SRC/IARPA WORKSHOP ON MASSIVE INFORMATION STORAGE

Inspired by DNA molecular structures

Abstract

DNA molecular structures can serve as storage media with a volumetric density (bits/unit volume) orders of magnitude greater than current storage media. This workshop explores the near-term feasibility of developing a massive information storage technology based on DNA organizing principles. The workshop is structured to encourage exploratory thinking with a goal of supporting an aggressive DNA-based memory system development cycle. In the following document, leading questions are offered to stimulate discussion for each of the workshop topical areas and these are followed by a set of straw five-year development objectives.

Contents

IARPA/SRC Workshop on DNA-based Massive Information Storage
April 27 & 28, 2016
Hyatt Arlington, 1325 Wilson Blvd., Arlington, VA.

Technical Program

Day 1 Sessions April 27, 2016 Morning

7:30 am—8:00 am	**Workshop Check-In**	**Senate Ballroom, Second Floor, Hyatt Arlington** (all Day 1 sessions will be held in this room)
8:00 am—8:30 am	**Welcoming Remarks**	Jason Matheny, Director, IARPA Tom Kalil, Deputy Director, OSTP Victor Zhirnov, Chief Scientist, SRC David Markowitz, Technical Advisor, IARPA
8:30 am—10:30 am	**Session 1**	**Nucleic Acid Memory for Massive Storage: Fundamentals and Application Perspectives**
	Keynote	George Church / Harvard University ***Next-Generation Digital Information Storage in DNA***
	Roundtable / Open Mic Discussions	**See Discussion Questions on Page 5** Karin Strauss / Microsoft Corp. Brian Bramlett / Intel Corp. John Cumbers / SynBioBeta Devin Leake / Gen9
10:30 am—10:45 am	**Break**	
10:45 am—12:45 pm	**Session 2**	**DNA Sequencing Technologies: Current Status and Future Directions**
	Keynote	Mostafa Ronaghi / Illumina, Inc. ***DNA Chips in Handsets***
	Roundtable / Open Mic Discussions	**See Discussion Questions on Page 6** Steve Turner / Pacific Biosciences John Kasianowicz / NIST Binquan Luan / IBM Timothy Lu / MIT

12:45 pm—2:00 pm	**Lunch (on your own)**	

Day 1 Sessions April 27, 2016 Afternoon

2:00 pm—4:00 pm	**Session 3**	**Toward A Practical DNA Hard Drive: Synthesis of Digitally Encoded Polymers**
	Keynote	Jean-François Lutz/ Inst Charles Sadron ***Design and synthesis of digitally encoded polymers That can be decoded and erased***
	Roundtable / Open Mic Discussions	**See Discussion Questions on Page 7** Sriram Kosuri / UCLA Bill Peck / Twist Biosciences Rob Carlson / Synthesis Will Hughes / Boise State U.
4:00 pm—4:15 pm	**Break**	
4:15 pm—6:15 pm	**Session 4**	**Operational Aspects of DNA-based Storage Systems**
	Keynote	Olgica Milenkovic / U. Illinois ***A Rewritable, Random-Access DNA-Based Storage System***
	Roundtable / Open Mic Discussions	**See Discussion Questions on Page 8** Gurtej Sandhu / Micron Inc. Jossy Sayir / European Bioinformatics Institute Luis Ceze / U. Washington Robert Grass / ETH Zurich Rafic Makki / GLOBALFOUNDRIES
6:15 pm	**Workshop Day 1 Concludes**	

Day 2 April 28, 2016 Morning Breakout Sessions

This morning, we will hold two rounds of parallel breakout sessions on three technical areas:

Group 1	**DNA Synthesis**	Senate Ballroom (Second Floor)
Group 2	**DNA Sequencing**	Marshall Room (Lobby Level)
Group 3	**Operational Aspects of DNA-based Storage Systems**	Douglas Room (Lobby Level)

Each technical area will be given a set of topics to discuss as a group. Each group also has a designated facilitator, who is responsible for summarizing the discussion and presenting to the reconvened general session after lunch. Please provide concrete, specific ideas!

9:00 am—10:30 am	**Breakout Round 1**	**Assigned Seating** Attendees will be notified of their group assignments prior to Day 2.
10:30 am—11:00 am	**Break**	
11:00 am—12:00 pm	**Breakout Round 2**	**Free Choice** Continue your first round discussion or join a new one.

Independent position papers are welcomed by any/all as a follow-up to these discussions. Please send no later than 11:59PM EST on 31 May 2016 to: david.markowitz@iarpa.gov
victor.zhirnov@src.org

12:00 pm—1:30 pm **Lunch (on your own)**

Day 2 Sessions April 28, 2016 Afternoon

1:30 pm—2:15 pm ***Breakout Group Reports (15 min each)*** Senate Ballroom (Second Floor)

2:15 pm—3:30 pm **Session 6** **Research Needs and Potential Responses** Senate Ballroom (Second Floor)

Facilitator Ralph Cavin / SRC

This session will examine technology gaps and gating research issues critical to enabling the future of DNA storage systems. The goal is to identify promising future directions for research from the perspectives of industry, university and government stakeholders.

Roundtable / Open Mic Discussions

See Discussion Questions on Page 9
Mitra Basu / NSF
David Markowitz / IARPA
Victor Zhirnov / SRC
Celia Merzbacher / SRC

3:30 pm—4:00 pm **Closing Comments / Adjourn** Senate Ballroom (Second Floor)

Session 1 Discussion Questions

Keynote speakers and discussion panelists are asked to directly address the following questions in their remarks, to help the workshop sponsors better understand opportunities and challenges in each problem domain.

Session 1 Nucleic Acid Memory for Massive Storage: Fundamentals and Application Perspectives

1) What is DNA-based massive information storage?
    a. Why is it important? (And to whom should it be important?)
    b. Why is it hard?
    c. What is the state of the art in this research area?
    d. What factors (e.g. funding, new technology) have driven research in this area until now?
    e. What are prospects for future progress, based on the current state of the field? Why?
2) What application drivers are supporting, or may arise from, progress in this research area?
    a. Suggest some early “ports of entry” (e.g. archival data storage)
    b. What are the limits of current practice in these application domains that DNA Storage technology could overcome?
    c. How could DNA Storage concepts provide new directions/inspiration for future information technologies, semiconductors, biomedical technologies, etc.? (e.g. by forcing us to think outside the confines of current paradigms)
3) What research areas stand to benefit from progress in DNA Storage research? How?
4) Is there more we can learn from biology to guide the development of DNA storage technologies?
5) The semiconductor industry has accumulated unique tools and experience that could be useful to DNA Storage research and development activities. What are some pathways for deployment of the expertise and infrastructure of the semiconductor industry in the DNA Storage domain?
6) What are some “far out” applications/consequences of practical DNA storage technology?
7) Please suggest companies, universities and government agencies that stand to benefit from direct participation in a collaborative applied research initiative on DNA Storage.
    a. What is the potential value of collaboration to each stakeholder?
    b. Are the incentives for each stakeholder properly aligned? If not, how to align them?

Session 2 Discussion Questions

Keynote speakers and discussion panelists are asked to directly address the following questions in their remarks, to help the workshop sponsors better understand opportunities and challenges in each problem domain.

Session 2 DNA Sequencing Technologies: Current Status and Future Directions

1) To the extent possible, address the Straw Target Goals on pages 10-13 of this document.
2) Is the speed of DNA sequencing on-track to achieve the goals for TA2?
    a. Are the fundamental physics limits of DNA sequencing speed known?
3) What are ideal sequencing capabilities to have 5 years from now?
    a. What applications would this enable?
    b. What is the time horizon for miniaturized (e.g. portable) DNA sequencing technologies?
4) Envision the properties of a sequencer that supports DNA memory
    a. Winning technology?
    b. Limits of scaling? (e.g. tradeoffs between size and capabilities)
    c. Integration and interfacing with semiconductor electronics?
        i. What electronic device structures are needed; can biological components be incorporated?
    d. What interfaces are possible between active semiconductor substrates and molecular chemistry? What are advantages and limitations of different types of physical stimuli and responses that could be used in bi-directional chemistry-to-semiconductor communication, e.g. Optical, Electrical, Chemical, Mechanical?

Session 3 Discussion Questions

Keynote speakers and discussion panelists are asked to directly address the following questions in their remarks, to help the workshop sponsors better understand opportunities and challenges in each problem domain.

Session 3 Toward A Practical DNA Hard Drive: Synthesis of Digitally Encoded Polymers

1) To the extent possible, address the Straw Target Goals on pages 10-13 of this document.
2) Is the speed of DNA synthesis on-track to achieve the goals for TA1?
    a. What needs to be done to obtain a radical decrease in the cost of DNA synthesis? When would this be possible?
    b. Is it possible for the parameters of practical DNA synthesis to approach the parameters achieved in living cells, e.g. the read/write latency $<100\mu s$/bit, operating power $<10^{-11}$W/GByte?
3) What are ideal synthesis capabilities to have 5 years from now?
    a. What applications would this enable?
    b. What is the time horizon for miniaturized (e.g. portable) DNA synthesis technologies?
4) Envision the properties of a synthesizer that supports DNA memory
    a. Winning technology?
    b. Limits of scaling? (e.g. tradeoffs between size and capabilities)
    c. Integration and interfacing with semiconductor electronics?
        i. What electronic device structures are needed; can biological components be incorporated?
    d. What interfaces are possible between active semiconductor substrates and molecular chemistry? What are advantages and limitations of different types of physical stimuli and responses that could be used in bi-directional chemistry-to-semiconductor communication, e.g. Optical, Electrical, Chemical, Mechanical?

Session 4 Discussion Questions

Keynote speakers and discussion panelists are asked to directly address the following questions in their remarks, to help the workshop sponsors better understand opportunities and challenges in each problem domain.

Session 4 Operational Aspects of DNA-based Storage Systems

1) To the extent possible, address the Straw Target Goals on pages 10-13 of this document.
2) Discuss expected retention time of DNA memory. Is perceived 'fragility' of biological matter a concern for DNA as a practical storage technology?
3) What is the envisioned 'mechanical' design/assembly of a hypothetical DNA hard drive?
4) How could one implement hierarchical DNA-based storage platforms that enable random access memory for rapid information retrieval on multiple scales?
5) Discuss approaches for DNA encodings that support reliable storage, error correction, accessibility, data compression, etc.
6) What is the optimum/preferred base for storing digital information in DNA? (e.g. binary, ternary…)
7) Is it possible to expand the concept of DNA storage beyond the ATGC code, e.g. by using other synthetic nucleotides? Are there potential benefits for the extension of the DNA alphabet?
8) Do the methods that are used by biological systems to protect information have implications for Cybercecurity and/or Biosecurity?

Session 5 Breakout Groups (Discussion questions to be determined after Day 1)

Session 6 Discussion Questions

Keynote speakers and discussion panelists are asked to directly address the following questions in their remarks, to help the workshop sponsors better understand opportunities and challenges in each problem domain.

Session 6 **Research Needs and Potential Responses**

1) Given the existing technology landscape, toward which Straw Target Goals do you think the research community could contribute most substantially?
    a. What are the greatest challenges?
    b. What are the most promising technical approaches for solving each challenge?
    c. What are the risks and the payoffs of each technical approach?
    d. What are the necessary ingredients for success?
    e. What are the major cost drivers of each technical approach? What is the likely cost?
    f. How long will each technical approach take?
    g. How should progress be measured? (i.e. what metrics and tests should be applied)
2) What are the most promising opportunities for collaborative R&D of future DNA information storage technologies by industry, university and government stakeholders?
    a. What would a successful collaboration look like, and what challenges would be faced?
3) What principles should such a collaborative model embrace?
    a. How should the research be implemented?
    b. What mechanisms should be used to develop research goals and ensure collaborative focus?
    c. How can agility be maintained (with respect to emerging research opportunities)?
    d. What mechanisms should be put in place to enable rapid transfer of knowledge and technical capabilities to the sponsors?

Straw Target Goals for "DNA Hard Drive" Concept

Here we suggest goals, technical areas, and metrics of success that could be used as a strategy to guide the development of "DNA Hard Drive" technology. This information is suggested as an initial point of departure for group discussions, and should not be treated as limiting constraints. The organizers want to hear your critiques, additions, refinements and suggested alternatives to this information.

Goal Within five years, develop a one Exabyte (10^{18}) storage technology with the following properties:

A. One Petabyte (10^{15}) per day can be written into the storage system
B. Ten Petabytes per day can be retrieved from the storage system
C. Operational energy consumption is less than 200 kW per day
D. Information can be stored without degradation for at least 100 years

Approach Explore the use of DNA technologies and methods as a basis for realizing massive information storage.

Technical Areas In the following, three primary technical areas (TAs) are identified and metrics are offered that, if met, would result in meeting the 5-year goals highlighted above.

TA1: DNA Synthesis and Storage

- **Motivation**
  - Future global networks will generate *O*(Petabytes/day) much of which needs to be saved.
  - The data must be organized for rapid access and the storage media must be removable for secure and separate information storage.
- **State of the Art**
  - DNA synthesis[1] is slow and expensive (~$149/1000 bits per second)[2]
- **Challenges**
  - Speed synthesis while driving down cost
  - DNA storage for rapid access
  - Technology-independent archival storage for long-term access

[1] **DNA synthesis** is the natural or artificial creation of deoxyribonucleic acid (DNA) molecules by either DNA replication, polymerase chain reaction, or by artificial gene synthesis

[2] Tabatabaei Yazdi, et. al., Sci. Rep., 2015

- **Phase 1 Metrics**
  - Speed
    - Synthesis at $\geq$ 1 Terabyte/day
    - Write time at $\leq$ 10x synthesis time
  - Error rate
    - < 1 nt[3] / Million nt (1Mnt) synthesized
  - Cost of reagents & consumables
    - < $1k/terabyte written & stored
  - Energy consumption
    - < 100 kW-hours /100 Terabytes written and stored
    - 0 kW-hours for ongoing storage
  - Storage capacity: 1 Petabyte
  - Stability of stored data: > 20 years
  - Storage media must be capable of removal/replacement
- **Phase 2 Metrics**
  - Speed:
    - Synthesis at >= 100 Terabytes / day
    - Write time <= 2x synthesis time.
  - Error rate: < 1 base-pair error / 100 Mbp synthesized
  - Cost of reagents and other consumables: <$1k / 100 Terabytes written and stored
  - Energy consumption:
    - < 100 kW hours / 100 Terabytes written and stored
    - 0 kW hours for ongoing storage
  - Storage capacity: 100 Petabytes
  - Estimated stability of stored data: > 50 years
  - Storage capabilities must be robust to removal and replacement of storage media.
- **Phase 3 Metrics**
  - Speed:
    - Synthesis at >= 1 Petabyte / day
    - Storage time <= 1x synthesis time
  - Error rate: < 1 bp error / 1 Gbp synthesized
  - Cost of reagents and other consumables: <$1k / 1 Petabyte written and stored
  - Energy consumption:
    - < 100 kW hours / 1 Petabyte written and stored
    - 0 kW hours for ongoing storage
  - Storage capacity: 1 Exabyte
  - Estimated stability of stored data: > 100 years
  - Storage capabilities must be robust to removal and replacement of storage media.

[3] One nucleotide (nt) is equivalent to two bits of data

TA2: DNA Sequencing[4]

- **Motivation:**
  - A random-access read mechanism is required to read specified segments of stored DNA non-destructively.
- **State of the art:**
  - Illumina Hi-Seq X Ten: 50 human genomes/day * 200 Gb/genome = 10 Tb/day
- **Challenges:**
  - Current technologies, such as DRAM, destroy data as they read, but this may be undesirable for a molecular storage system that operates over slower time scales. Is it possible to sequence DNA *in situ* without disrupting stored information?
- **Phase 1 Metrics:**
  - Sequencing speed: >= 100 Terabytes / day
  - Unrecoverable Bit-Error rate: 10^{-13}
  - Cost of reagents and other consumables: <$1k / 10 Terabytes read
  - Energy consumption: < 100 kW hours / 10 Terabytes read
  - Sequencing capabilities must be robust to removal and replacement of storage media.
- **Phase 2 Metrics:**
  - Sequencing speed: >= 1 Petabyte / day
  - Unrecoverable Bit-Error rate: < 10^{-14}
  - Cost of reagents and other consumables: <$1k / 1 Petabyte read
  - Energy consumption: < 100 kW hours / 1 Petabyte read
  - Sequencing capabilities must be robust to removal and replacement of storage media.
- **Phase 3 Metrics:**
  - Sequencing speed: >= 10 Petabytes / day
  - Unrecoverable Bit-Error rate < 10^{-15}
  - Cost of reagents and other consumables: <$1k / 10 Petabytes read
  - Energy consumption: < 100 kW hours / 10 Petabytes read
  - Sequencing capabilities must be robust to removal and replacement of storage media.

[4] **DNA sequencing** is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases—adenine, guanine, cytosine, and thymine—in a strand of DNA.

TA3: DNA File System

- **Motivation:**
  - Integrate TA1 and TA2 technologies into a system that provides fast and reliable random-access read/write capabilities.
- **State-of-the-art:**
  - (Bornholt et al, ASPLOS 2016) demonstrates a DNA storage system that provides random access using a key-value store, and controllable redundancy that trades off reliability for density.
- **Challenges:**
  - Identify the most appropriate data encoding/decoding schemes and their implementation. These should support error-correction and ideally compression.
  - Determine the design and implementation for a file system to manage lookups/reads/writes.
- **Phase 1 Metrics:**
  - Speed: Write 1000 files (1 Tb total) and read each file 10 times, all in < 24 hours
- **Phase 2 Metrics:**
  - Speed: Write 10k files (100 Tb total) and read each one 10 times, all in < 24 hours
- **Phase 3 Metrics:**
  - Speed: Write 1M files (1 Pb total) and read each one 10 times, all in < 24 hours