

11th Annual Conference on Foundations of Nano-Science: Self-Assembled Architectures and Devices (FNANO14)

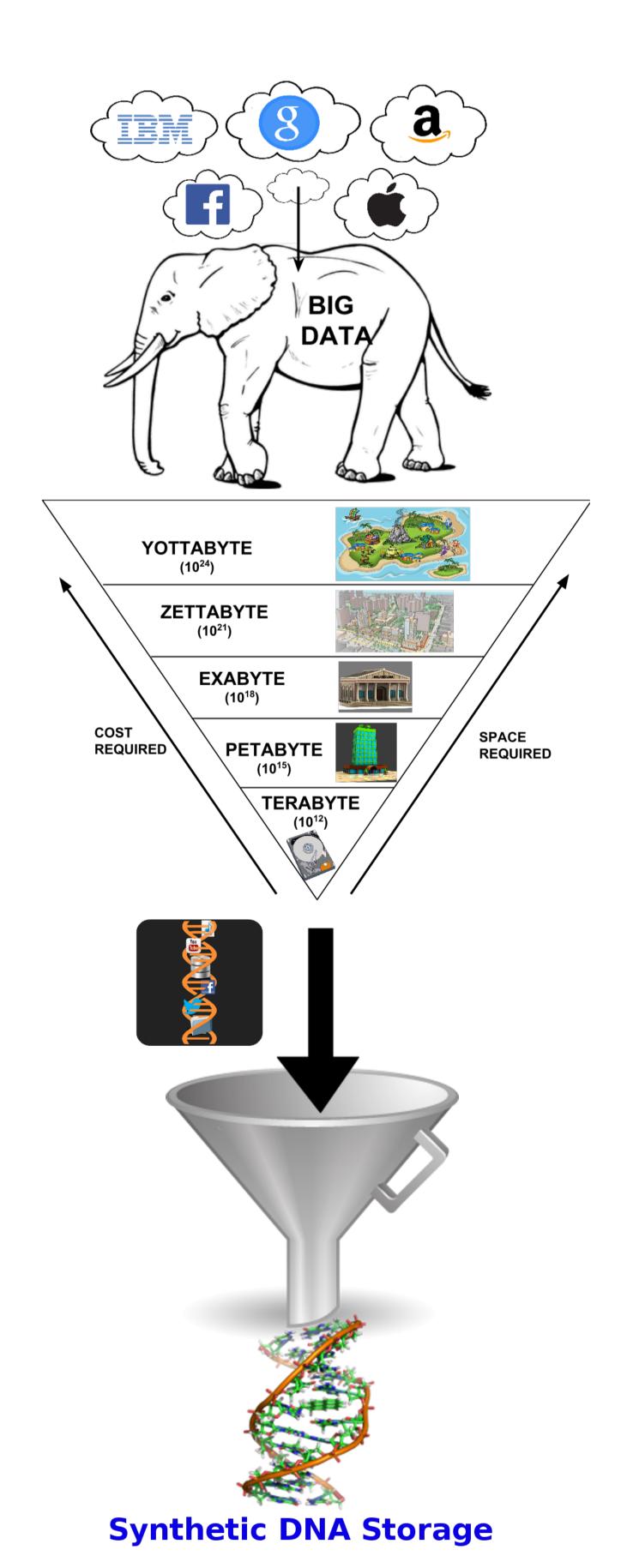
DNACloud: A Potential Tool to Store Big Data on DNA

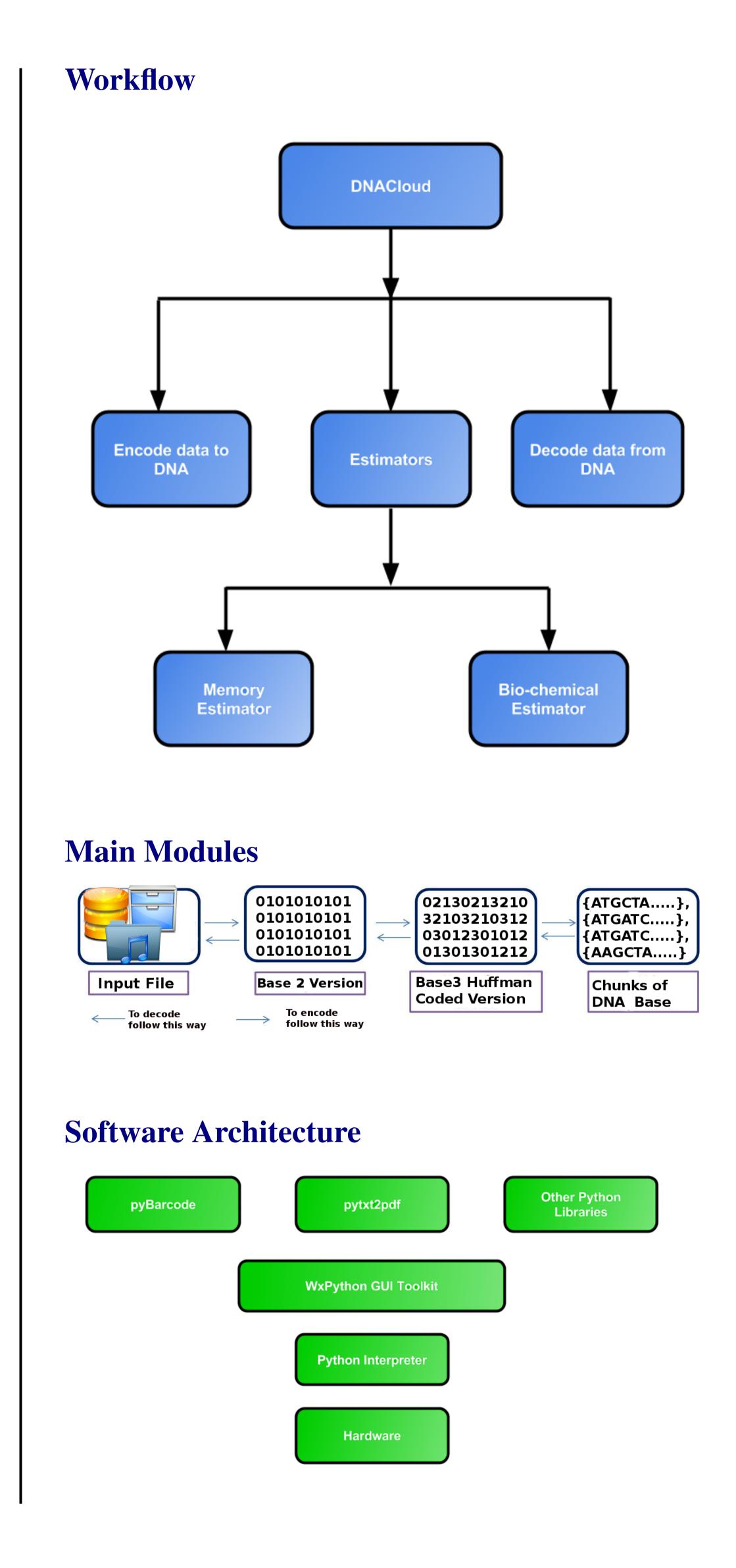
Shalin Shah, Dixita Limbachiya and Manish K. Gupta

Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India.

Abstract

The term Big Data is usually used to describe huge amount of data that is generated by humans from digital media such as cameras, internet, phones, sensors etc. However before one can use the data, one has to address many issues for big data storage. Motivated by Goldman and his team [1], we have developed a software called DNACloud which makes it easy to store the data on the DNA.





Software Output

DNACloud produces three different types of output files apart from .dnac file one of them is the figure . Other files which are generated by the software include output text file of Memory and pdf file generated by Biochemical Estimator.

Output of LATEX file generated using DNACloud

Seq ID	DNA Chunk Sequence		
1	ATCTGCGATA CAGATGAGATCATGTCACTAGCTGCGACAGCACGACTATCATGCGACGACGACGTAGTAGTAGCGATGCGACGACGATGCGACGATGCGATGCGATGCGATGCGATGCGACGTACGT		
2	*TTGATCGACGCTGTGGTGGTGATAGTACGCTGCTGCATCATACGCTACGCTGCGCTGCGTATCATACGCTACGCTAGACGATGGTACGTGAGCAGTACGTAC		
3	'ATGCGACGACGTAGTATGCGATGCGACGCACTAGTAGTATGCGATGCGATGCGATGCGACTCGTGATCGACGCGCGCG		
4	TGGTGCGTATCATCATCATCAGACGTAGACGTTACGCTGAGACATAGCTCGCGCGGGTATATACCTGACGTCACATCACATCACGATCGTATCCTGCGATACGTACG		
K	*ATCTGCTACGATGCGACTCGTGATCGACGCGCGCTATATATGCACTCCACTGTAGTGCTAGCATGCAT		
6	*TTGCGCGCGGATATATACGTGACGTGCACATCACGATCGTATCGTGCGTACGGTTCGGTTCAGGTACAGGACTATACGGTGCTGCTGCTGCTGCTACGTACG		
7	AGTA GTGTAGTGCTAGCATAGCACGCATGTCATGTCATG		
8	*TTA CAGTA CAGTA CAGTA CAGTATA CGCTGAT CTCCTCGTCTA CAGTA CTCACGTGACGTG		
9	'ACTAGACGACGAGCACATGTCATGAGTGCACTGCACTGC		
10	TACGTGACGTGACGTGAGGTCGTATCGTATCGTATCGCGATGACGCGTACGTGATGATGATGTACAGTACAGTACAGCACTGTACAGTACAGTACAGTACAGTTACAGCTACGGTACAGGTACG		
11	AGCATAGCATAGCGCTACTGCGCGCATGCACTACTACATGTCATGTCATGTCGTGACATGTCATGTCATGTCATGTCATGTCATGTCATGTCTGTC		
12	TACGTGATGATGTACAGTACAGTACAGTACAGTACAGTA		
13	ACCTGA CATCTCATCTCATCTCATCTCATCTCATCTCAT		
14	TGTA CAGTA CAGTA CAGTA CAGA CAGTA CACTGCATCGCTATCGCTATCTCGTACAGTACA		
15	'ACCTAGOGAGATCGCATAGAGCATGTCATCTCGTGAGCATAGCAT		
16	TGTA CA GCACTCGTA TCCTATCACATA CGTCTGTA CAGTGTACGTATCGTAT		
17	'ATGCAGACATGTCACATAGCATAGCATAGCATAGACTGGTCTAGAGACTCGATGTGACGTTATGCAGACATGTCATATGCATAGCATAGCATAGCTAGC		
18	**TTCGTATCTGACGAGATCTCTGAGCTACACTGCACATACGTCTGTACACTATACGTATCGTATCGTATCGTACCGTACACTACACTACAGAGCGTATCACGTACCGTACGTA		
19	'ATGTGACGTGTATGCAGACATGTCATATGCATAGCATAG		
20	TIA CGTATCGTATCGTATCGTATCGTACAGTACAGTACAGT		
21	AGTCATGTCATGTCTCGCATAGTGCATGTCATGTCATGT		
22	TAGA TACGCTACACGTACACTACAGTACAGTACAGTACA		
23	*ATGTCATATCATGTCATGTCATGTCATGTCATGTCATGT		
24	TACAGTACAGTACAGTACAGTATGACGATCGTGCGCAGGACAGCACAGCTGCGCACCTACAGCGTCGTACAGACGTGTACAGCTGTATCGCTATCAGCGTACGCTACACACAC		
25	A CTAGCACGCGTCGTCGA CGACGCGTGCATGTCGCAGCATGTCTCGCAGATGTCTAGCATGTCGACATAGCATAGCTGCACTGTCTACGACGTACGT		
26	TCACCTACA GCCTCCTACA GACCTCTACA GATCCTACGCTATCCTATC		
27	ATTETA GCATGCG ACATGCG ACTA GCATA CGACTGCTCTA CCACGATGCACTACTA CGACTGCACTACTA TA CTACACACCTACTA GCATTGCACTCGTCTA GCCTACCTA GCATACCTA GCCATACTA GCCTACCTA GCCTACCTA GCCATACCTA GCATACCTA GCATACCTACATAC		
28	TTCGACGTGACAGATGCTGCTATCTACGCTACGTACGTATGATGTGTGCATCATCGATACGTACG		
29 30	*AGCGATCGATGGACTACTATACTACACACCTAGTAGCTAGC		
31	'ACTAGGCTACGAGGTAGTAGAGTGTAGTACGACGCTGTATACTACGTCTGTAGTACACGAGTACTACACGAGGTACTACTACTACGATCACGAGTACGTAGTACGATCACGTAGTACGATCACGTAGTACGATCACGTAGTACGATCACGTAGTACGTAGTACGATCACGTAGTACACGTACGT		
32	TTCATGCTGCACATATGATGCAGCACATGTGCTGCATCATGTGTAGTA		
33	ATACAGGACGTAGTACACATCATCACAGATA CTATACTATCGCATGCGATGCG		
34	TGTCTATGATATGATACGCTACGCTACGCTACGCTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTACGTCG		
35	AGGGACATCA CATCA C		
36	*TTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTA		
37	A ATCACATCA CATCACATCACATCA CATCACATCA CATCACATCATCA CATCACATCA CATCACATCA CATCACATCA CATCACATCA CATCACATCA CATCACATCA CATCACATCA CA		
38	**TTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGT		
39	A ATCACATCA CATCACATCACATCA CATCA CATCACATCACATCACATCA CATCACATCA CATCA CATCACATCA CATCACATCA CATCATCTCGTAGTGCATA GCTGTA CGTACGAGTG		
40	TEAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAGTGTAG		
41	A ATCACATCA CATCACATCACATCA CATCA CATCACATCTGCTA GTGCATA GCACAGCATA GTATCTCTGGGA GCACATA CATCTA GCACGCA TA CATCTATGA TA CGTACGTACGTCTCT		
42	TEAGTGTAGTAGAGGATCAGGTATCGTGTGGTATCATAGAGAGGCGTGTGTGT		
43	A ACAGCATA GTATCTCTGCG AGCACATACATGTAGCACGCATACATGTATGCGTGCACGCATACATGTATGCGTGCACGCATACATGTATGCGTGCACGCAGCGTACGTA		
44	TATCTA CATCCTCCTATCTA CATACCCA CCTCCCTATCTACATA CCCACCTCCCTATCTACATA CCCACCTCCCTATCTACATACCCACCTCCCTATCTACATACCCACCTCCCTATCTACATACCCACCTCCCTATCTACATACCCACCTCCCTATCTACATACCCACCTCCCTATCTACATACCTAC		
45	ACCTCCACGCATACATCTATGCCTGCACGCATACATCTATGCGTGCACGCATACATCTATGCGTGCACGCATACATCTATGCGTACGCTGCACGCATACATCTATGATACGTACG		
46	TTGCGTATGTA CATACGCACGTGCGTATGTA CATACGCACGTGCGTATGTACATA CGCACGTGCGTATGTACATA CGCACGTGCGTATGTACATACGCACGATACGTACGTCACC		
47	*ATACATGTATGCGTGCACGCATACATGTATGCGTGCACGCATACATGTATCGCTGCACGCATACATGTATGCGTGCACGCATACATGTATGCGTGCACGCAC		
48	TCATAOGCACGTGCGTATGTACATACGCACGTGCGTATGTACATACGCACGTGCGTATGTACATACGCACGTGCGTATGTACATACGCACGTGCGTATGTACGTAC		
49	*ACGTGCACGCATACATGTATGCGTGCACGCATACATGTATGCGTGCACGCATACATGTATGCGTGCACGCATACATGTATGCGTGCACGATGTATGCGTGCACACAACAACAACAACAACAACAACAACAACAACAACA		
80	TTGCCTATCTACATACGCACCTCCCTATCTACATACCCCACCTCCCTATCTACATACCCCACCTCCT		
81	1ATA CATGTATGCCTGCAGCATA CATGTATGCGTGCACGATGTATGCGTGCACCCATACACACTACACTACACTATACGTGCTGCAGCCTCATACTGATACGTACCTACGTACG		
52	TCATACGCACGTGCTACATACGCACGTGCGTATGTGTGATGTCATGTCATACGCACACACTCGCAGTATGACGCACTATGACGCACGTGCGCGATGTCACGTACGT		
53 54	A ACCCATAC ACA GTACA GTACA GTATG TOTATG COTTGTGA GCCTCATACTGCCTGATA CTGCCTGCA CCCGCTACATA CA CGACCTATCTGTGC ATATGA CGCACTA CCTACCTA CGTACCTACCTACCTACCTACCTACCTACCTACCTACCTA		
134	Continued on next page, generate by DNA-Cloud, http://www.guptalab.org/dnacloud		

Testing

We have performed some encoding and decoding operation on some sample data files, results for which are shown in the table below.

File_Type	File size(Bytes)	DNA required
Text	48680	$1.06 \times 10^{-16} \text{gms}$
Audio	85799	$1.88 \times 10^{-16} \text{ gms}$
Video	33745571	$7.41 \times 10^{-14} \text{ gms}$
Image (HD)	473206	$1.04 \times 10^{-15} \text{ gms}$

Comparison of the file formats encoded by DNACloud. Different file types were encoded and decoded using this tool.

Advantages

DNA data storage is a stellar technology when it comes to data store, some of them are as follows:

- 1. Dense storage medium
- 2. No maintenance, electricity requirement
- 3. Long term data storage
- 4. Portable

Future Challenges

DNA data storage technology has many forthcoming challenges, some of them are as follows:

- 1. Re-writable
- 2. Cost-effective
- 3. Cheap synthesizing and sequencing techniques
- 4. Length of DNA string
- 5. Data security
- 6. Efficient encoding and decoding algorithm

Estimator

Our tool, DNACloud, consists of two storage estimators which are described as below:

- 1. Memory Estimator: This estimator provides approximate values for the amount of physical memory required, amount of secondary memory required, amount of DNA required etc.
- 2. Biochemical Properties Estimator: This module provides the minimum and maximum boiling point among-st all the oligonucleotides residing in the .dnac files.

References

[1] Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M LeProust, Botond Sipos, and Ewan Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 2013.



Web: www.guptalab.org/dnacloud Email: dnacloud@guptalab.org