancer Biomarker Signatures using Array Technology' featuring a DNA double helix image.

DOI: 10.1126/science.1226355

<https://www.sciencemag.org/content/early/2012/08/15/science.1226355.full>

DNA Encoding

Text Hello, world!

ASCII

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040	 	Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	\$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	((72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051))	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[[123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174	|	
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135]]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177		DEL

DNA Encoding

Text Hello, world!

ASCII 72 101 108 108 111 44 32 119 111 114
 108 100 33

Bits and Bytes

- Bit: off or on, 0 or 1
- Byte: eight bits, 01010101
- Each position represents a power of 2
- All characters in the ASCII set can be represented by one byte (0 – 127 (= 2^7-1))

DNA Encoding

Text Hello, world!

ASCII 72 101 108 108 111 44 32 119 111 114
108 100 33

Binary 01001000 01100101 01101100 01101100
01101111 00101100 00100000 01110111
01101111 01110010 01101100 01100100

DNA Encoding

- 0 becomes A or C
- 1 becomes T or G

DNA Encoding

Text	Hello, world!												
ASCII	72	101	108	108	111	44	32	119	111	114	108	100	33
Binary	01001000	01100101	01101100	01101100	01101111	00101100	00100000	01110111	01101111	01110010	01101100	01100100	
DNA	AGCAGCCC	ATTCCGAT	CTTATTAC	CTTCTTCC	CGGAGGGG	AATATGCC	ACTACCCA	ATGTATTT	ATTCTTGT	ATTTAATC	CTGCGGAA	CGTAAGCC	

How we did it

1. Downloaded a reference sequence from NCBI
2. Excised out a chunk
3. Encoded the quote into DNA
4. Randomly inserted the quote-DNA into the excised chunk from the reference
5. Generated simulated reads from the new reference

What you get

<code>dna-encode.pl</code>	Perl script to encode/decode text to/from DNA
<code>i2x100f180.1.fq</code>	Read 1 of Illumina 2x100 reads from 180+/-20 bp fragments
<code>i2x100f180.2.fq</code>	Read 2 of Illumina 2x100 reads from 180+/-20 bp fragments
<code>i2x50f2000.1.fq</code>	Read 1 of Illumina 2x50 reads from 2+/-0.2 kbp fragments
<code>i2x50f2000.2.fq</code>	Read 2 of Illumina 2x50 reads from 2+/-0.2 kbp fragments
<code>i2x250f700.fq</code>	Interleaved reads 1 and 2 of Illumina 2x250 reads from 700+/-50 bp fragments

dna-encode.pl

NAME

dna-encode – encode and decode ASCII text into DNA

SYNOPSIS

dna-encode [OPTIONS]... [FILE]...

DESCRIPTION

This script encodes a string of characters first into big endian (network order) binary and then into DNA where zero become A or C and one becomes G or T.

This implementation is based on the algorithm described in George M. Church, Yuan Gao, and Sriram Kosuri. Next-Generation Digital Information Storage in DNA. Science 2012. DOI: 10.1126/science.1226355.

dna-encode.pl

OPTIONS

- d, --decode
Decode a DNA sequence into a message rather than the default of encoding a message into DNA.
- l, --little-endian
Encode/decode characters using little endian byte order rather than the default big endian byte order.
- r, --reverse-complement
Reverse complement the DNA after encoding (or before decoding).
- verbose
Output intermediate binary when encoding/decoding.

Pause for Questions?