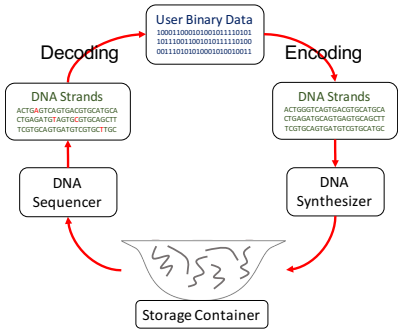
שם הקורס: סמינר על אחסון מידע בדנא
Seminar on DNA Data Storage
מספר הקורס 236804

מרצה:	פרופ' איתן יעקובי
שעות הרצאה:	ראשון 10:30-12:30
דרישות קדם:	ראה תאור הקורס

תאור הקורס

DNA-based storage has attracted significant attention due to recent demonstrations of the viability of storing information in macromolecules. The potential for using macromolecules for ultra-dense storage was recognized as early as in the 1960s, when the celebrated physicist Richard Feynman outlined his vision for nanotechnology in the talk "*There is plenty of room at the bottom*". DNA molecules, which may be abstracted as strings over the four-symbol alphabet A, C, G, T, have been successfully used as a building block of several small scale self-assembly based computers. DNA lends itself to implementations of non-volatile recoding media of outstanding integrity due to a few unique properties: (i) self-assembly potential (DNA has been successfully used in several small scale self-assembly based computers), (ii) stability (DNA can be recovered from 30,000 years old Neanderthal), and iii) capacity (a single human cell, with a mass of roughly 3 pgrams, hosts DNA strands encoding 6.4GBs of information). Furthermore, the technologies for synthesizing artificial DNA and for sequencing have reached unprecedented levels of efficiency and accuracy and given the trends in cost decreases of DNA synthesis and sequencing, it is estimated that within the next 5-10 years DNA storage will become a highly competitive archiving technology.

A DNA storage system consists of three important entities; The first is a DNA synthesizer that produces the strands that encode the data to be stored in DNA. In order to produce strands with acceptable error rate the length of the strands is typically limited to no more than 250 nucleotides. The second part is a storage container with compartments that stores the DNA strands, however unordered. Lastly, a DNA sequencer reads back the strands and transfers them back to digital data. The encoding and decoding stages are two external processes to the storage systems which convert the binary user data into strands of DNA in such a way that even in the presence of errors (the nucleotides in red), it will be possible to revert back to the original binary data of the user. DNA as a storage system has several attributes which distinguish it from any other storage system. The most outstanding one is that the strands are not ordered and thus it is not possible to know the order in which they were stored. Usually, this constraint can be overcome by using indices, that are stored as part of the strand. Note that this limitation already imposes the capacity of DNA storage to be strictly less than 2 bits per nucleotide. This structure also prevents random access to the stored data since it is not possible to read a given strand in the pool and most of the proposed systems must read the entire pool to retrieve even a single strand.



The goal of this seminar is to cover and address coding-theoretic challenges and solutions arising in the context of synthesis, storage, and sequencing of DNA strands as well as other related problems.

דרישות הקורס

The course will combine lectures by the instructors with independent reading in a seminar format. The students will read important papers in the field, will reason about the results in a critical way, and will present them in class along with their own ideas for extending the results.

דרישות קדם

The course combines some techniques from the fields of linear and modern algebra, coding theory, and algorithms. There are no formal pre-requisites, but the following courses provide good background:

104134 Modern Algebra

234237 Algorithms 1

To register, email your name and ID to the instructor. Please state whether you are a graduate or undergraduate student, and other information you think is relevant, such as related courses you took.

תוצרי למידה

By the end of the seminar, the student will understand the concept of coding and algorithms for DNA storage and how to design codes for error correction and for constraints in DNA.

ספרות

The course will survey recent literature with the state-of-the-art works on DNA storage, for example:

1. T. Shinkar, E. Yaakobi, A. Lenz, and A. Wachter-Zeh, *Clustering-Correcting Codes*, to appear I*EEE Trans. Inform. Theory*.
2. M. Levy and E. Yaakobi, *Mutually Uncorrelated Codes for DNA Storage*, *IEEE Trans. Inform. Theory*, vol. 65, no. 6, 3671–3691, Jun. 2019.
3. A. Lenz, P.H. Siegel, A. Wachter-Zeh, and E. Yaakobi, *Coding over Sets for DNA Storage*, IEEE Trans. Inform. Theory, vol. 66, no. 4, pp. 2331–2351, Apr. 2020.
4. O. Sabary, D. Bar-Lev, Y. Gershon, A. Yucovich, and E. Yaakobi, *On The Decoding Error Weight of One or Two Deletion Channels*, submitted to I*EEE Trans. Inform. Theory*.
5. V.I. Levenshtein, *Efficient reconstruction of sequences*, *IEEE Trans. on Inform. Theory*, vol. 47, no. 1, pp. 2–22, Jan. 2001.
6. C. Rashtchian, K. Makarychev, M. Racz, S. Ang, D. Jevdjic, S. Yekhanin, L. Ceze, and K. Strauss, *Clustering billions of reads for DNA data storage*, *NIPS*, 2017.
7. L. Anavy, I.Vaknin, O. Atar, R. Amit, and Z.Yakhini, *Improved DNA based storage capacity and fidelity using composite DNA letters*, *Nature Biotechnology*, vol. 10, pp. 1229–1236, 2019.
8. S.K. Tabatabaei, B.Wang, N.B.M. Athreya, B. Enghiad, A.G. Hernandez, J.-P. Leburton, D. Soloveichik, H. Zhao, and O. Milenkovic, *DNA punch cards: Encoding data on native DNA sequences via topological modifications*, *Nature Communications*, vol. 11, no. 1742, Apr. 2020.
9. S. Chandak, K, Tatwawadi, B. Lau, J. Mardia, M.Kubit, J. Neu, P. Griffin, M. Wootters, T.Weissman, and H. Ji, *Improved read/write cost tradeoff in DNA-based data storage using LDPC codes*, *BioRxiv*, Sep. 2019.
10. Y. Erlich and D. Zielinski, *DNA fountain enables a robust and efficient storage architecture*, *Science*, vol. 355, no. 6328, pp. 950–954, 2017.
11. M. Blawat, K. Gaedke, I. H¨utter, X.-M. Chen, B. Turczyk, S. Inverso, B.W. Pruitt, and G.M. Church, *Forward error correction for DNA data storage*, *Int. Conf. on Computational Science*, vol. 80, pp. 1011–1022, 2016.
12. Choi, Y., Ryu, T., Lee, A.C. *et al. High information capacity DNA-based data storage with augmented encoding characters using degenerate bases*, *Scientific Report*, vol. 9, no. 6582, Apr. 2019.
13. Lee, H.H., Kalhor, R., Goela, N. *et al. Terminator-free template-independent enzymatic DNA synthesis for digital information storage*, *Nature Communications*, vol. 10, no. 2383, Jun. 2019.
14. S. H. T. Yazdi, R. Gabrys, and O. Milenkovic, *Portable and error-free DNA-based data storage*, *Scientific Reports*, vol. 7, sp. 5011, Jul. 2017.
15. Organick, L., Ang, S., Chen, YJ. *et al. Random access in large-scale DNA data storage*, *Nature Biotechnology*, vol. 36, pp. 242-248, Feb. 2018.