STORAGE DEVELOPER CONFERENCE

SDC 22

Fremont, CA
September 12-15, 2022

BY Developers FOR Developers

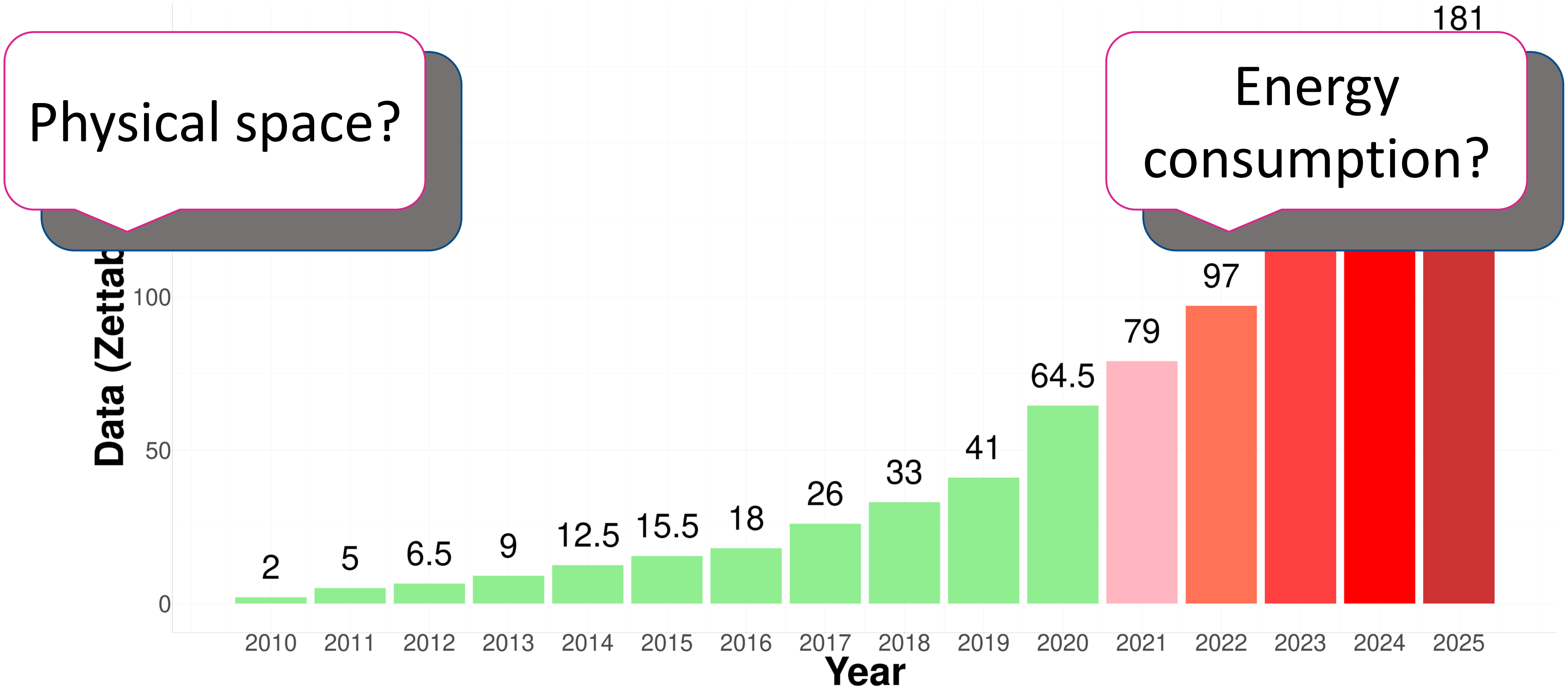
A SNIA Event

DNA Data Storage: a Decade of Coding and Decoding, How Far Have We Got?

João Henrique Diniz Brandão Gervásio (Jay Gervasio)

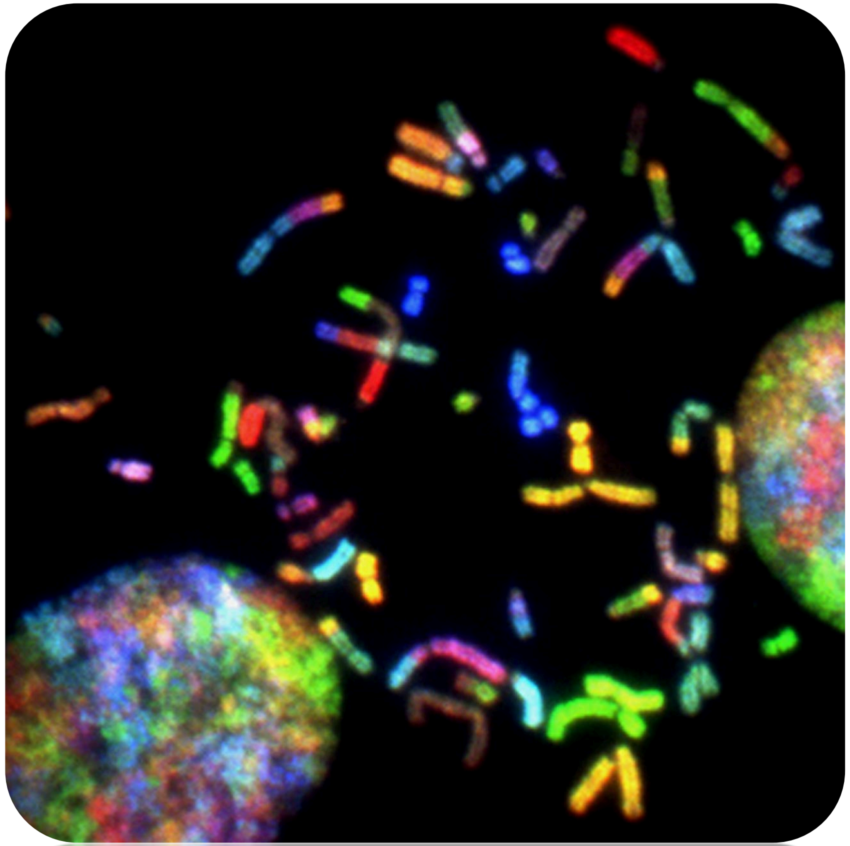
Adriano Galindo Leal, PhD, EE

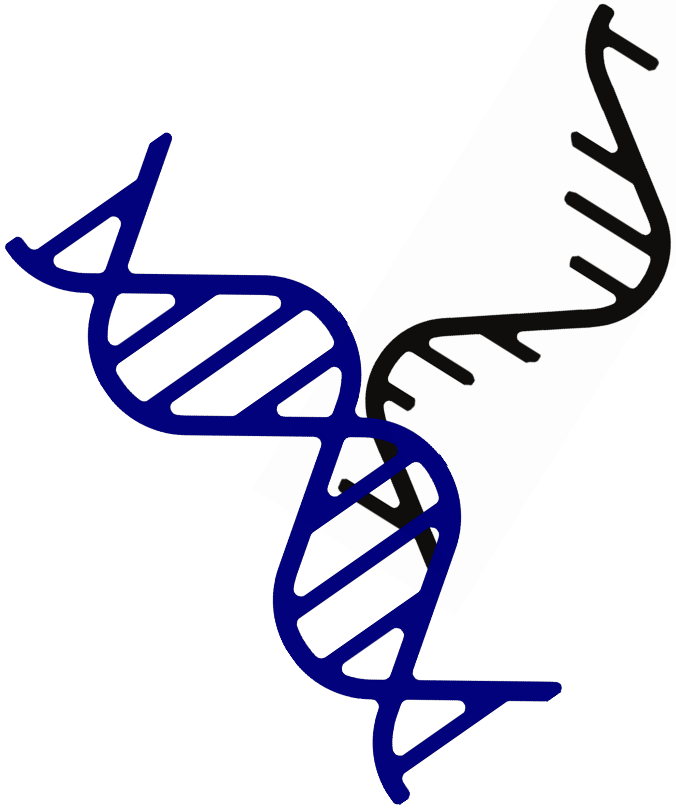
The problem we all face



Source: IDC report Worldwide Global DataSphere Forecast, 2021-2025

Biology and Data Storage





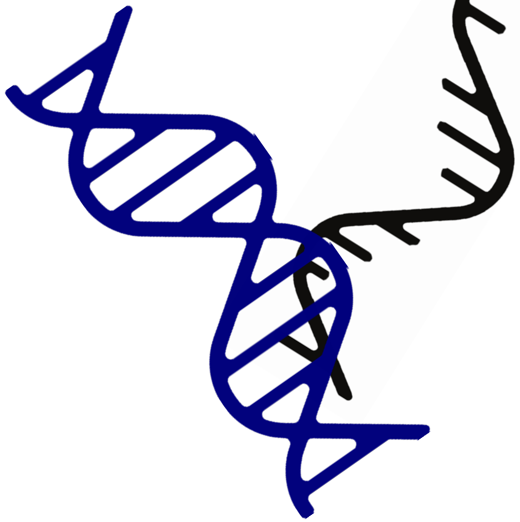
Biology and Data Storage

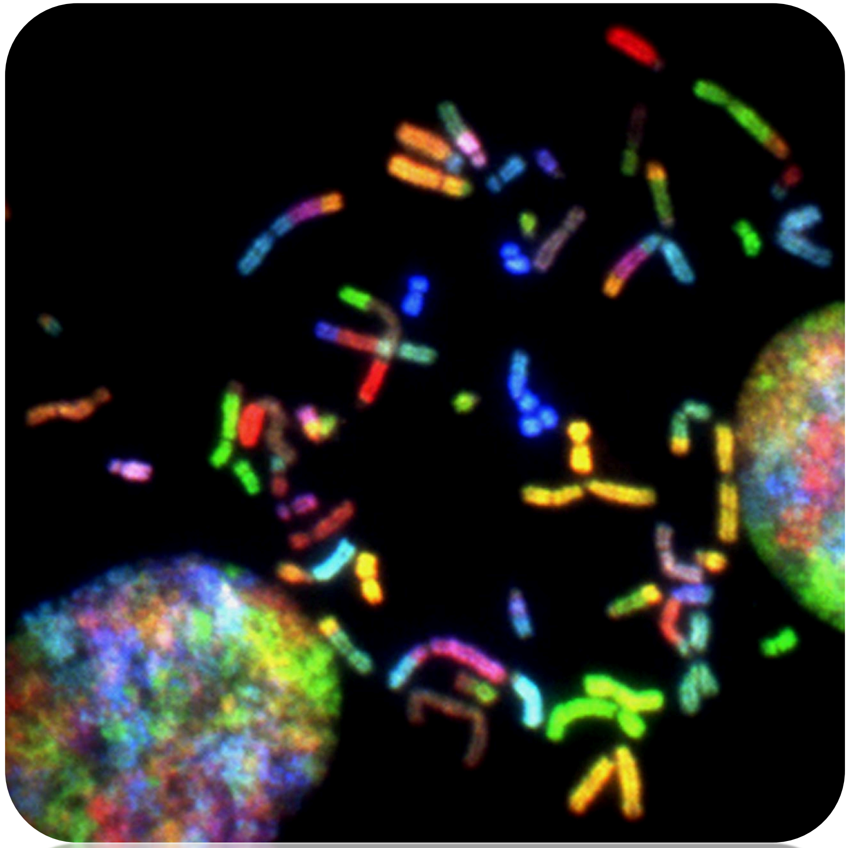
DNA and **RNA** already store information

- Billions of bases in a genome
- A, C, T, G

A=T

C=G





Can we store digital data into DNA?

Richard Feynman (1958)

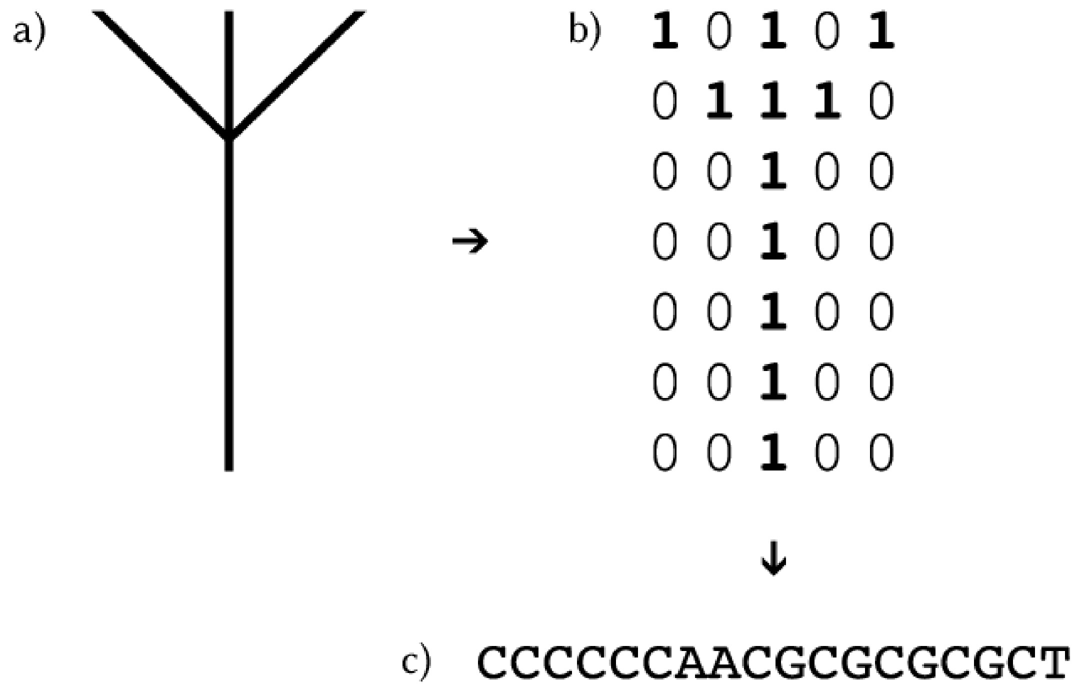
"There's plenty of room at the bottom"

Can we store digital data into DNA?

Microvenus

Phase-change coding

C = X
T = XX
A = XXX
G = XXXX

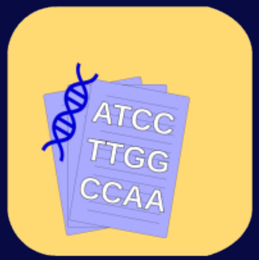


Davis, J. (1996). Microvenus. *Art Journal*, *55*(1), 70. https://doi.org/10.2307/777811

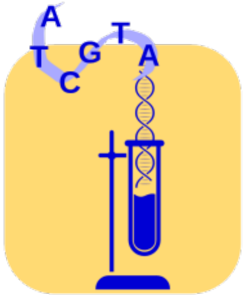
How is it done?

1. Converting information to binary code

2. Converting binary code to DNA code



3. DNA Synthesis



4. Storage

5. Recovery

6. Sequencing

7. Decoding

Church, Gao & Kosuri, 2012

Bit	Base
0	A or C
1	T or G

5.27 Megabits

2012

Next-Generation Digital Information Storage in DNA

George M. Church,[1,2] Yuan Gao,[3] Sriram Kosuri[1,2*]

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

Lenovo | ipt

STORAGE DEVELOPER CONFERENCE SDC 22

DNA data storage constraints

Synthesis of short sequences

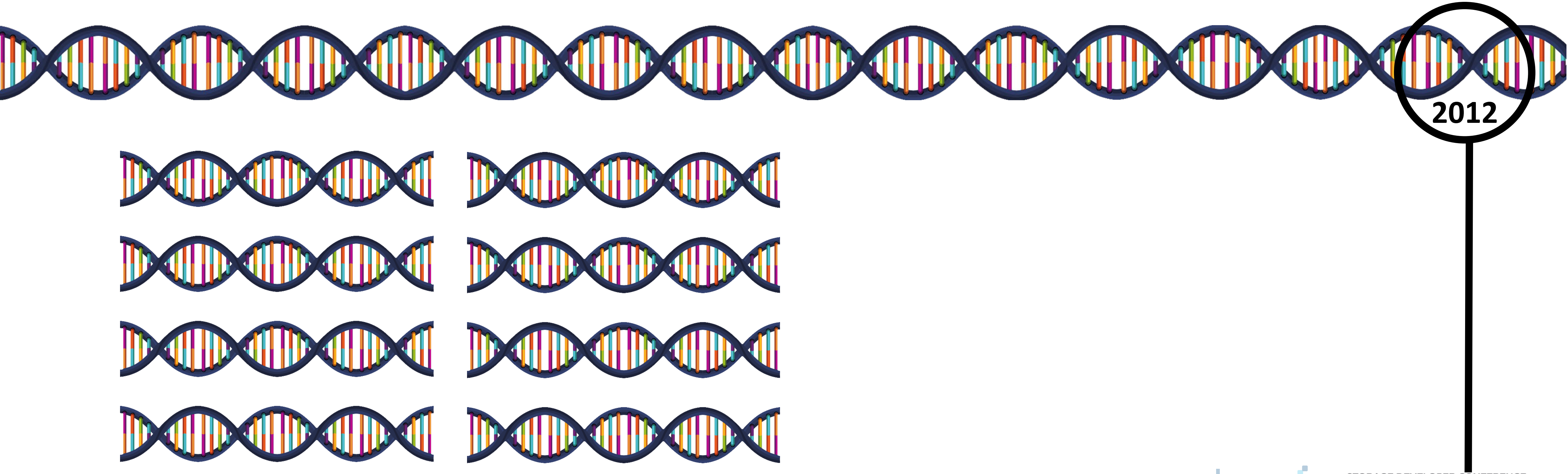
Avoid repetition of the same base (homopolymer)

Keep GC-content around 50%

2012

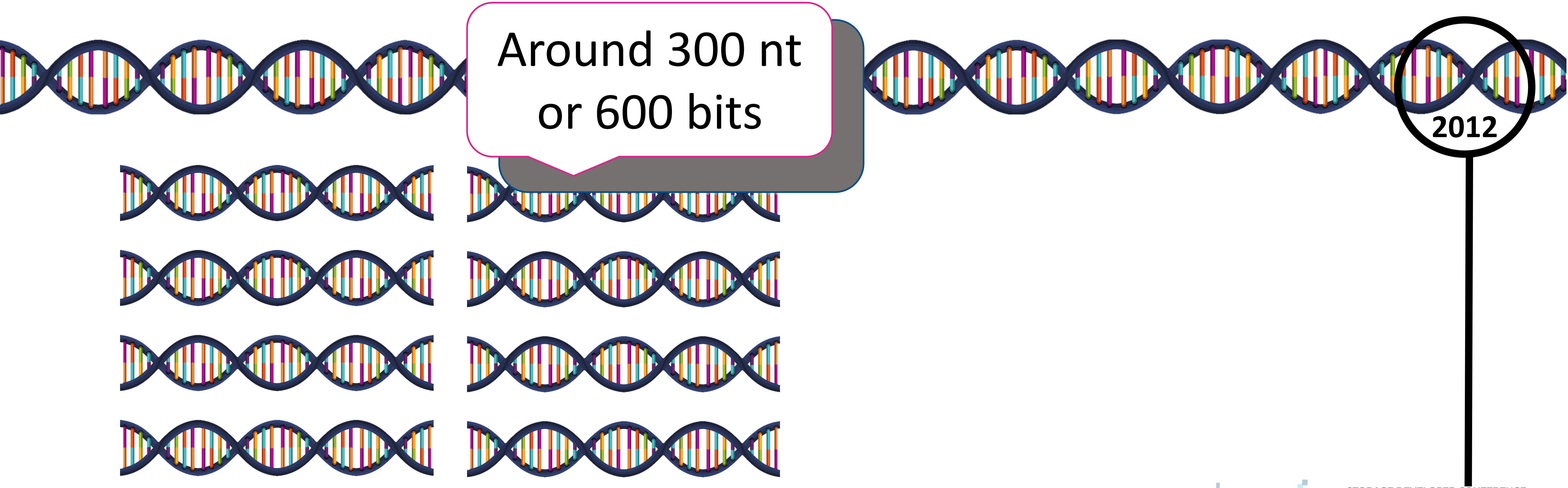
DNA data storage constraints

Synthesis of short sequences



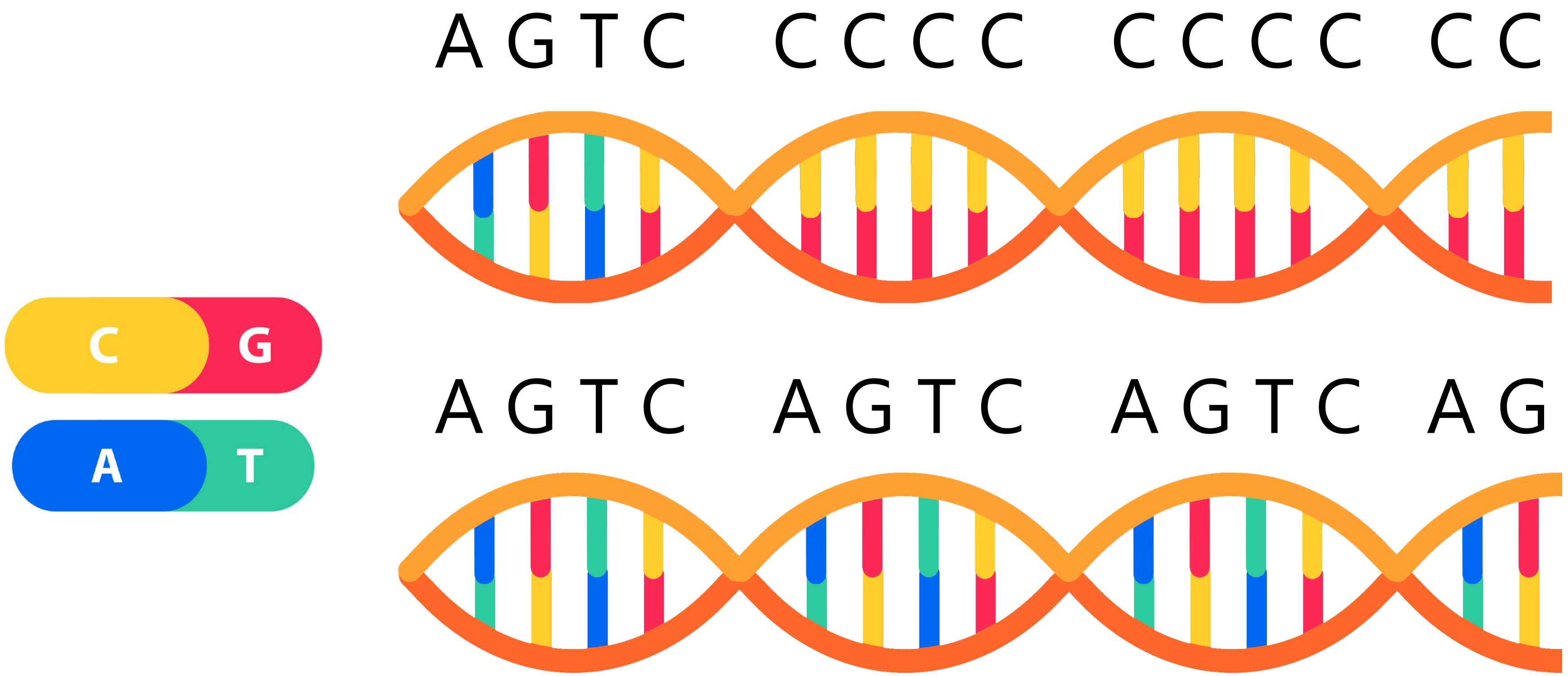
DNA data storage constraints

Synthesis of short sequences



DNA data storage constraints

Avoid repetition of the same base (homopolymer)



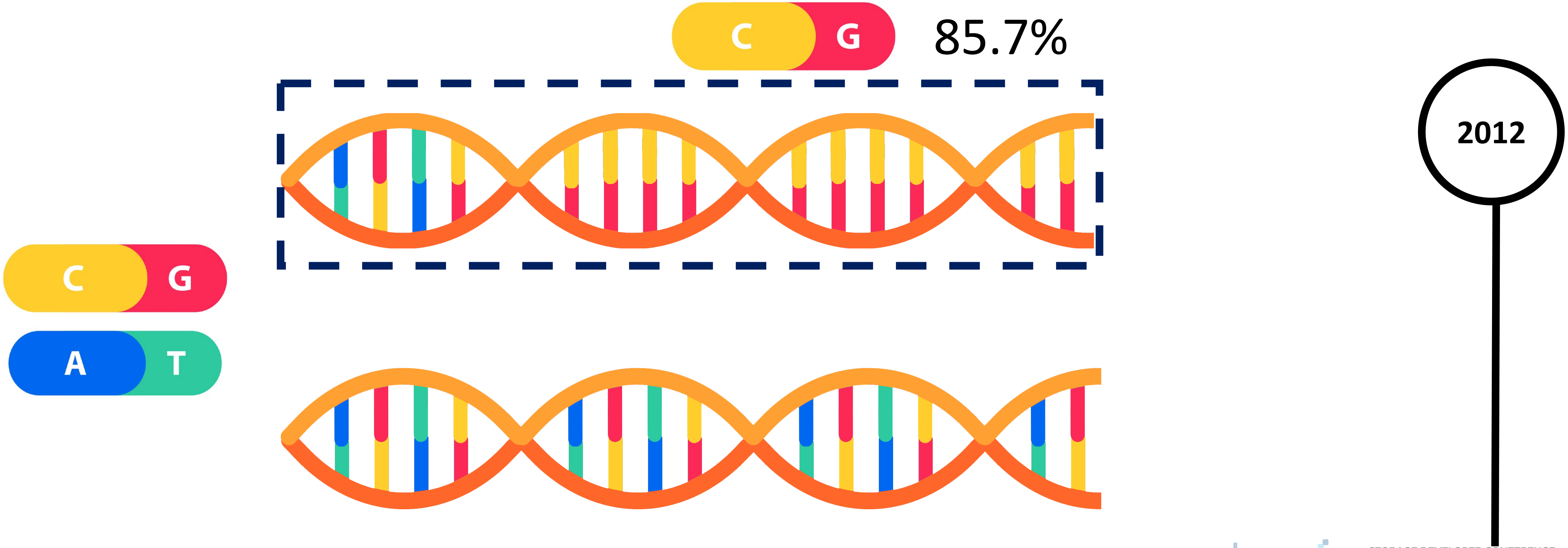
2012

Lenovo | ipt

STORAGE DEVELOPER CONFERENCE SDC 22

DNA data storage constraints

Keep GC-content around 50%



Church, Gao & Kosuri, 2012

Binary: 01101000011011111101110111

DNA general: ATTATAAAATTATTTTATTTATTT

DNA const: ATGCTACACTGCTGTGATGTCTGT

2012

No homopolymer and 45% of GC!!!

Bit	Base
0	A or C
1	T or G

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

Avoid all homopolymers!

Previous base	next trit to encode		
	0	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

2013

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

STORAGE DEVELOPER CONFERENCE

Avoid all homopolymers!

Base	Number	Digits
16	17	2
2	010111	6
3	0212	4

2013

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

Lenovo | ipt

STORAGE DEVELOPER CONFERENCE SDC 22

Avoid all homopolymers!

Base 3: 0 2 1 2

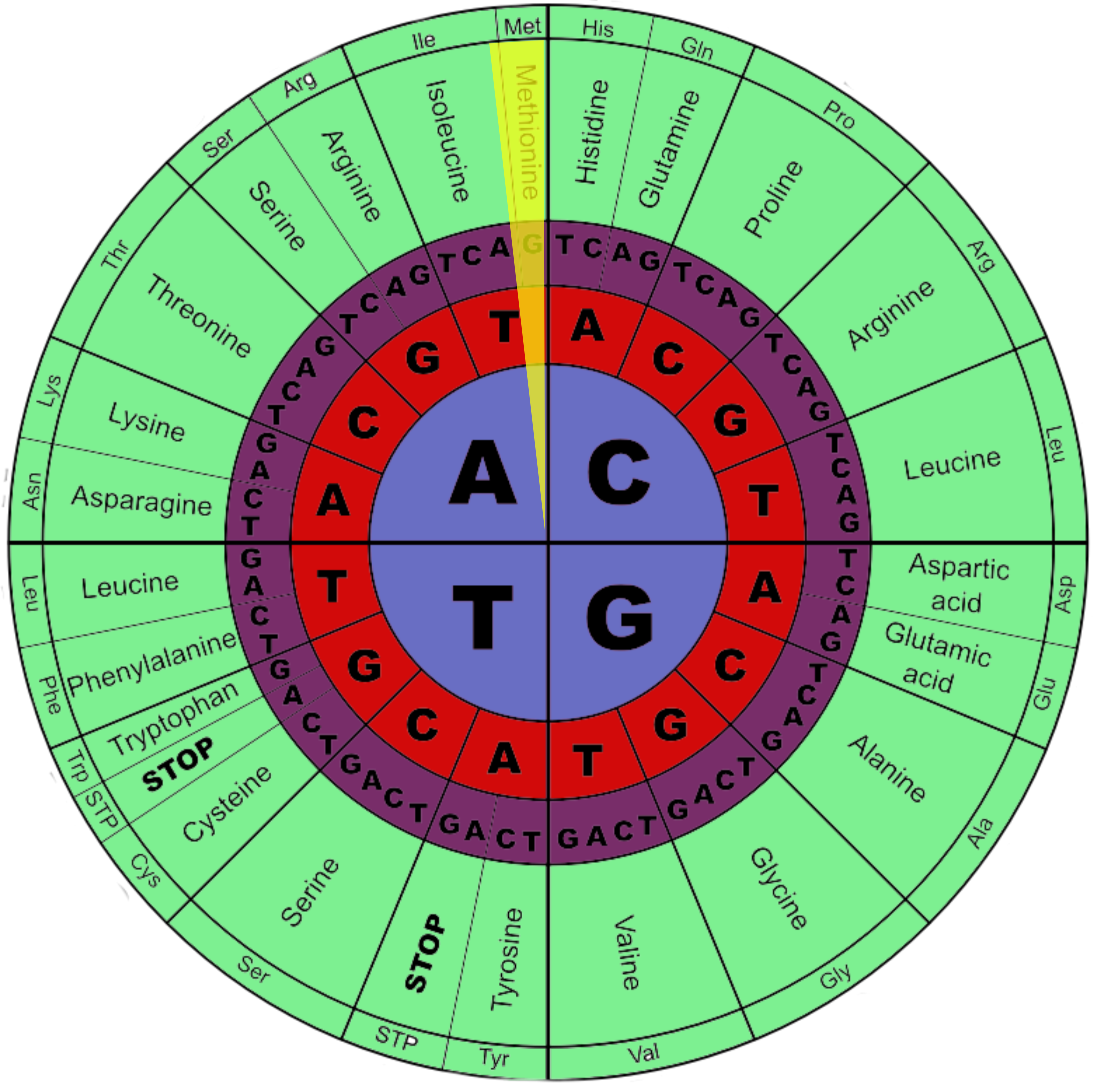
DNA: C A G C

Previous base	next trit to encode: 0	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

2013

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

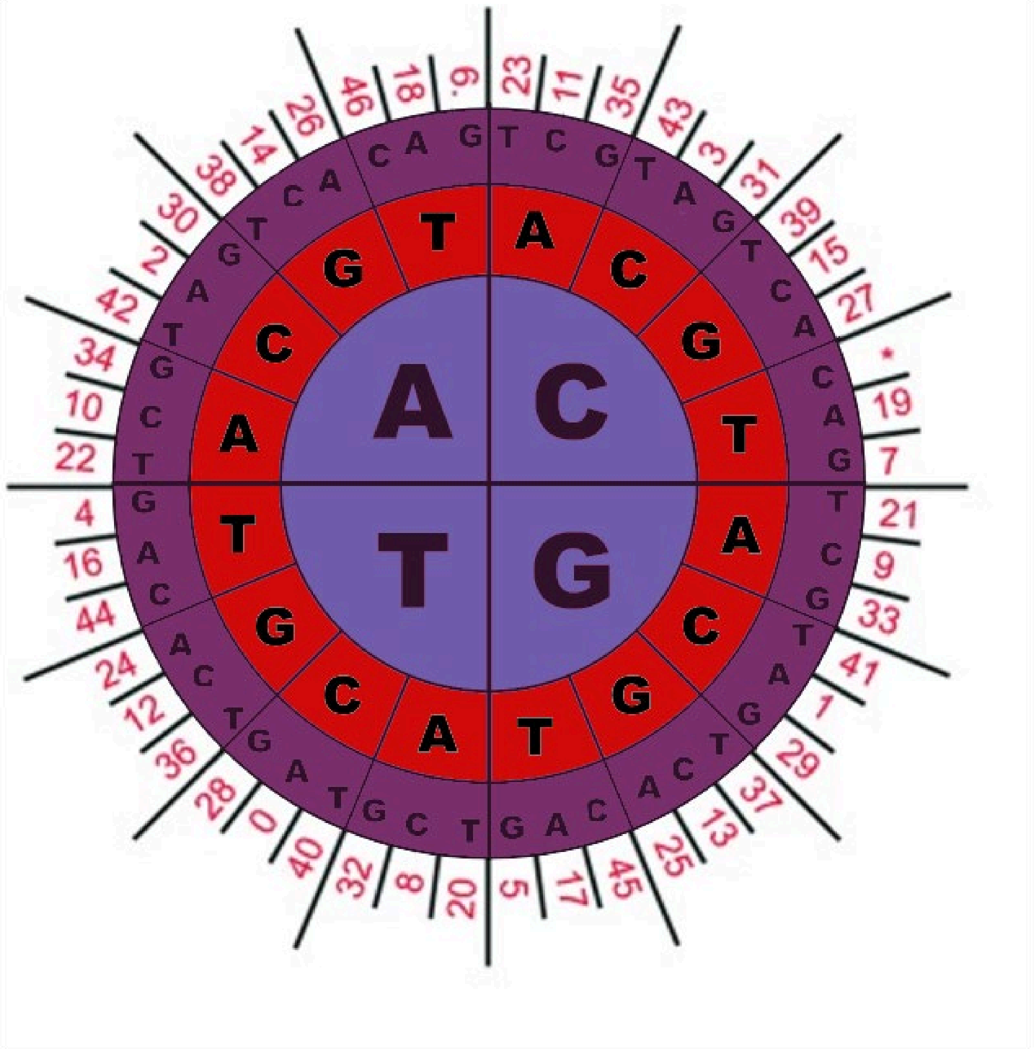
Another biology class, I CANNOT BELIEVE IT!



GENETIC CODE

2015

Let's get natural!



Error detecting and correcting codes.

Reed-Solomon

2015

Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., & Stark, W. J. (2015). Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angewandte Chemie International Edition*, *54*(8), 2552–2555. https://doi.org/10.1002/anie.201411378

Can we access individual files?

Random-access

Many files in the same "tube"

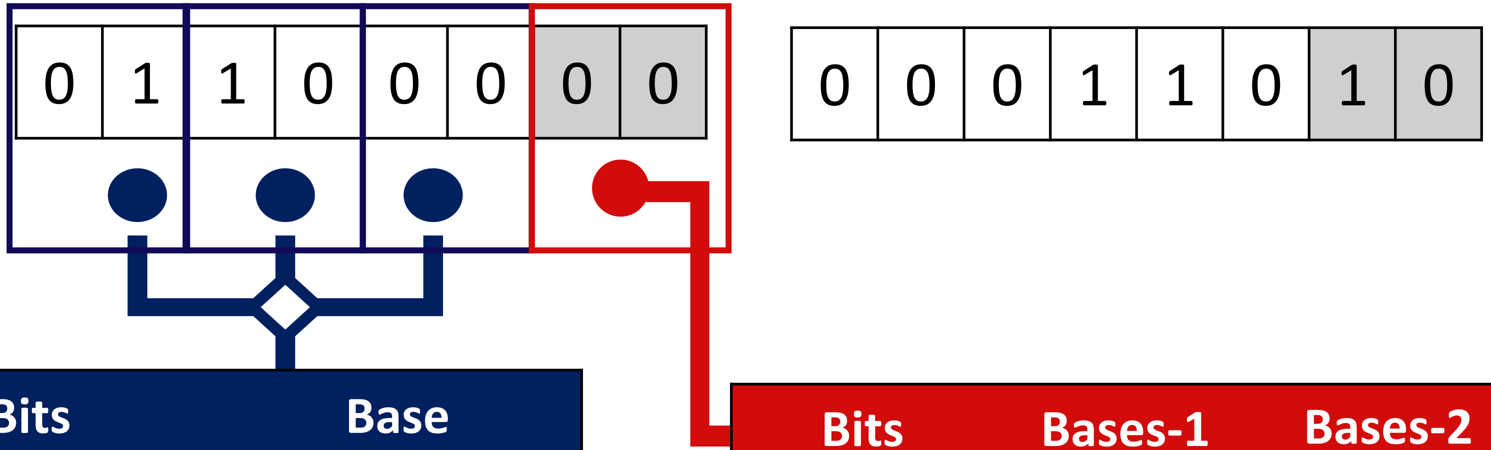
Get only those we want without sequencing

2015

Yazdi, S. M. H. T., Kiah, H. M., Garcia-Ruiz, E., Ma, J., Zhao, H., & Milenkovic, O. (2015). DNA-Based Storage: Trends and Methods. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, *1*(3), 230–248. https://doi.org/10.1109/tmbmc.2016.2537305

STORAGE DEVELOPER CONFERENCE

Run Length Limit



Bits	Base
00	A
01	C
10	G
11	T

Bits	Bases-1	Bases-2	Bases-3	Bases-4
00	AA	CC	GG	TT
01	AC	CG	GT	TA
10	AG	CT	GA	TC
11	AT	CA	GC	TG

2016

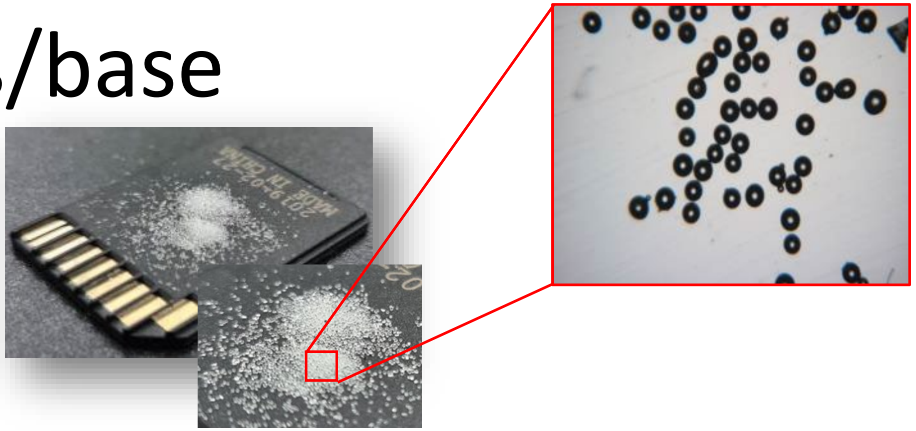
Blawat, M., Gaedke, K., Hütter, I., Chen, X. M., Turczyk, B., Inverso, S., Pruitt, B. W., & Church, G. M. (2016). Forward Error Correction for DNA Data Storage. *Procedia Computer Science*, *80*, 1011–1022. https://doi.org/10.1016/j.procs.2016.05.398

Data density

Bits	Base
00	A
01	C
10	G
11	T

$\log2(4) = 2$

2 bits/base



2017

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

Actual physical storage capacity

5 years

0.001 Exabytes/ gram
2012

210x denser

0.21 Exabytes/ gram
2017

2017

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

Hundreds of megabytes

200 MB

35 files

Sequence clustering!

2018

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

Lenovo | ipt

STORAGE DEVELOPER CONFERENCE SDC 22

Huffman

ABRACADABRA => 11 - 01000001 01000010 01010010 01000001 01000011 01000001 01000100 01000001 01000010 01010010 01000001

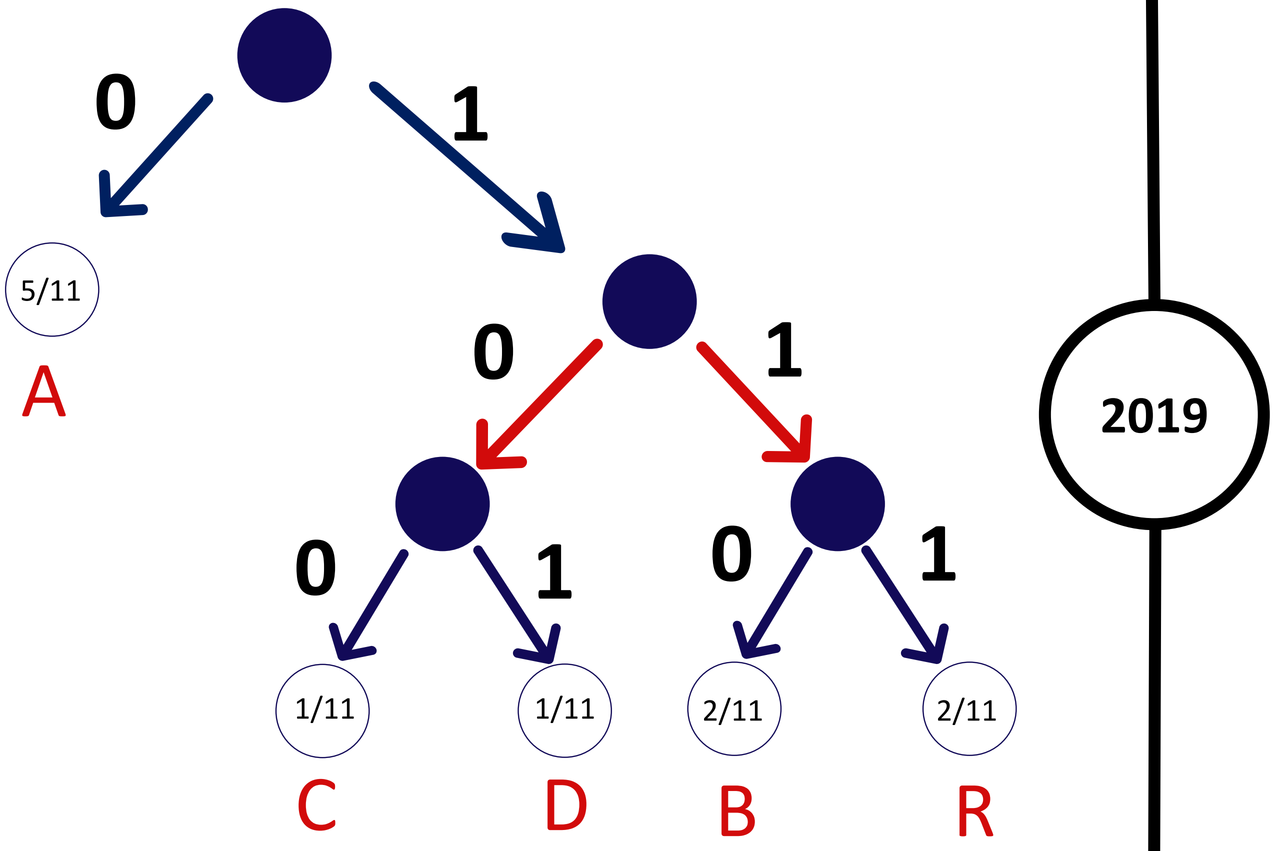
A => 5 – 01000001 – 0

B => 2 – 01000010 – 110

R => 2 – 01010010 – 111

C => 1 – 01000011 – 100

D => 1 – 01000100 – 101

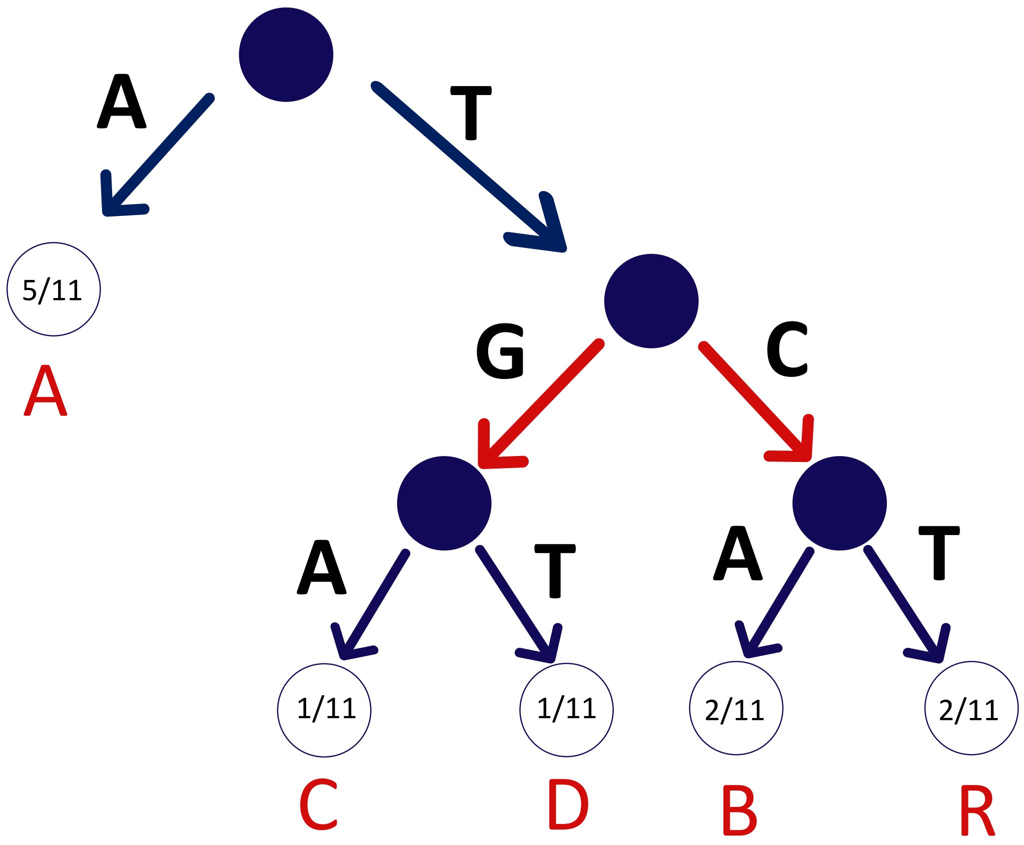


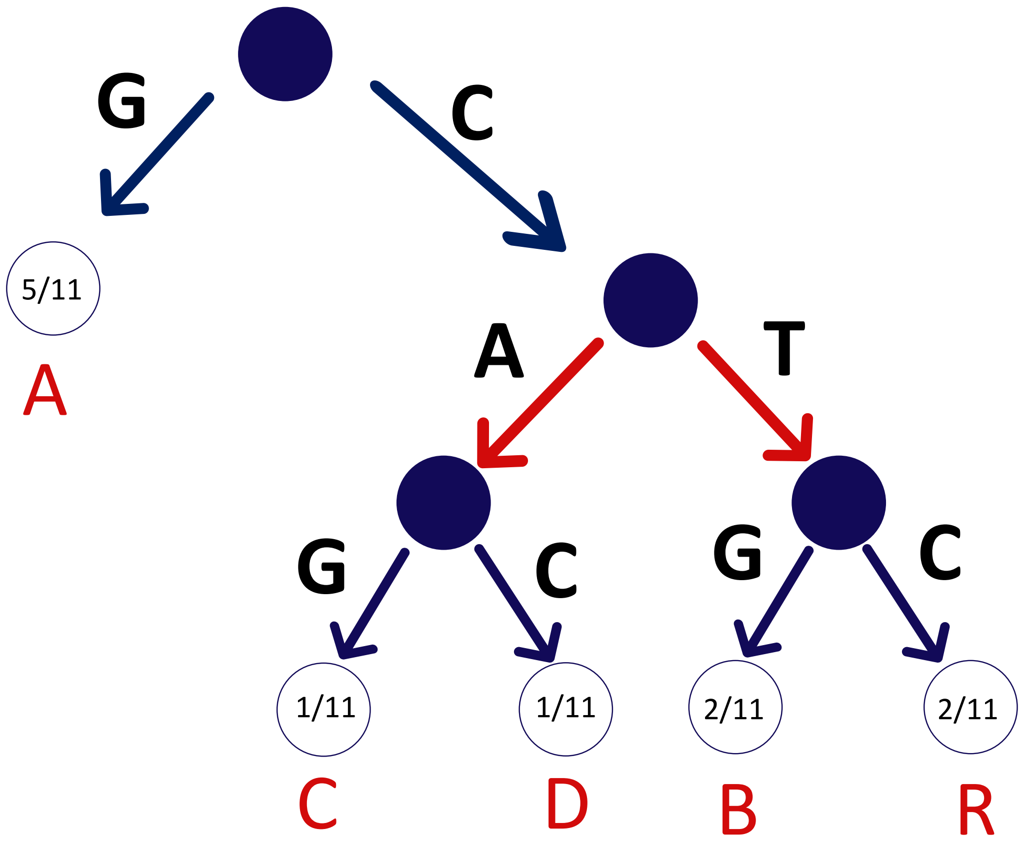
Mishra, P., Bhaya, C., Pal, A. K., & Singh, A. K. (2020). Compressed DNA Coding Using Minimum Variance Huffman Tree. In IEEE Communications Letters (Vol. 24, Issue 8, pp. 1602–1606). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/lcomm.2020.2991461

Lenovo | ipt

STORAGE DEVELOPER CONFERENCE SDC 22

Huffman





2019

Mishra, P., Bhaya, C., Pal, A. K., & Singh, A. K. (2020). Compressed DNA Coding Using Minimum Variance Huffman Tree. In IEEE Communications Letters (Vol. 24, Issue 8, pp. 1602–1606). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/lcomm.2020.2991461

Huffman

ABRACADABRA => 11 - 01000001 01000010 01010010 01000001 01000011 01000001 01000100 01000001 01000010 01010010 01000001 – A CTG TCT G TGA G TGT G TCA CTC A

A => 5 – 01000001 – A – G

B => 2 – 01000010 – TCA – CTG

R => 2 – 01010010 – TCT – CTC

C => 1 – 01000011 – TGA – CAG

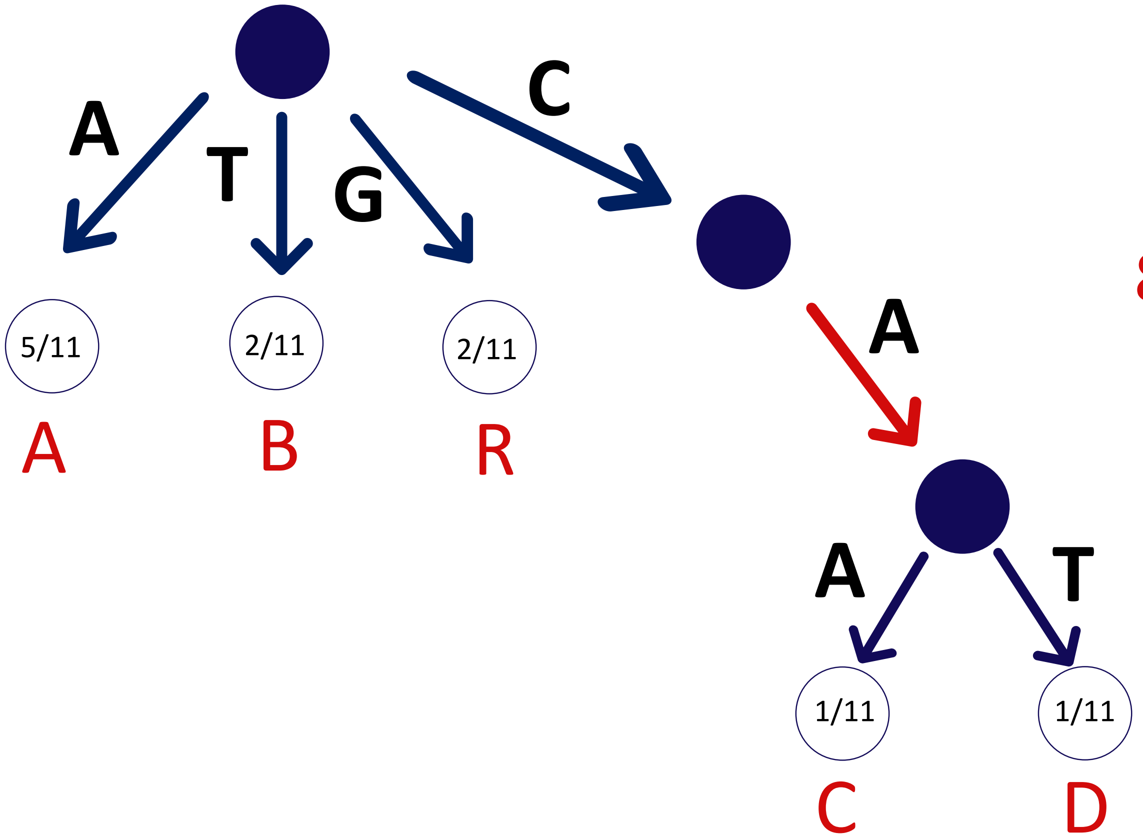
D => 1 – 01000100 – TGT – CAC

88 Bits/23 bases = 3,82 bits/base

2019

Mishra, P., Bhaya, C., Pal, A. K., & Singh, A. K. (2020). Compressed DNA Coding Using Minimum Variance Huffman Tree. In IEEE Communications Letters (Vol. 24, Issue 8, pp. 1602–1606). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/lcomm.2020.2991461

Huffman



ABRACADABRA => 88 bits

A T G A CAA A CAT A T G A – 15 bases

88 Bits/15 bases = 5,82 bits/base

2019

Zhang, S., Huang, B., Song, X., Zhang, T., Wang, H., & Liu, Y. (2019). A high storage density strategy for digital information based on synthetic DNA. In 3 Biotech (Vol. 9, Issue 9). Springer Science and Business Media LLC. https://doi.org/10.1007/s13205-019-1868-4

"Expand" the alphabet

symbol	base mix
R	A,G
Y	C,T
M	A,C
K	G,T
S	C,G
W	A,T
H	A,C,T
B	C,G,T
V	A,C,G
D	A,G,T
N	A,C,G,T

15 pseudo-bases!

2019

Choi, Y., Ryu, T., Lee, A. C., Choi, H., Lee, H., Park, J., Song, S. H., Kim, S., Kim, H., Park, W., & Kwon, S. (2019). High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-43105-w

"Expand" the alphabet

Bits	Base
00	A
01	C
10	G
11	T

log2(4) = 2

2 bits/base

Base	
A	R
C	Y
G	M
T	K
	S
	W
	H
	B
	V
	D
	N

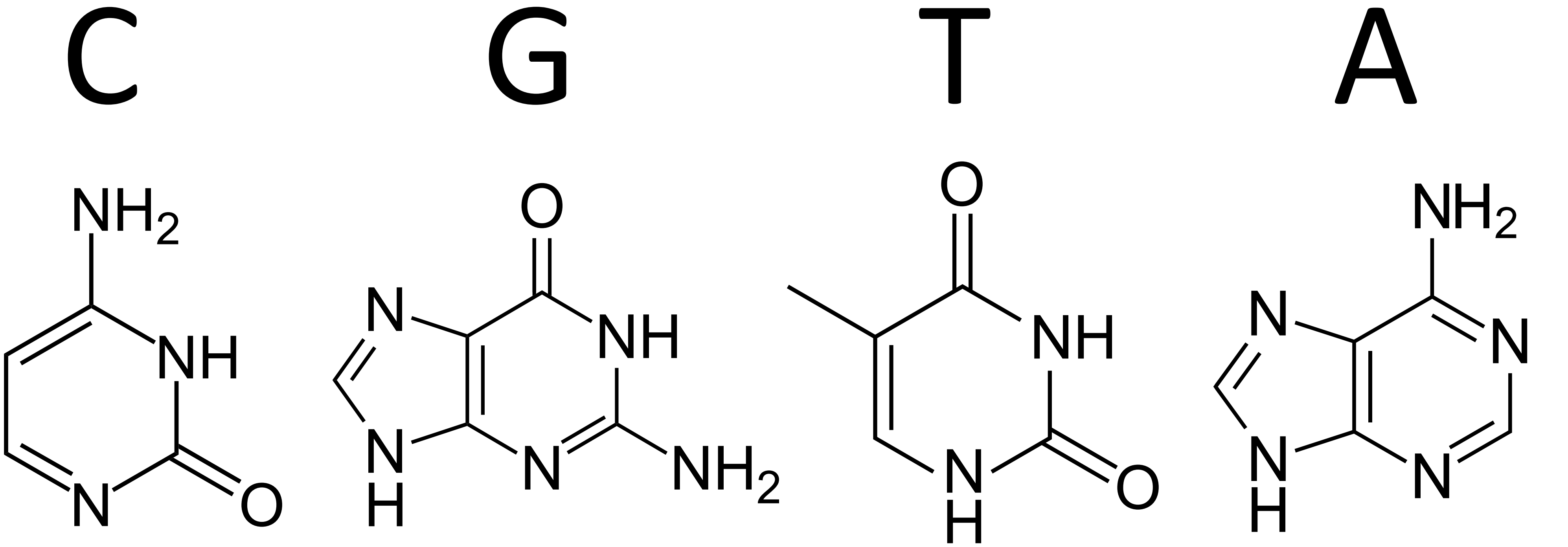
log2(15) = 3.9

3.9 bits/base

2019

Choi, Y., Ryu, T., Lee, A. C., Choi, H., Lee, H., Park, J., Song, S. H., Kim, S., Kim, H., Park, W., & Kwon, S. (2019). High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-43105-w

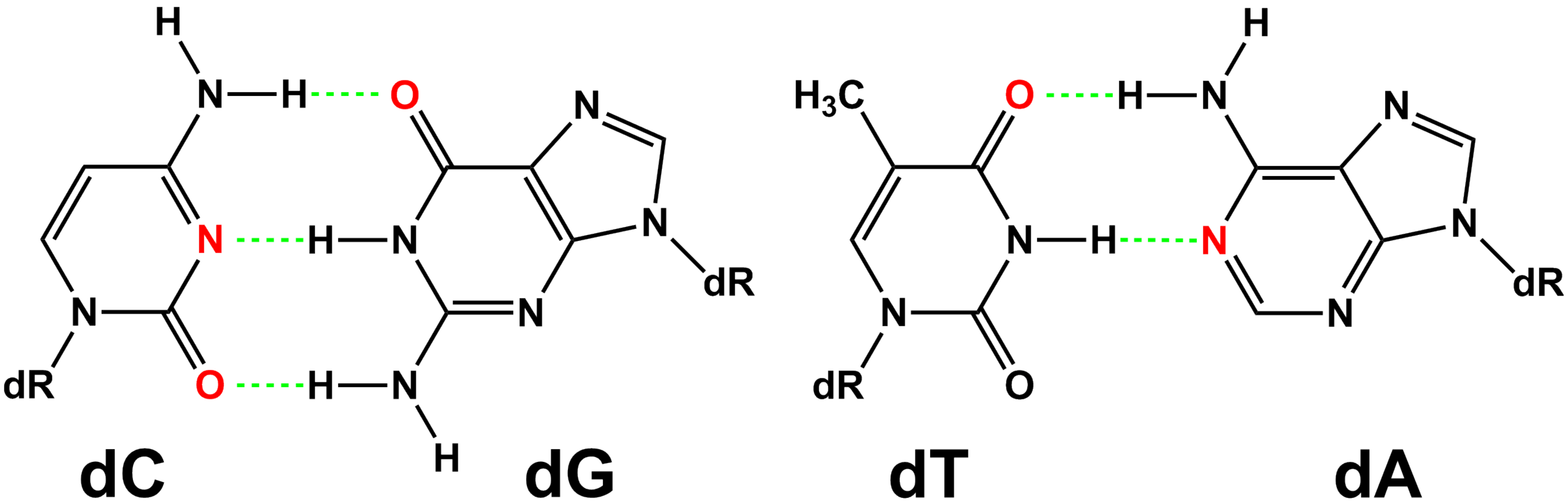
"Expand" the alphabet



2020

Biswas, S., Nath, S., Sing, J. K., & Sarkar, S. K. (2020). Extended nucleic acid memory as the future of data storage technology. *International Journal of Nano and Biomaterials*, *9*(1/2), 2. https://doi.org/10.1504/ijnbm.2020.107412

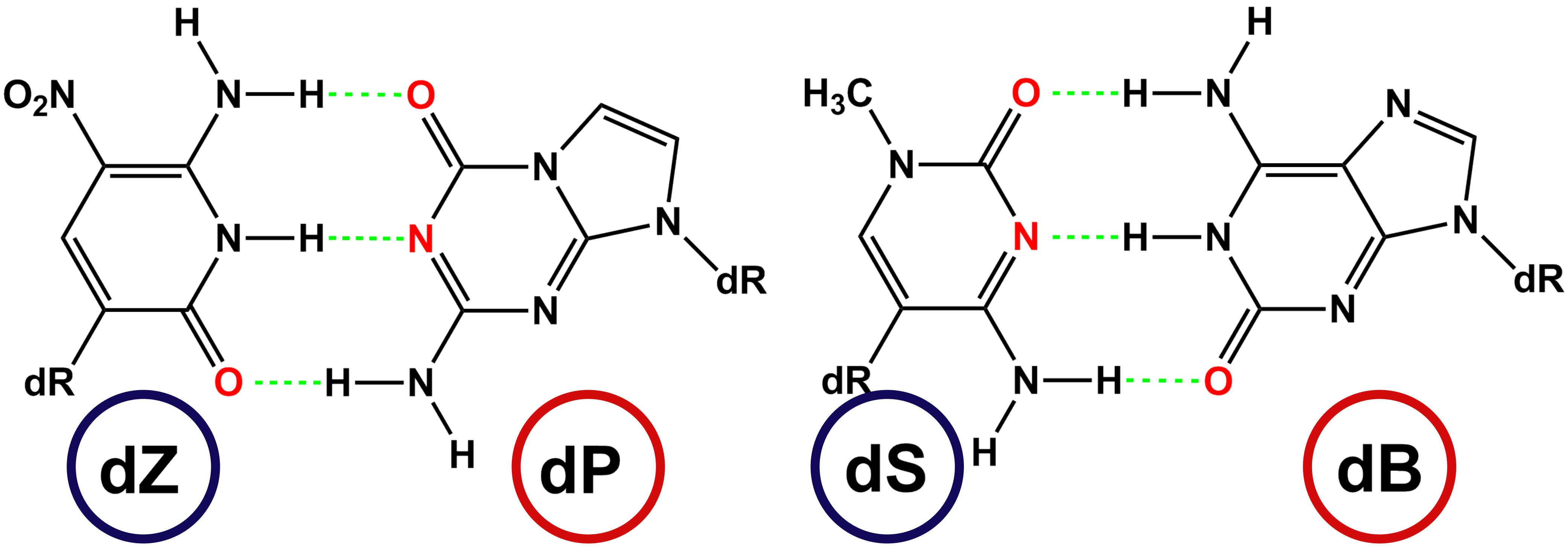
“Expand” the alphabet

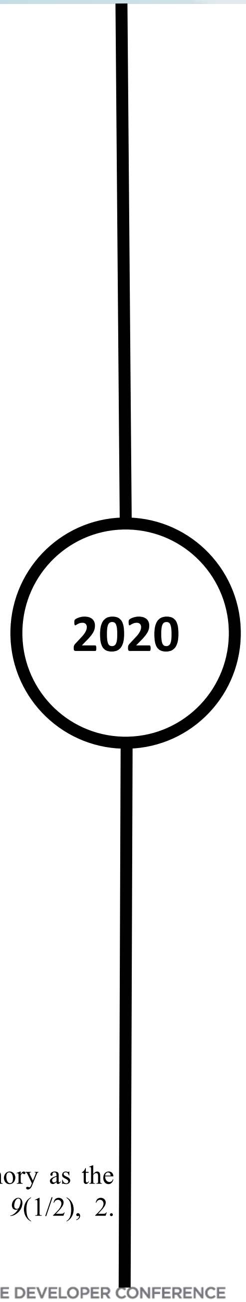


2020

Biswas, S., Nath, S., Sing, J. K., & Sarkar, S. K. (2020). Extended nucleic acid memory as the future of data storage technology. *International Journal of Nano and Biomaterials*, *9*(1/2), 2. https://doi.org/10.1504/ijnbm.2020.107412

Expand the alphabet





Biswas, S., Nath, S., Sing, J. K., & Sarkar, S. K. (2020). Extended nucleic acid memory as the future of data storage technology. *International Journal of Nano and Biomaterials*, *9*(1/2), 2. https://doi.org/10.1504/ijnbm.2020.107412

Expand the alphabet

Bits	Base
000	A
001	C
010	G
011	T
100	Z
101	S
110	P
111	B

log2(8) = 3

3 bits/base

What if...

2020

Biswas, S., Nath, S., Sing, J. K., & Sarkar, S. K. (2020). Extended nucleic acid memory as the future of data storage technology. *International Journal of Nano and Biomaterials*, *9*(1/2), 2. https://doi.org/10.1504/ijnbm.2020.107412

Lenovo | ipt

STORAGE DEVELOPER CONFERENCE SDC 22

Expand the alphabet

Bits	Base
000	A
001	C
010	G
011	T
100	Z
101	S
110	P
111	B

log2(8) = 3

3 bits/base

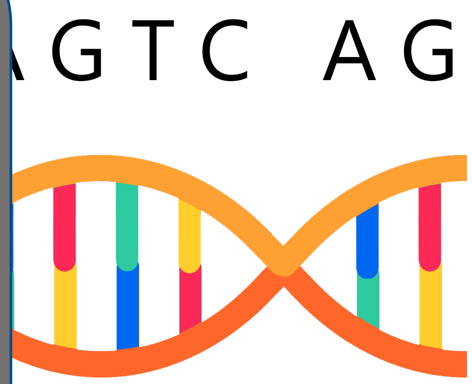
7.99 bits/base

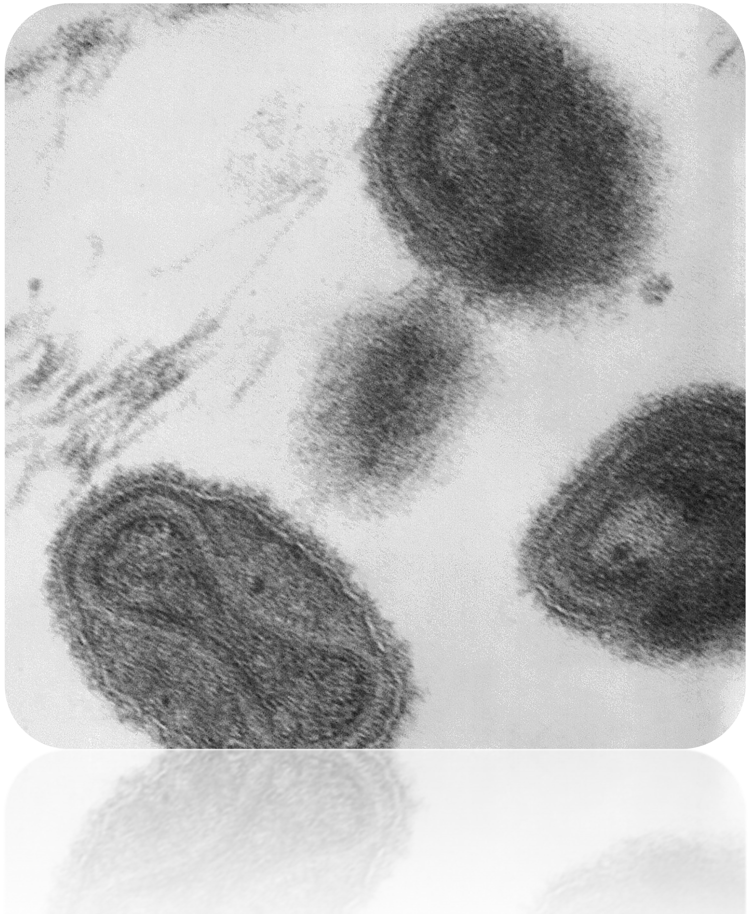
2020

STORAGE DEVELOPER CONFERENCE SDC22

Every challenge can be overcome

- Avoid homopolymers
- Control GC content
- Escape from unwanted sequences





2020

Liu, Q., Wang, P., Cui, J., & Qi, H. (2020). MRC: A High Density Encoding Method for Pratical DNA-based Storage. *2020 Eighth International Conference on Advanced Cloud and Big Data (CBD)*. https://doi.org/10.1109/cbd51900.2020.00012

Data inception

The message behind the message behind the DNA molecule.

2022

Ping, Z., Chen, S., Zhou, G., Huang, X., Zhu, S. J., Zhang, H., Lee, H. H., Lan, Z., Cui, J., Chen, T., Zhang, W., Yang, H., Xu, X., Church, G. M., & Shen, Y. (2022). Towards practical and robust DNA-based data archiving using the yin–yang codec system. *Nature Computational Science*, *2*(4), 234–242. https://doi.org/10.1038/s43588-022-00231-2

Data inception

0.21 Exabytes/ gram
2017

5 years

2058x denser

432.2 Exabytes/ gram
2022

2022

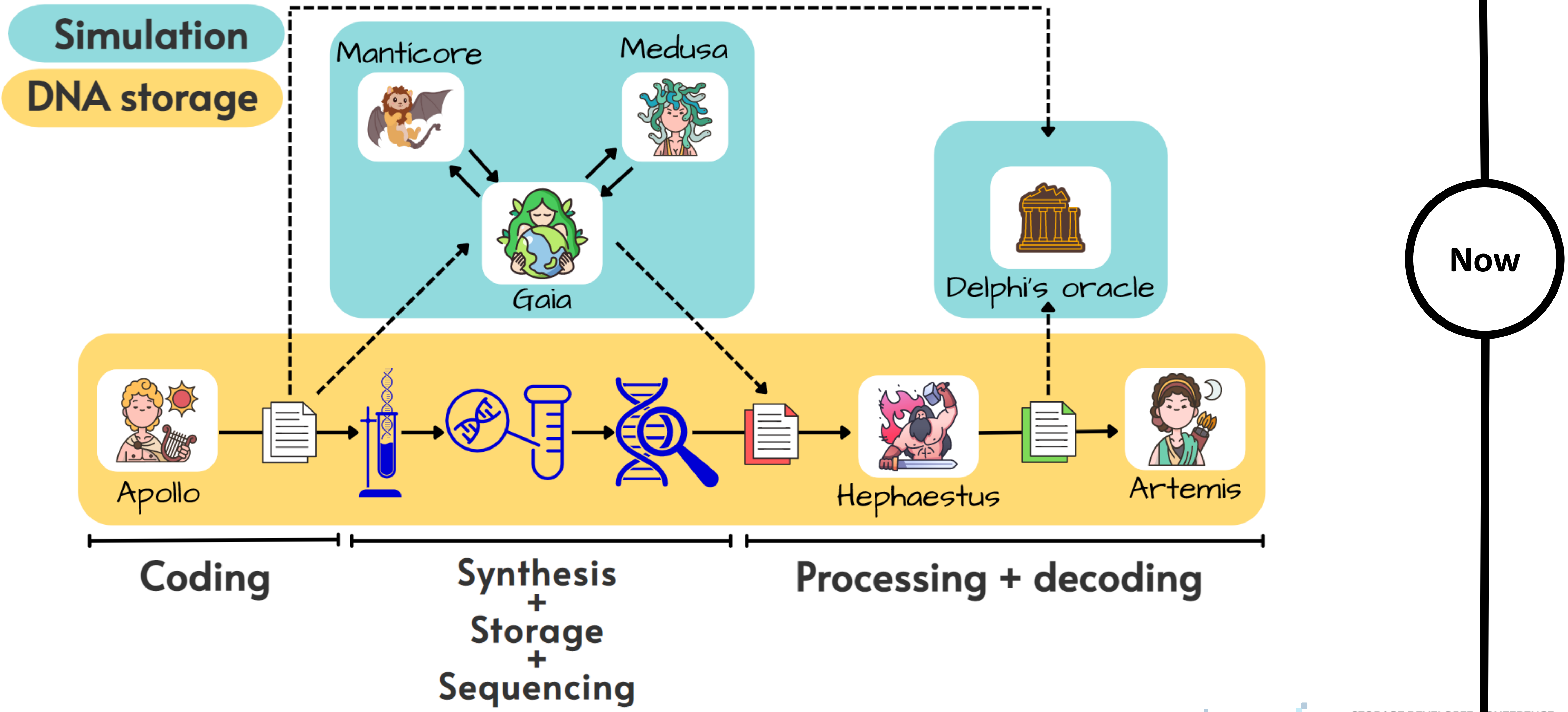
Ping, Z., Chen, S., Zhou, G., Huang, X., Zhu, S. J., Zhang, H., Lee, H. H., Lan, Z., Cui, J., Chen, T., Zhang, W., Yang, H., Xu, X., Church, G. M., & Shen, Y. (2022). Towards practical and robust DNA-based data archiving using the yin–yang codec system. *Nature Computational Science*, *2*(4), 234–242. https://doi.org/10.1038/s43588-022-00231-2

What about us?

PROMETHEUS

Now

What about us?

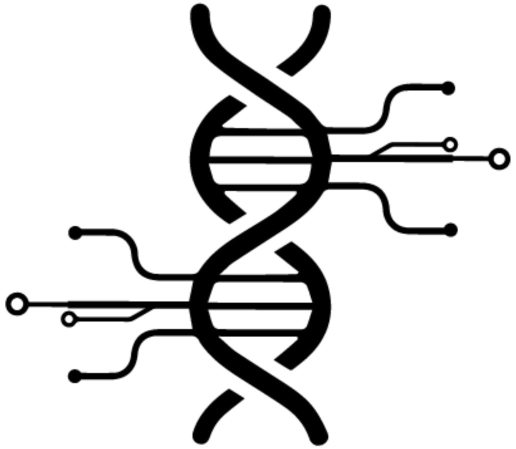


How far will we go?!

Working together we go further!

Future

JPEG



DNAEG?

How far will we go?!

Efforts to create industry standards should be a priority. It is the way to shape a reliable staple solution.

Future

Lenovo | ipt

STORAGE DEVELOPER CONFERENCE SDC 22

Thank you!

Jay Gervasio (joaodiniz@ipt.br)

Adriano Galindo Leal, PhD, EE (leal@ipt.br)

Data density

```
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
```

symbol	base mix
R	A,G
Y	C,T
M	A,C
K	G,T
S	C,G
W	A,T
H	A,C,T
B	C,G,T
V	A,C,G
D	A,G,T
N	A,C,G,T

2019

Choi, Y., Ryu, T., Lee, A. C., Choi, H., Lee, H., Park, J., Song, S. H., Kim, S., Kim, H., Park, W., & Kwon, S. (2019). High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-43105-w

STORAGE DEVELOPER CONFERENCE

Data density

symbol	base mix
R	A,G
Y	C,T
M	A,C
K	G,T
S	C,G
W	A,T
H	A,C,T
B	C,G,T
V	A,C,G
D	A,G,T
N	A,C,G,T


```
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTAGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
CGACGCACCTGTTTACACGTCGTATGCATAAACGAGCCGCACGAACCAGAGAGCATAAAGAGGACCTCTAGTTCCTTTA
```

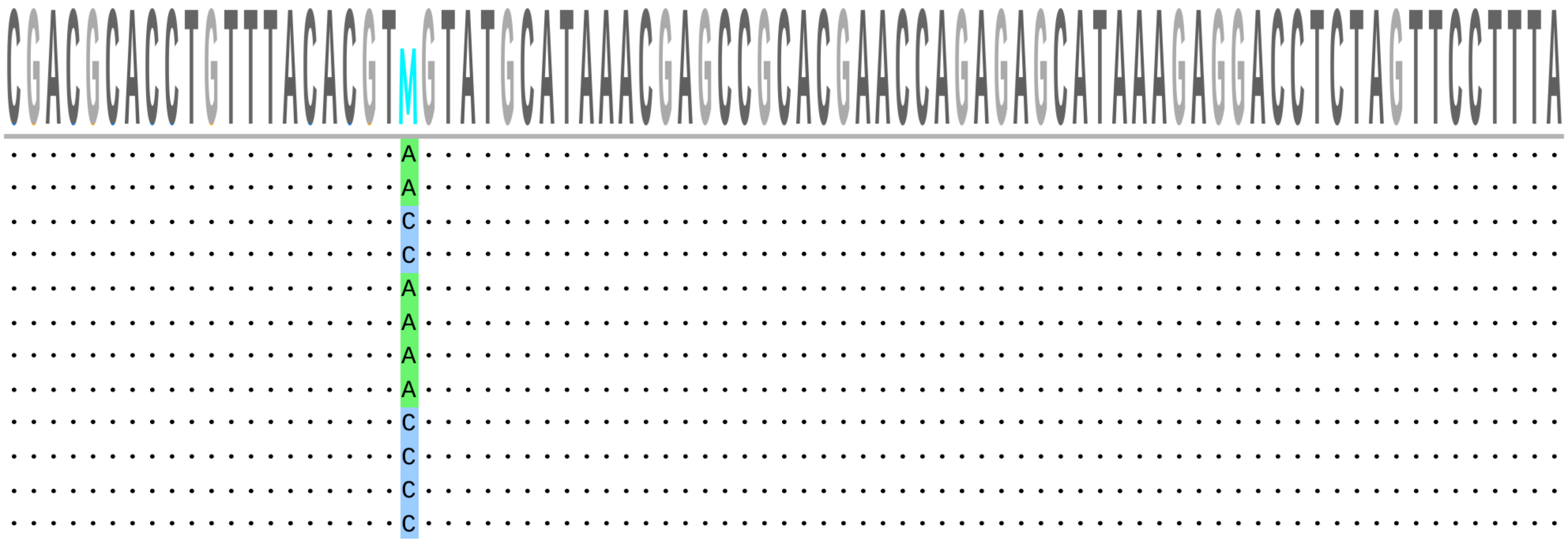
2019

Choi, Y., Ryu, T., Lee, A. C., Choi, H., Lee, H., Park, J., Song, S. H., Kim, S., Kim, H., Park, W., & Kwon, S. (2019). High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-43105-w

Lenovo | ipt

STORAGE DEVELOPER CONFERENCE SDC 22

Data density



symbol	base mix
R	A,G
Y	C,T
M	A,C
K	G,T
S	C,G
W	A,T
H	A,C,T
B	C,G,T
V	A,C,G
D	A,G,T
N	A,C,G,T

2019

Choi, Y., Ryu, T., Lee, A. C., Choi, H., Lee, H., Park, J., Song, S. H., Kim, S., Kim, H., Park, W., & Kwon, S. (2019). High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-43105-w

Another biology class, I CANNOT BELIEVE IT!

SECOND LETTER

FIRST LETTER	T		C		A		G		THIRD LETTER
T	TTT TTC TTA TTG	Phenyl-alanine Leucine	TCT TCC TCA TCG	Serine	TAT TAC TAA TAG	Tyrosine Stop	TGT TGC TGA TGG	Cysteine Stop Tryptophan	T C A G
C	CTT CTC CTA CTG	Leucine	CCT CCC CCA CCG	Proline	CAT CAC CAA CAG	Histidine Glutamine	CGT CGC CGA CGG	Arginine	T C A G
A	ATT ATC ATA ATG	Isoleucine Methionine	ACT ACC ACA ACG	Threonine	AAT AAC AAA AAG	Asparagine Lysine	AGT AGC AGA AGG	Serine Arginine	T C A G
G	GTT GTC GTA GTG	Valine	GCT GCC GCA GCG	Alanine	GAT GAC GAA GAG	Aspartic acid Glutamic acid	GGT GGC GGA GGG	Glycine	T C A G

2015

 |