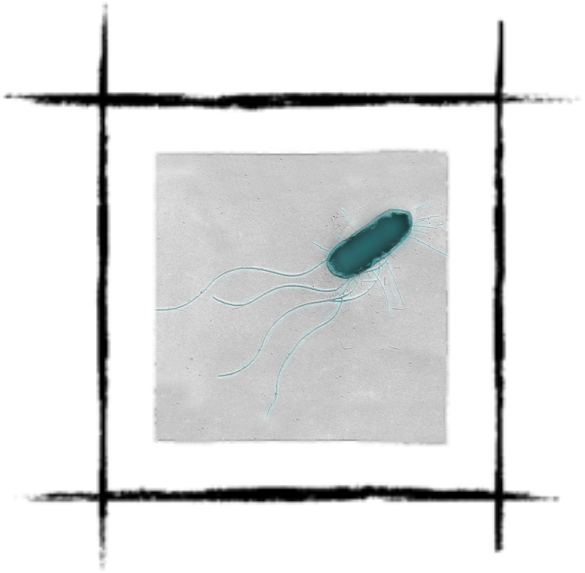
Molecular Storage System(MoSS)

Storing digital data in the world's most resilient molecule - **The DNA**



E Coli :

"DNA is the best molecule that we've worked with, we barely live without it and we use it every day - 5 Stars"

Butterfly :

"We make nano structures that give us amazing colours and textures by following our DNA codex - 10/10 for letting us accessorise"

Chimpanzee :

"It's proven that we're almost human using our DNA - BEST EVIDENCE EVR"

What is MoSS ?

We store digital data in form or electrical/magnetic ‘signatures’ as in the net charge or magnetic potential of a particular location in the substrate, such as silicon - represents either 0 or 1 of the binary system.

Traditionally, the base potential is regarded as 0, and if the potential is turned into a non-zero entity - it is regarded as 1. A simpler way of understanding is that our hard disks have hypothetical tiny light bulbs, if the bulb is on its 1, if it’s off its a 0.

MoSS - Short form for Molecular Storage System, MoSS stores data in the form of a chemical sequence, a DNA sequence. Both 0 and 1 are represented chemically using MoSS and there will be a sequence identifier to denote which direction to read the sequence in.

Lets take an example of a classical byte of information stored on hard disks our computers.

For a 32-bit system:

32 bit address	1Byte	1Byte	1Byte	1Byte

Majority of operating systems have a 32-bit address that act like mail boxes to 32-bits of information (or 4x1Bytes of information), the operating systems uses file systems that acts like a ledgers to these memory addresses. If you want to open a cat video thats on your hard disk, the operating system uses the file system and display the file instantaneously upon your click.

In the same way MoSS also stores data with a 32 bit address and 4 bytes of information, but in a DNA sequence.

An example binary to MoSS representation

01110111 01101111 01110010 01101011	00110001	00110010	00110011	00110100

ATTTATTT ATTATTTT ATTTAATA ATTATATT	AATTAAAT	AATTAATA	AATTAATT	AATTATAA

DNA is a combination of majorly 4 chemicals (**A**denine, **G**uanine, **C**ytosine and **T**hymine) - famously **AGTC**. MoSS only uses 2 nucleotides for representing the binary system (**A=0, T=1** duh!), the **G**uanine and **C**ytosine are used only for stabilising the structure of DNA and for stepwise extension of MoSS, which will be explained in detail in the successive sections of this document.

Structure of MoSS

Bits of MoSS are in a specific arrangement and in designated blocks, the number of blocks of MoSS can be easily calculated with the formula $2*2^n$, where n is number of bits per blocks.

For example if you want to store 4 bits of information per blocks, there will be a total of 32 blocks in the MoSS assembly system. 16 of the would be forward blocks, 16 of them would be reverse blocks.

A 4-bit MoSS system:

AAAA	AAAT	AATA	AATT	ATAA	ATAT	ATTA	ATTT
TAAA	TAAT	TATA	TATT	TTAA	TTAT	TTTA	TTTT

Both forward and reverse blocks contain the same 16 bit combinations. The reverse blocks are the unary operation for the forward blocks. If the bits 0000 are represented as AAAA in the forward blocks, they are represented as TTTT in the reverse blocks.

Each block of MoSS contains three fields, A **bit field** that comprises of **A** and **T** representing **0** and **1** respectively and two **+ and - stability fields** on either terminus of the bit field. The stability fields only comprise of **G** and **C** nucleotides The **forward block stability field is G rich and the reverse block stability field is C rich**.

A visual representation of a 4-bit (0000) MoSS forward block (5' to 3'):

(+) Stability field	Bit field	(-) Stability field
GCGGGGC	**AAAA**	**CGGGGCG**

A visual representation of a 4-bit (0000) MoSS reverse block (3' to 5'):

(+) Stability field	Bit field	(-) Stability field
GCCCCGC	**TTTT**	**CGCCCCG**

Assembly of MoSS

The forward and reverse blocks of MoSS are assembled and extended by hydrogen bonding. Usage of modified nucleotides like LNA's in the stability fields would result in enhanced hydrogen bonding. Various versions of stability fields can be designed as per requirement. An efficient design of the stability fields comprises of primarily 3 design parameters.

1. No or low chance of G-quadruplex formation
2. No or low chance of self ligation
3. Efficient hydrogen bonding to the target

Some examples are given below.

1. High cost efficient design (0000 MoSS forward block - 5' to 3') - 18 bp
   GCGGGGCAAAACGGGGCG
2. Balanced design (0000 MoSS forward block - 5' to 3') - 16 bp
   CGGGGCAAAACGGGGC
3. Low secondary structure design (0000 MoSS forward block - 5' to 3') - 14 bp
   CGGGGAAAAGGGGC
4. No secondary structure design (0000 MoSS forward block - 5' to 3') - 12 bp
   GGGGAAAAGGGG

Modified bases like LNA's can be introduced to increase the stability and decrease the concentration necessary for stable bonding.

An example of LNA included MoSS forward block. (+ prefix denotes an LNA nucleotide)
+G+C+G+G+G+G+CAAAACGGGGCG

An example of all LNA modified MoSS forward block (+ prefix denotes an LNA nucleotide)
+G+G+G+GAAAA+G+G+G+G

The assembly of MoSS blocks would **start with a (-) reverse block stability field** immobilised to a surface and an **alternating forward and reverse blocks** would be assembled over it. The assembly would **end with a (+) forward block stability field**.

An assembly to represent 16 bits of information (01110111 01101111) in MoSS

- +FSF = (+) Forward block stability field
- -FSF = (-) Forward block stability field
- +RSF = (+) Reverse block stability field
- -RSF = (-) Reverse block stability field

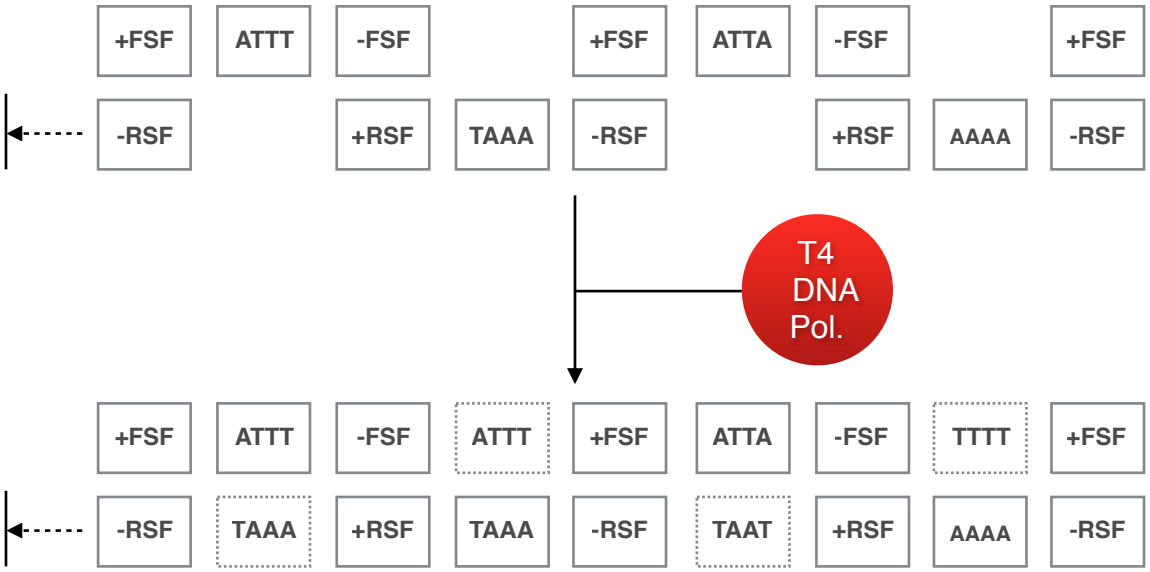
The forward blocks are bound in 5' to 3' orientation and the reverse blocks are bound in 3' to 5' orientation. The initial (-) revere stability field is bound to a immobilised substrate via covalent linkage such as streptavidin and biotin interaction.

Polishing of MoSS

The assembling of MoSS blocks can be extended as per the requirement and the blocks would be polished (gap-filled) after the final assembly step.

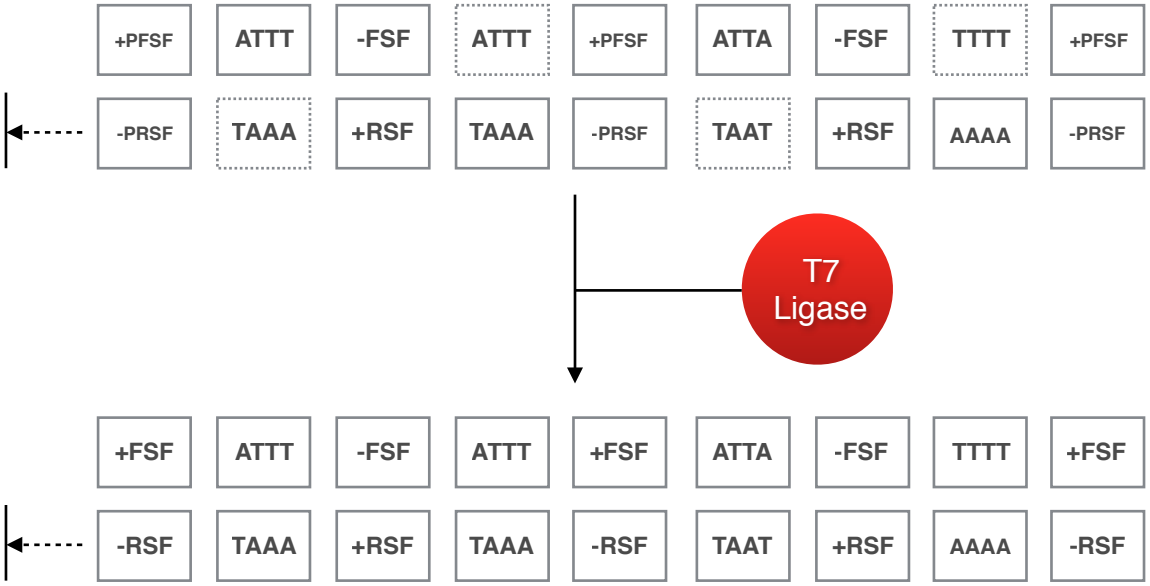
Use of high fidelity gap filling polymerase like T4 DNA polymerase would fill the gaps in the MoSS assembly.

The gap filling step would be as follows.



Post the gap-filling step the extended blocks are needed to be connected with the consecutive block. A **pre 5' phosphorylation step to the (+) FSF and (-) RSF ends** would enable the MoSS blocks to be permanently ligated to each other.

The ligation reaction would be as follows.

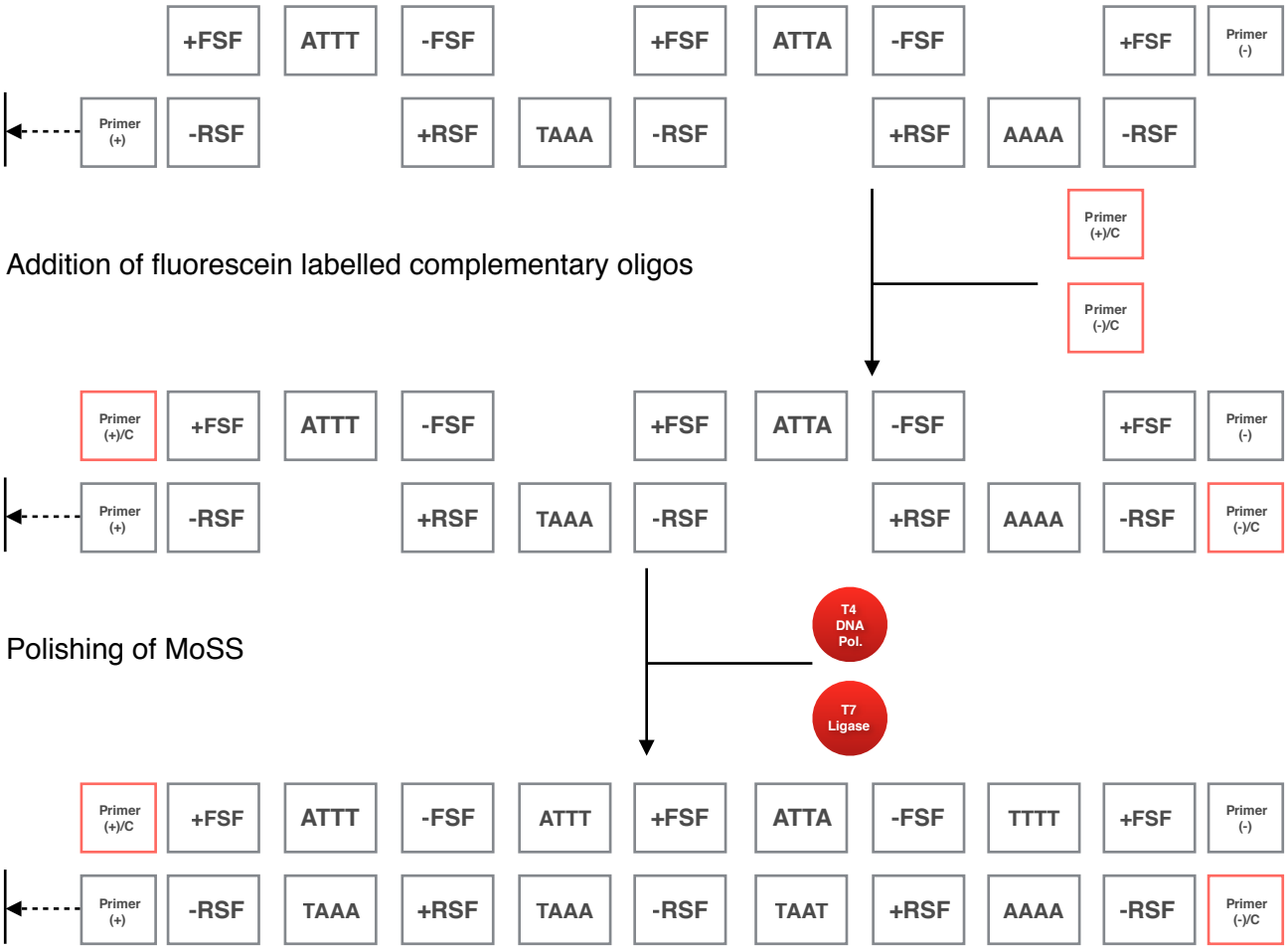


Amplification of MoSS

DNA can be amplified using polymerase chain reaction (PCR), PCR requires primer regions to be present in the sequence which aid in the extension of DNA polymerase for making multiple copies of the template DNA.

Addition of a forward primer region to the initial (-) reverse block stability field and a reverse primer region to the final (+) forward block stability field would introduce priming regions to the MoSS fragments. Multiple priming oligos could be used for using multiple primers or a single set of priming oligos could be used if all the MoSS blocks are to amplified in a single reaction.

MoSS assembly and amplification with priming regions

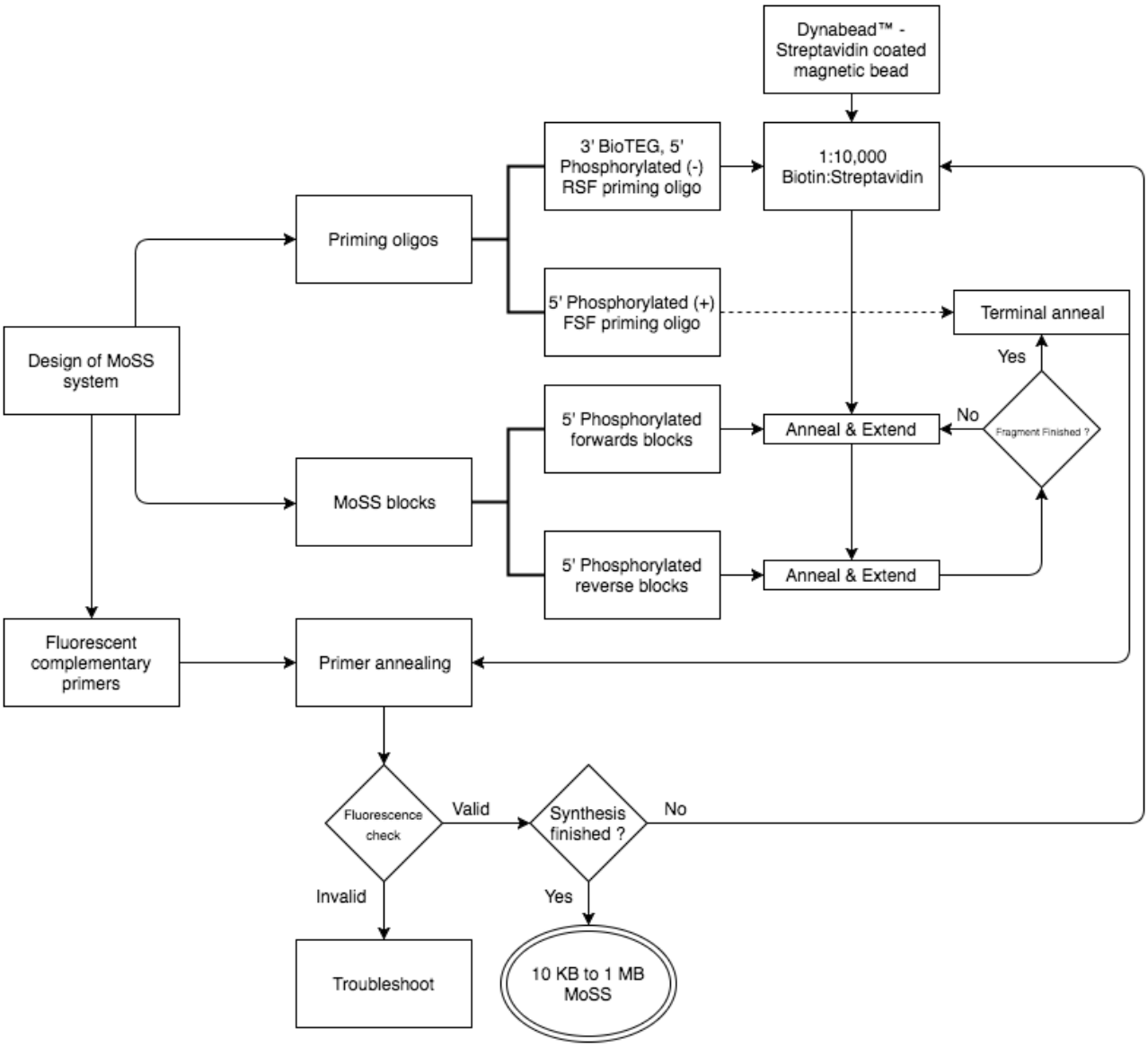


The **fluorescein labelling** of the primer regions would aid in **identifying the stability** of MoSS fragments as well as **validation** of individual MoSS fragment synthesis. As in the emitted florescence would increase linearly with each new MoSS fragment synthesised. The fluorescein labelled oligos could also be used for calibration purposes.

The constructed MoSS fragments could be directly used in a PCR with forward and reverse primers for making multiple copies of the MoSS fragments.

Workflow of MoSS

A workflow of MoSS assembly is visualised below. The capacity of MoSS directly depends on the design of MoSS blocks, total available capacity of streptavidin and amount of biotin anchor per assembly fragment



Experimental preferences for the above workflow.

1. Total streptavidin capacity on 1ul of Dynabead™ = 100 pmole
2. Anchor oligo [(-)RSF priming oligo] per fragment = 10 femtomole
3. Capacity of MoSS fragment = 4 bytes/fragment
4. Total capacity of 1 ul of Dynabead™ = 10,000*4 bytes = 40 KB
5. Preferred MoSS capacity = 4-bit
6. Preferred MoSS design = Balanced design (**CGGGGCXXXXCGGGGC** - 16 bp)
7. Preferred sequencing system = Illumina™ MiSeq (2x250)

Sequencing of MoSS

Sequencing is an integral part of MoSS design as high stability versions of MoSS systems can enable long block extensions which can lead to data storage upto 100 bytes/MoSS fragment. This enables a theoretical maximum capacity of 1 MB MoSS per 100 pmole of streptavidin-biotin anchor.

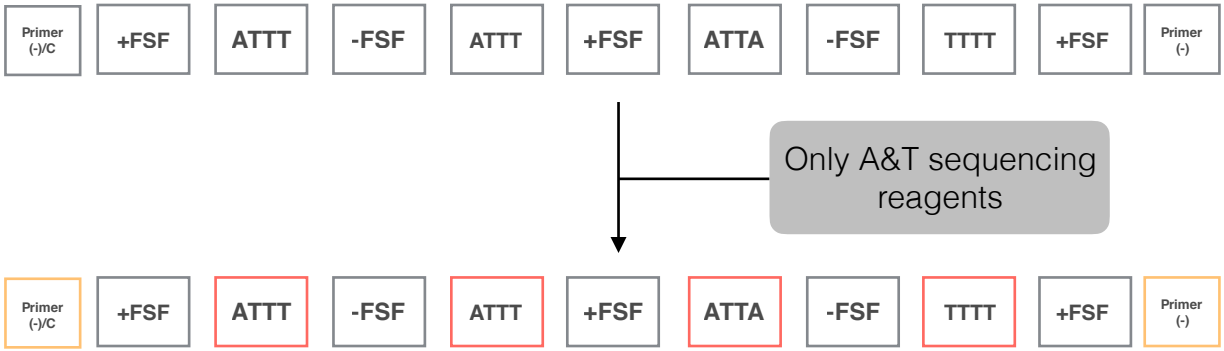
Selection of sequencing systems depends on the use case as some sequencing systems support long fragments with lower coverage, where as other sequencing platforms provide smaller reads with high coverage. Cost of sequencing also has impact on the choice of sequencing platforms.

Amplification would also affect with longer MoSS systems as suitable PCR kits must be selected to amplify long DNA strands as Taq based PCR kits would no longer be ideal.

Sufficient modifications can be made to the sequencing reagents as the data can be read from a MoSS platform using only two nucleotides. **Sequencing only Adenine and Thymine from the MoSS systems would retrieve all the data stored in them**. This could lead to decrease in the overall sequencing costs. Novel sequencing methods can be introduced that can sequence only one basepair where the bit fields of the MoSS blocks contain only a single nucleotide and the data is represented by the number of repetitions of the nucleotide before occurrence of another base.

The use of designated DNA regions as stability fields and bit fields would aid innovative sequencing methods that cater MoSS systems much efficiently in a cost effective manner than general sequencing methods that read all the 4 nucleotides.

Sequencing of a MoSS block



The sequencing with only A and T reagents would partially read the priming regions, which would be sufficient enough as they are predetermined and uniform across MoSS fragments. The bit fields would be completely sequenced and the information archived in the MoSS system would be read by the user. The assembly software would rebuild the reads into binary data and the data would be completely retrieved.

Sequencing systems that are portable, cost efficient and with good long read speeds are ideal for MoSS systems. The sequencing systems would be needed to evolve in a different arc for taking the full potential of the MoSS platform.

Features of MoSS

Traditional methods and experiments surrounding DNA data storage have looked upon how much of the information could be crammed in DNA (via compression). There were several methods evolved on how to avoid repeats and efficiently synthesise DNA cost effectively for data archiving purposes.

MoSS has been engineered to efficiently anneal and make stable DNA data blocks that are extendable and store data in an uncompressed, true-to-binary fashion. This approach has exponentially decreased the cost to store data in DNA. The enzyme requirement has also been lowered which further decreases the total incurred costs.

A table explaining the differences between a MoSS system and traditional DNA data archiving methods.

Criterion	MoSS	Traditional DNA
Cost per byte	0.5 ¢	12.5 ¢
Total number of combinations	16 (4-bit MoSS system)	Millions (Data dependent)
Block assembly	Pipette handling robot (Magnetic pick and place)	None
Sequencing cost	<50% of normal	Normal
Manual oversight	None	High
Automation level	End-to-end automated	None
DNA stability	Relatively high (>60% GC)	Normal
Archival format	True-to-binary	Compressed
***in vivo* activity**	Inert	Normal
***in vivo* replication**	Normal	Normal

A table comparing MoSS with traditional data archiving medium such as magnetic tapes

Criterion	MoSS	Magnetic tapes
Read and write speeds	Very slow	Relatively lightening fast
Areal data density	2.5 Petabits/mm^2	0.1 Gigabits/mm^2
Cost per GB	$ 500,000	1 ¢
Copies per run	100 Billion copies (Typical PCR)	1 Copy
Cost per copy	0.0005 ¢	1 ¢
Cost to make 100 Billion copies	$ 100 (Typical PCR)	-Does not compute-
Radiation shielding	Nominal	High (Random bit decays)
Maintenance cost	Extremely low	High
Maintenance volume	1 cm^3 per Zettabyte (1 Million Petabytes)	10 cm^3 per Gigabyte

Experimental MoSS

An experimental setup to synthesise 40 KB of MoSS is shown in the following table.

Property	Entity	Category
Environmental control	Laminar flow hood	Equipment
Hardware	OpenTrons™ (OT-Hood)	Equipment
Bead transfer	3D printed magnetic pipette tip with removable magnetic flea	Equipment
Sequencing platform	Illumina™ Mi-Seq	Equipment
Bead immobilisation	Dynabeads™ (Streptavidin based)	Reagents
MoSS system	4-bit blocks (LNA™ based)	Reagents
Gap filling polymerase	T4 DNA Polymerase (NEB™)	Reagents
Stick end DNA ligase	T7 DNA ligase (NEB™)	Reagents
PCR polymerase	Taq Pol	Reagents
Bead wash and bind	IDT™ duplex buffer	Consumable

The reaction follows a pick and place strategy of the magnetic flea without any reagent transfer during the experiment. The magnetic beads would be attached to the magnetic flea and the pipette handler would move to the respective reagent well and move the reaction forward. The reaction starts with biotin anchored priming oligos followed by MoSS blocks.

Individual MoSS fragments would contain **1 byte of MoSS version information, 1 byte of data extension information, 64-bit memory address and 4 bytes of data**. A cluster of 10,000 different MoSS fragments can be immobilised on 1 ul of magnetic beads leading to a final storage of 40 KB of MoSS.

A Taq based PCR reaction, followed by sequencing would reveal the MoSS integrity and data validation. The final PCR yield could be purified and stored in cold storage using various methods of long term DNA storage.

The MoSS fragments would be stable for longer capacities, care should be taken while longer strand synthesis as data errors could creep up due to random loss of MoSS blocks if they were held by hydrogen bonding only. The best method for long fragment assembly would be to have a gap filling and ligation reaction (MoSS polishing) for individual fragments, opposed to a single pot reaction for lower capacity MoSS fragments.

The terminology of MoSS components is given below.

Tools and components of MoSS

1. DNA duplex stability prediction - http://biophysics.idtdna.com
2. G-quadruplex prediction - http://bioinformatics.ramapo.edu/QGRS/index.php
3. Oligo secondary structure prediction - https://eu.idtdna.com/calc/analyzer
4. LNA melting temperature - https://www.exiqon.com/ls/Pages/ExiqonTMPredictionTool.aspx
5. OpenTrons - https://opentrons.com
6. Exiqon LNA oligos - http://www.exiqon.com/custom-lna-oligos
7. IDT oligos - http://eu.idtdna.com/Site/Order/oligoentry
8. Thermo Dynabeads - https://www.thermofisher.com/order/catalog/product/65001
9. IDT buffers - https://www.idtdna.com/pages/products/reagents/buffers-and-solutions
10. NEB T4 DNA polymerase - https://www.neb.com/products/m0203-t4-dna-polymerase
11. NEB T7 DNA ligase - https://www.neb.com/products/m0318-t7-dna-ligase
12. Oxford Nanopore - https://nanoporetech.com
13. Illumina - http://www.illumina.com

About MoSS

MoSS - Molecular Storage System is a DNA data storage platform developed by Helixworks Technologies (https://helix.works).