Beyond the Genome 2013

Informatics Challenge

Michael Schatz
Sven-Eric Schelhorn, October 2013

The prize

iPad mini

The problem, 1

Reads

A metagenomic sample was generated by mixing portions of the reference sequences of several microbial species. Sequence reads were simulated from these portions. Within each portion of a reference sequence, a foreign insert *(i.e, not originating from any of the microbial species) was placed. These inserts encode a* message.

The problem, 2

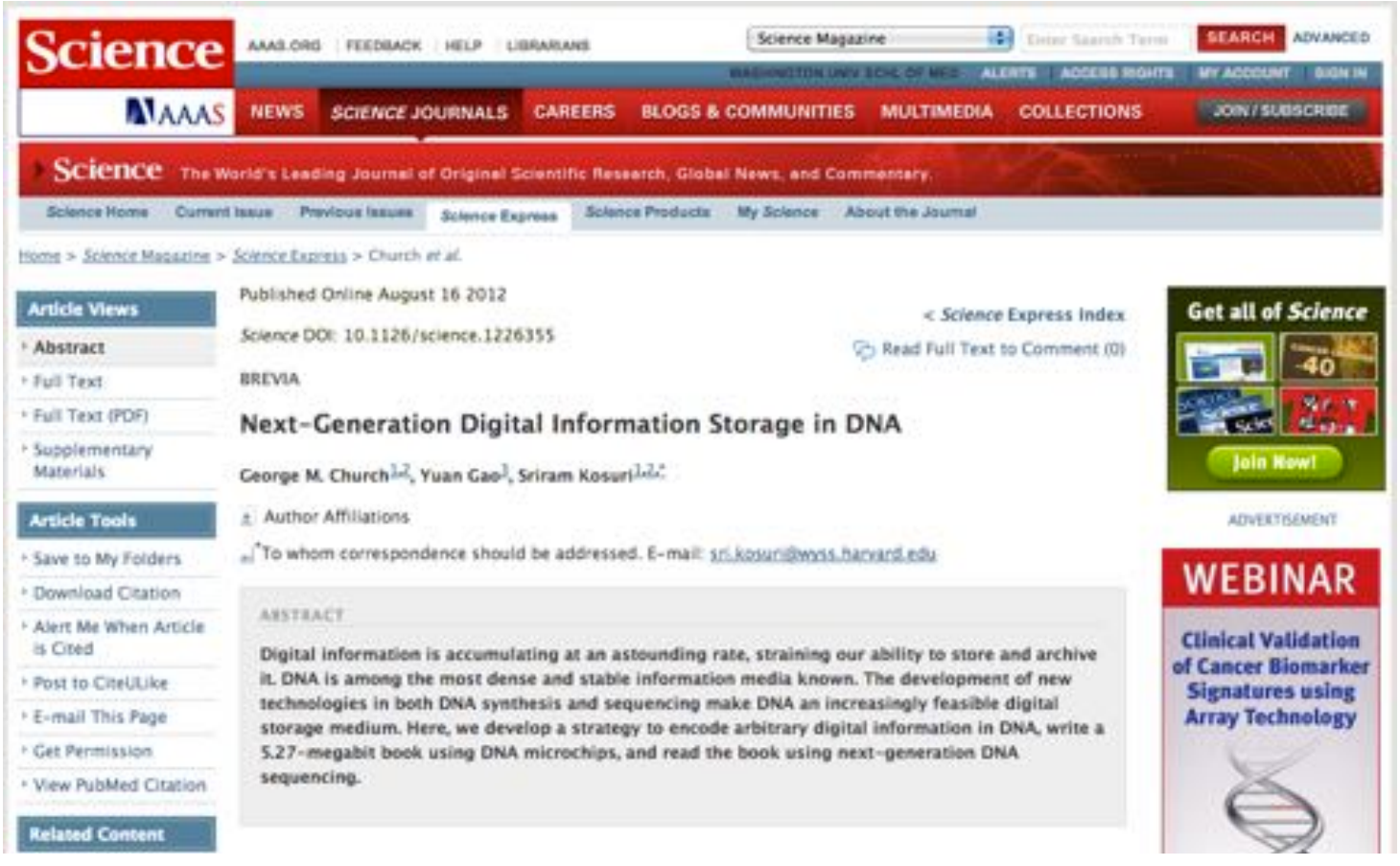
Message

One of the inserts corresponds to the ‘wildtype’, *as deposited in public sequence databases such as NCBI nt. The other inserts are slight variations of this wildtype (>90% nuc. similarity).*

The message we are seeking is encoded as nucleotide variants of the non-wildtype inserts with respect to the wildtype insert.

Consequently, there is one message for each non-wildtype insert in the read data. All messages together yield a quote that is the solution to the challenge.

How to encode a message into DNA



Science AAAS.ORG | FEEDBACK | HELP | LIBRARIANS

Science Magazine | SEARCH | ADVANCED

ALERTS | ACCESS RIGHTS | MY ACCOUNT | SIGN IN

AAAS NEWS | SCIENCE JOURNALS | CAREERS | BLOGS & COMMUNITIES | MULTIMEDIA | COLLECTIONS | JOIN / SUBSCRIBE

Science The World's Leading Journal of Original Scientific Research, Global News, and Commentary.

Science Home | Current Issue | Previous Issues | Science Express | Science Products | My Science | About the Journal

Home > Science Magazine > Science Express > Church et al.

Article Views
- Abstract
- Full Text
- Full Text (PDF)
- Supplementary Materials

Article Tools
- Save to My Folders
- Download Citation
- Alert Me When Article is Cited
- Post to CiteULike
- E-mail This Page
- Get Permission
- View PubMed Citation

Related Content

Published Online August 16 2012

Science DOI: 10.1126/science.1226355

< Science Express Index

Read Full Text to Comment (0)

BREVIA

Next-Generation Digital Information Storage in DNA

George M. Church[1,2], Yuan Gao[1], Sriram Kosuri[1,2,*]

Author Affiliations

*To whom correspondence should be addressed. E-mail: sri.kosuri@wyss.harvard.edu

ABSTRACT

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.

Numbers are bits

Bit: off or on, 0 or 1
Byte: eight bits, 01001000

Each position represents a power of 2 :

$$01001000 = 8 + 64 = 72$$

All characters in the ASCII set can be represented by one byte (0 – 127 (= 27-1)

ASCII code table

Dec	Hx	Oct	Char
0	0	000	NUL (null)
1	1	001	SOH (start of heading)
2	2	002	STX (start of text)
3	3	003	ETX (end of text)
4	4	004	EOT (end of transmission)
5	5	005	ENQ (enquiry)
6	6	006	ACK (acknowledge)
7	7	007	BEL (bell)
8	8	010	BS (backspace)
9	9	011	TAB (horizontal tab)
10	A	012	LF (NL line feed, new line)
11	B	013	VT (vertical tab)
12	C	014	FF (NP form feed, new page)
13	D	015	CR (carriage return)
14	E	016	SO (shift out)
15	F	017	SI (shift in)
16	10	020	DLE (data link escape)
17	11	021	DC1 (device control 1)
18	12	022	DC2 (device control 2)
19	13	023	DC3 (device control 3)
20	14	024	DC4 (device control 4)
21	15	025	NAK (negative acknowledge)
22	16	026	SYN (synchronous idle)
23	17	027	ETB (end of trans. block)
24	18	030	CAN (cancel)
25	19	031	EM (end of medium)
26	1A	032	SUB (substitute)
27	1B	033	ESC (escape)
28	1C	034	FS (file separator)
29	1D	035	GS (group separator)
30	1E	036	RS (record separator)
31	1F	037	US (unit separator)

Dec	Hx	Oct	Html	Chr
32	20	040		Space
33	21	041	!	!
34	22	042	"	"
35	23	043	#	#
36	24	044	$	$
37	25	045	%	%
38	26	046	&	&
39	27	047	'	'
40	28	050	(	(
41	29	051	)	)
42	2A	052	*	*
43	2B	053	+	+
44	2C	054	,	,
45	2D	055	-	-
46	2E	056	.	.
47	2F	057	/	/
48	30	060	0	0
49	31	061	1	1
50	32	062	2	2
51	33	063	3	3
52	34	064	4	4
53	35	065	5	5
54	36	066	6	6
55	37	067	7	7
56	38	070	8	8
57	39	071	9	9
58	3A	072	:	:
59	3B	073	;	;
60	3C	074	<	<
61	3D	075	=	=
62	3E	076	>	>
63	3F	077	?	?

Dec	Hx	Oct	Html	Chr
64	40	100	@	@
65	41	101	A	A
66	42	102	B	B
67	43	103	C	C
68	44	104	D	D
69	45	105	E	E
70	46	106	F	F
71	47	107	G	G
72	48	110	H	H
73	49	111	I	I
74	4A	112	J	J
75	4B	113	K	K
76	4C	114	L	L
77	4D	115	M	M
78	4E	116	N	N
79	4F	117	O	O
80	50	120	P	P
81	51	121	Q	Q
82	52	122	R	R
83	53	123	S	S
84	54	124	T	T
85	55	125	U	U
86	56	126	V	V
87	57	127	W	W
88	58	130	X	X
89	59	131	Y	Y
90	5A	132	Z	Z
91	5B	133	[	[
92	5C	134	\	\
93	5D	135	]	]
94	5E	136	^	^
95	5F	137	_	_

Dec	Hx	Oct	Html	Chr
96	60	140	`	`
97	61	141	a	a
98	62	142	b	b
99	63	143	c	c
100	64	144	d	d
101	65	145	e	e
102	66	146	f	f
103	67	147	g	g
104	68	150	h	h
105	69	151	i	i
106	6A	152	j	j
107	6B	153	k	k
108	6C	154	l	l
109	6D	155	m	m
110	6E	156	n	n
111	6F	157	o	o
112	70	160	p	p
113	71	161	q	q
114	72	162	r	r
115	73	163	s	s
116	74	164	t	t
117	75	165	u	u
118	76	166	v	v
119	77	167	w	w
120	78	170	x	x
121	79	171	y	y
122	7A	172	z	z
123	7B	173	{	{
124	7C	174			|
125	7D	175	}	}
126	7E	176	~	~
127	7F	177	&#127;	DEL

Example

Text `Hello, world!`

Text can be numbers

Text	`Hello, world!`
ASCII	`72 101 108 108 111 44 32 119 111 114 108 100 33`

Binary text

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

Full example

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 prepared the data

1. *Downloaded a couple of reference sequences from NCBI*
2. *Excised out a chunk each*
3. *Took a DNA sequence (the 'wildtype insert') and inserted it into one of the chunks*
4. *Made copies of the wildtype insert and encoded messages as nucleotide variants with respect to the wildtype*
5. *Inserted one of the resulting variant inserts into each of the remaining reference sequences*
6. *Generated simulated reads from all new references at different coverages (metagenomics: non-uniform coverages)*

Variant encoding

(This example encodes only one letter, the real messages are longer parts of sentences)

OLD WAY: Full length encoding (absolute)

A	G	C	A	G	C	C	C

= 01001000 = H…

TODAY: variant encoding (relative)

wild

var1

A	G	C	A	G	C	C	C

= 01001000 = H…

What you get

dna-encode.pl	Perl script to encode/decode text to/ from DNA
sh_end_1.fastq.gz sh_end_2.fastq.gz	Paired end read data from the mixed references, fastq-format, 2x250bp from 1000+/-50bp fragments (inner distance 500+/-50bp)
lo_end_1.fastq.gz lo_end_2.fastq.gz	Paired end read data from the mixed references, fastq-format, 2x150bp from 5300+/-500bp fragments (inner distance 5000+/-500bp)

The decoder script

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.

The decoder script

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.
```

Tip 1

It's a *metagenomic* sample.
Choose your tools accordingly.

Tip 2

After you identified an insert, you need to identify the insert *wildtype*. There are several ways to distinguish it from the variants. BLAST, consensus, pairwise similarities…

Tip 3

NCBI Blast may be unreliable due to the Government Shutdown.
If yes, try to use the public BLAST server at EBI/EMBL (WU-BLAST).

Questions?

Are there any questions?

Otherwise, the link to the read data and the email address to send the answer to will follow next.

The data

Read data (~10mb)

Can be obtained now at:

http://schatzlab.cshl.edu/btg2013.tgz

Answer (quote and author of quote) should be sent to:

beyondthegenome2013@gmail.com

Winner is announced today at about 4pm