NANO LETTERS

pubs.acs.org/NanoLett

Letter

Expanding the Molecular Alphabet of DNA-Based Data Storage Systems with Neural Network Nanopore Readout Processing

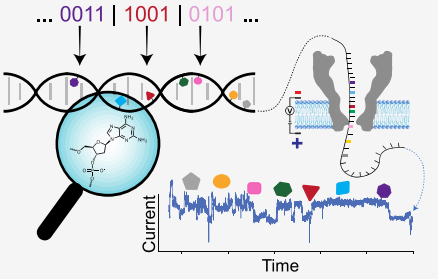
S. Kasra Tabatabaei,[∞] Bach Pham,[∞] Chao Pan,[∞] Jingqian Liu,[∞] Shubham Chandak, Spencer A. Shorkey, Alvaro G. Hernandez, Aleksei Aksimentiev,[*] Min Chen,[*] Charles M. Schroeder,[*] and Olgica Milenkovic[*,∞]

Cite This: *Nano Lett.* 2022, 22, 1905−1914

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: DNA is a promising next-generation data storage medium, but challenges remain with synthesis costs and recording latency. Here, we describe a prototype of a DNA data storage system that uses an extended molecular alphabet combining natural and chemically modified nucleotides. Our results show that MspA nanopores can discriminate different combinations and ordered sequences of natural and chemically modified nucleotides in custom-designed oligomers. We further demonstrate single-molecule sequencing of the extended alphabet using a neural network architecture that classifies raw current signals generated by Oxford Nanopore sequencers with an average accuracy exceeding 60% (39× larger than random guessing). Molecular dynamics simulations show that the majority of modified nucleotides lead to only minor perturbations of the DNA double helix. Overall, the extended molecular alphabet may potentially offer a nearly 2-fold increase in storage density and potentially the same order of reduction in the recording latency, thereby enabling new implementations of molecular recorders.



KEYWORDS: *DNA Data Storage, Unnatural Nucleotides, Nanopores, Single-Molecule, Neural Networks*

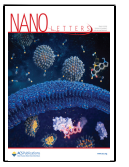
INTRODUCTION

DNA is emerging as a robust data storage medium that offers ultrahigh storage densities greatly exceeding conventional magnetic and optical recorders. Information stored in DNA can be copied in a massively parallel manner and selectively retrieved via polymerase chain reaction (PCR).[1−10] However, existing DNA storage systems suffer from high latency caused by the inherently sequential writing process. Despite recent progress, a typical cycle time of solid-phase DNA synthesis is on the order of minutes, which limits the practical applications of this molecular storage platform.[11] Using current technologies, writing 100 bits of information requires nearly 2 h [11] and costs more than U.S. $1,[12] assuming that each nucleotide stores its theoretical maximum of two bits. To overcome these challenges, new synthesis methods and information encoding approaches are required to accelerate the speed of writing large-volume data sets.[13]

Expanding the alphabet of a DNA storage media by including chemically modified DNA nucleotides can both increase the storage density and the writing speed because more than two bits are recorded during each synthesis cycle. However, designing chemically modified nucleotides as new letters for the DNA storage alphabet must be tightly coupled to the process of reading the encoded information via DNA sequencing, because current DNA sequencing methods, including single-molecule nanopore sequencing, have been developed and optimized to read natural nucleotides. Prior work reported an expanded nucleic acid alphabet of synthetic DNA and RNA nucleotides that can be replicated and transcribed using biological enzymes,[14] but this alphabet was not designed for molecular storage applications and was not accurately read using a nucleic acid sequencing method. Aerolysin nanopores were used to detect synthetic polymers flanked by adenosines, where each monomer of the polymer carries one bit of information.[15] Prior work has reported successful detection of base pairs containing single chemically modified nucleotides[16,17] or discrimination of single nucleotides in natural versus modified states.[18] Despite recent advances, single-molecule detection and sequencing of an expanded molecular alphabet based on a library of chemically diverse modified nucleotides has not yet been demonstrated.

Here, we report an expanded molecular alphabet for DNA data storage comprising four natural and seven chemically modified nucleotides that are readily detected and distinguished using nanopore sequencers (Figure 1 and Table 1). Our results show that *Mycobacterium smegmatis* porin A

Received: November 1, 2021
Revised: February 22, 2022
Published: February 25, 2022



https://doi.org/10.1021/acs.nanolett.1c04203
Nano Lett. 2022, 22, 1905−1914

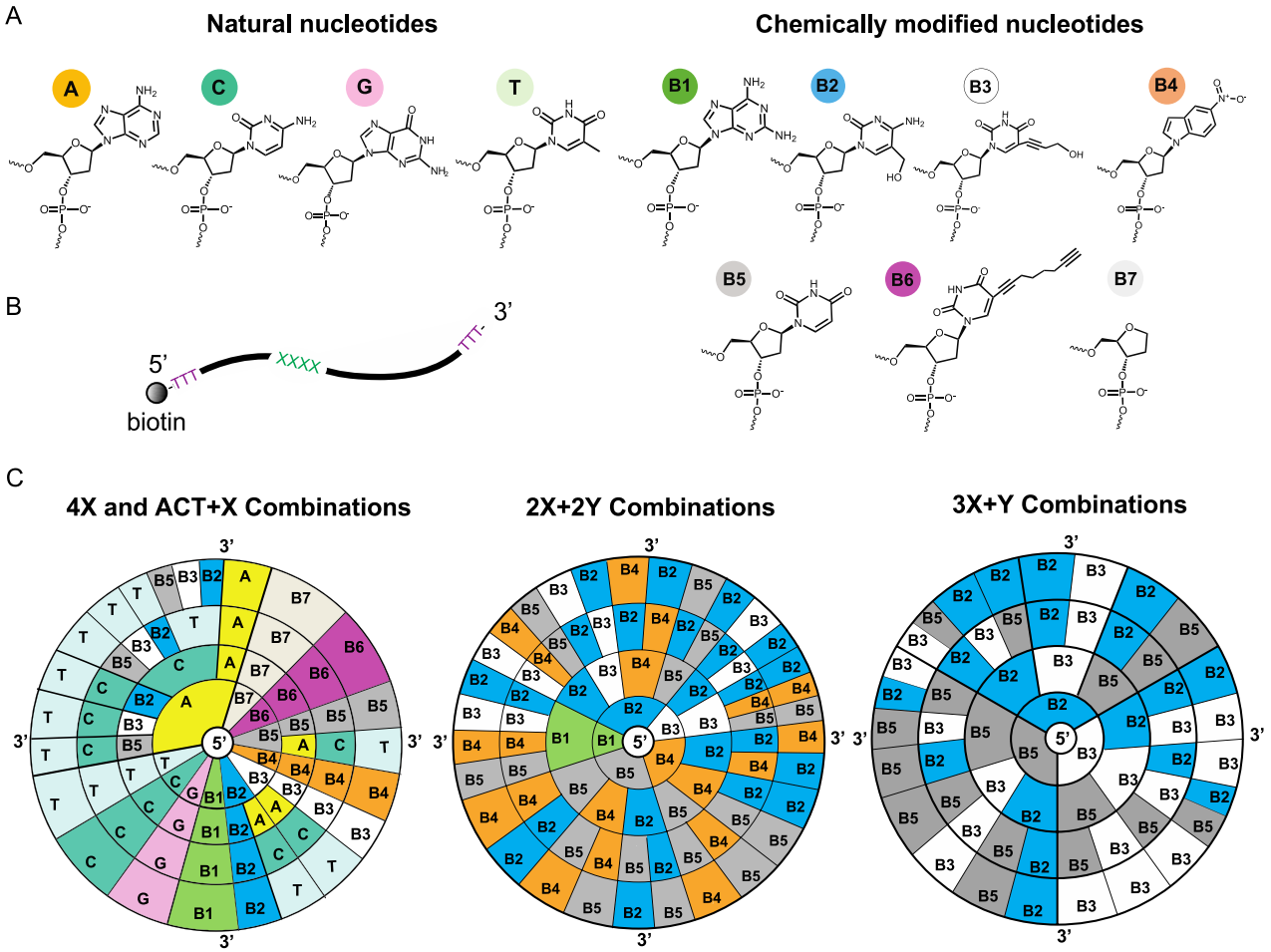


Figure 1. DNA data storage using natural and chemically modified nucleotides. (A) Chemical structures of natural DNA nucleotides (A, C, G, T) and the selected chemically modified nucleotides employed in our study (B1−B7). (B) Schematic of the ssDNA oligo used in MspA nanopore experiments. The length of the oligos is 40 nucleotides (nts), with biotin attached at the 5′ terminus. Homo- or heterotetrameric sequences are located at positions 13−16, flanked by two polyT regions of length 12 nt and 24 nt on the 5′ and 3′ ends, respectively. (C) Sequence space for DNA homotetramers or heterotetramers used in MspA nanopore experiments. The notation $aX + bY$, where a and b take values in {2, 3, 4} so that $a + b = 4$, indicates that "a" symbols of the same kind are combined with "b" symbols of another kind and arranged in an arbitrary linear order. In total, 77 distinct tetrameric sequences were synthesized and tested experimentally. (Left) Circular diagram showing all 11 homotetramers and 12 tetrameric sequences of the form ACT + X, where X is a chemically modified nucleotide from the set {B2, B3, B5}. (Middle) Circular diagram showing all 30 tested combinations of tetrameric sequences with total composition 2X + 2Y using chemically modified monomers from the set {B1, B2, B3, B4, B5}, including sequence patterns XXYY, XYYX, and XYXY. (Right) Circular diagram showing the remaining 24 combinations of tetrameric sequences with total composition 3X + Y using the set {B2, B3, B5}. Five chemically modified nucleotides form stable base pairs with natural nucleotides via hydrogen bonds (B2−G, B3−A, B5−A, B6−A, B6−C), based on the results from molecular dynamic (MD) simulations.

(MspA) nanopores, which are widely used for ssDNA sensing and single molecule chemistry studies,[19−21] can accurately discriminate 77 combinations and orderings of chemically diverse monomers within homo- and heterotetrameric sequences (Figures 1, 2, S1, and S2 and Tables S1−S3). We further demonstrate that highly accurate classification (exceeding 60% on average) of combinatorial patterns of natural and chemically modified nucleotides is possible using deep learning architectures that operate on raw current signals generated by GridION of Oxford Nanopore Technologies (ONT)[22] (Figures 3, S3, and S4). We further study the stability of DNA duplexes containing modified nucleotides using all-atom molecular dynamics (MD) simulations[23−26] (Figures 4, S5, and S6 and Table S5). Overall, the extended molecular alphabet has the potential to offer a nearly 2-fold increase in storage density and potentially the same order of reduction in recording latency, thereby providing a promising path forward for the development of new molecular recorders.

RESULTS AND DISCUSSION

To determine whether natural and chemically modified DNA nucleotides can be distinguished using the biological nanopore MspA, we designed a series of single-stranded DNA (ssDNA) molecules with the general sequence 5′-biotin-$(dT)_{12}$-XXXX-$(dT)_{24}$-3′, where X = {A, T, C, G, B1−B7} (Figure 2, Figures S1 and S2, Tables S1−S3). We hypothesized that specific chemical modifications to nucleobases such as amines, alkynes, or indole moieties can alter polymer−amino acid interactions in biological nanopores, thereby generating distinct signals in nanopore readouts. In the process, we also considered the stability of base pairing and base stacking interactions between natural and chemically modified nucleotides using a combination of MD simulations and experiments (Tables 1 and S5, Figures 4 and S5−S7).

Following molecular design and synthesis of ssDNA oligos (the chemical characterization and mass spectrometry analysis

https://doi.org/10.1021/acs.nanolett.1c04203
Nano Lett. 2022, 22, 1905−1914

Table 1. Chemically Modified Nucleotides Used in the DNA Data Storage System, Along with Their Chemical Properties[a]

Symbol	B1	B2	B3	B4	B5	B6	B7
Name	2,6-Diamino-purine 2′-deoxyriboside	5- Hydroxymethyldeoxy-cytidine	5-Hydroxy-butynl-2′-deoxyuridine	5-Nitroindole-2′-deoxy-riboside	Deoxyuridine	5-Octadiynyl-deoxyuridine	1,2-Dideoxy-ribose
Structurally most similar nucleo-tide	dA	dC	dT	dA	dT	dT	-
Pairing mate/in-teraction type (experiment*)	dT H bonds[28−30]	dG H bonds[31]	dA-	All natural nucleotides stacking[28,32]	dA H bonds[33]	dA H bonds[34,35]	-
Pairing mate/in-teraction type (simulation**)	dT H bonds	dG H bonds	dA H bonds	dG stacking	dA H bonds	dA, dC H bonds	-

[a]The symbols and the names of the chemically modified nucleotides are shown in the first and second rows, and the molecular structures are depicted in Figure 1. Structurally similar natural nucleotides are shown in the third row. In general, distinct chemical functional groups and molecular charges play an important role in discriminating nucleobases using MspA and ONT sequencers. The last two rows show pairing properties of the modified bases. * denotes data from Integrated DNA Technologies[28] or experiment data from previous work,[29−35] while ** denotes results from molecular dynamics simulations reported in Figure 4 and the Supporting Information (Figures S5 and S6, Table S5). Short dashes indicate that pairing is inherently impossible (e.g., B7) or that no specific information is published (e.g., interaction type of B3-dA pairing).

of oligos containing chemically modified nucleotides are provided in the Supporting Information (Figures S8−S84)), we performed MspA nanopore experiments where ssDNA oligos containing biotin at the 5′ terminus were electrophoretically attracted inside MspA nanopores. The bulky streptavidin protein prevents the oligos from fully translocating through the pore without appreciably affecting the measured ionic currents.[27] Consequently, ssDNA molecules are effectively immobilized within MspA nanopores, exposing the four nucleotides at positions 13−16 from the tethering point to the constriction of the MspA pore (Figure 2A).[36] In this assay, streptavidin holds ssDNA in the MspA constriction in a similar fashion to a helicase enzyme that steps through double-stranded (dsDNA) in an ONT sequencer, thereby enabling long duration current readings for each sequence tetramer (Figure S1).

We next used MspA nanopores to determine residual currents for homotetrameric sequences of all natural and chemically modified monomers (Figure 2B). Our results show that MspA accurately discriminates all four natural (A, G, C, T) and nearly all chemically modified nucleotides (B1−B7) at an applied bias of 150 mV. The abasic nucleotide B7 shows the largest residual current, which likely arises due to its small molecular size and reduced ability to interact with the reading head of MspA. The residual current levels are sensitive to the chemical identity of the nucleotides but do not directly correlate with their molecular size (Figure 2B). For example, current signals from B6 and B2 overlap at 150 mV, but B6 is well separated from B3 despite being structurally similar. We further studied the effect of the applied bias on the resolution of nucleotide bases. At 150 mV, four chemically modified nucleotides (B2, B3, B4, B5) showed well-resolved signals from each other and the natural nucleotides, but the current levels from B6 exhibited around 68% overlap with B2. Upon increasing the applied bias to 180 mV, the resolution between B2 and B6 was significantly improved, with an overlap area of the fitted Gaussian curves of 18%. In addition, at 180 mV, resolution in the I_{res} region exceeding 20% decreased, as may be seen from the residual currents of B4, A, and G which have Gaussian readout distributions that overlap in area by more than 90% (Figure 2B).

We further used MspA to detect and identify heterotetrameric sequences with compositions 2X + 2Y, where X, Y = {B2, B3, B4, B5} (Figure 2C, Figures S1 and S2, Tables S1−S3). Our results show that MspA can distinguish all heterotetrameric sequences with the same nucleotide composition when measurements at all three applied biases (150 mV, 180 mV, 200 mV) are performed. Due to the large sequence space explored, here we focus our discussion on representative tetrameric combinations of B2 and B3 (Figure 2C). In most cases, the residual currents of heterotetramers fall between those of two corresponding homotetramers. For example, the tetramer 3223 has an I_{res} of 12.3%, whereas those of B2 and B3 are 10.2% and 12.6%, respectively (at 180 mV). However, some combinations of B2 and B3, including 2232, 2322, 2333, 3233, 2323, 2332, and 2233, showed significant decreases in residual currents compared to homotetramers B2 and B3 (Figure 2C), whereas the residual current of tetramer 3322 is larger than homotetramers of B2 and B2 at either 150 mV or 180 mV. Importantly, all tetrameric sequences were resolved by adjusting the applied bias.[37] At a higher applied bias of 200 mV, tetramers that were unresolved at lower bias were readily resolved, including 2322, 2332, and 2322 (Figure 2C). Overall, these results are consistent with the observation that the residual current levels of DNA tetramers are not directly correlated with molecular size, similar to the case of natural nucleotides[38] where the blockade current was found to be determined by the competition of steric and base stacking interactions.[39]

We next investigated the ability of MspA pores to resolve different tetramers containing both natural and chemically modified nucleotides (Figure 2D). Here, we specifically focused on heterotetramers containing a single chemically modified nucleotide (B2, B3, or B5) added in different positions of the directional sequence ACT.[38] Our results clearly show that different positions of the chemically modified nucleotide in the tetramer generate distinct residual currents. For example, the residual current of heterotetrameric sequences of ACT containing four different positions of B2 (2ACT, A2CT, AC2T, and ACT2) are readily resolved at both 150 mV and 180 mV (Figure 2D). Although the residual current of homotetramer B2 and heterotetramer 2ACT overlap by ∼29% in their Gaussians at 150 mV, they are distinguishable at 180 mV. In addition, nearly all heterotetrameric sequences of ACT containing four different positions of B3 were resolved from the homotetramer B3 at 150 and 180 mV, whereas the residual currents of 3ACT and ACT3 were only distinguishable at 180 mV (Figure 2D). These results are

https://doi.org/10.1021/acs.nanolett.1c04203
Nano Lett. 2022, 22, 1905−1914

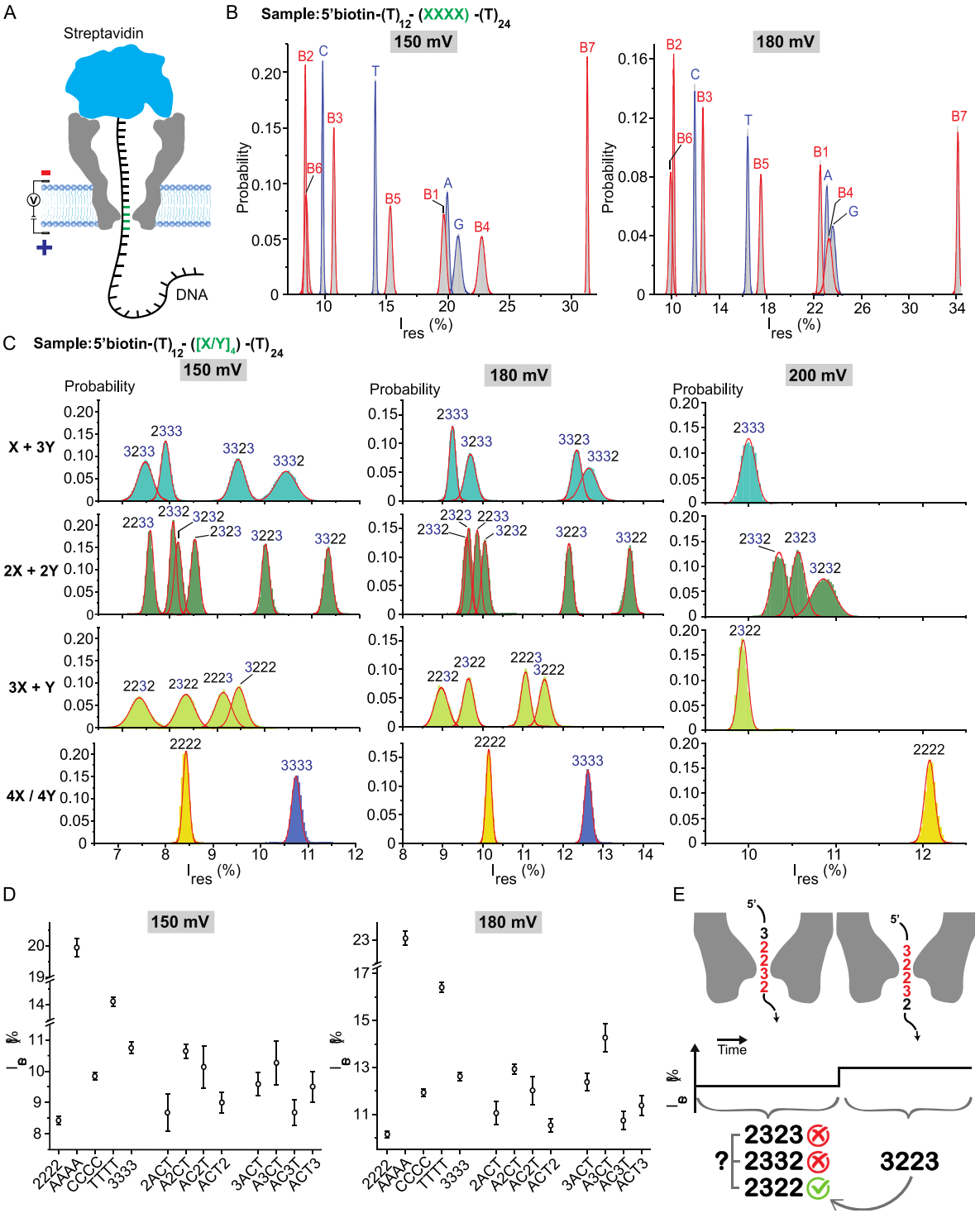


Figure 2. Identification of chemically modified DNA using MspA nanopores. (A) Schematic diagram of ssDNA immobilized in a MspA nanopore, where ssDNA containing a biotin−streptavidin interaction at the 5′ terminus prevents translocation through the pore. Residual ion current generated by four nucleotides at positions 13−16 from the 5′ terminus is recorded for ssDNA immobilized in the pore. (B) Histograms of average residual ionic currents I_{res} shown in gray for different homopolymers (A, T, C, G, and B1−B7). The fitted Gaussian curves are depicted in red for natural nucleotides (A, T, C, G) and in blue for chemically modified nucleotides (B1−B7). (C) Histograms of the average residual ionic currents and the fitted Gaussian curves at various applied voltages for tetramers involving different combinations and orderings of B2 and B3. (D) Peak values (points) and confidence intervals (bars) of the fitted Gaussians with mean residual ionic currents corresponding to tetramers obtained by inserting one of the monomers B2 and B3 into the sequence ACT, at applied biases of 150 mV and 180 mV. (E) Schematic of the shift reconciliation method for resolving ambiguities in the readouts of different tetramers.

consistent with prior work reporting that tuning the applied bias is a useful approach to enhance the accuracy of nanopore-based sequencing methods.[40] In summary, these results show the ability of MspA nanopores to accurately identify sequences containing chemically modified nucleotides.

https://doi.org/10.1021/acs.nanolett.1c04203
Nano Lett. 2022, 22, 1905−1914

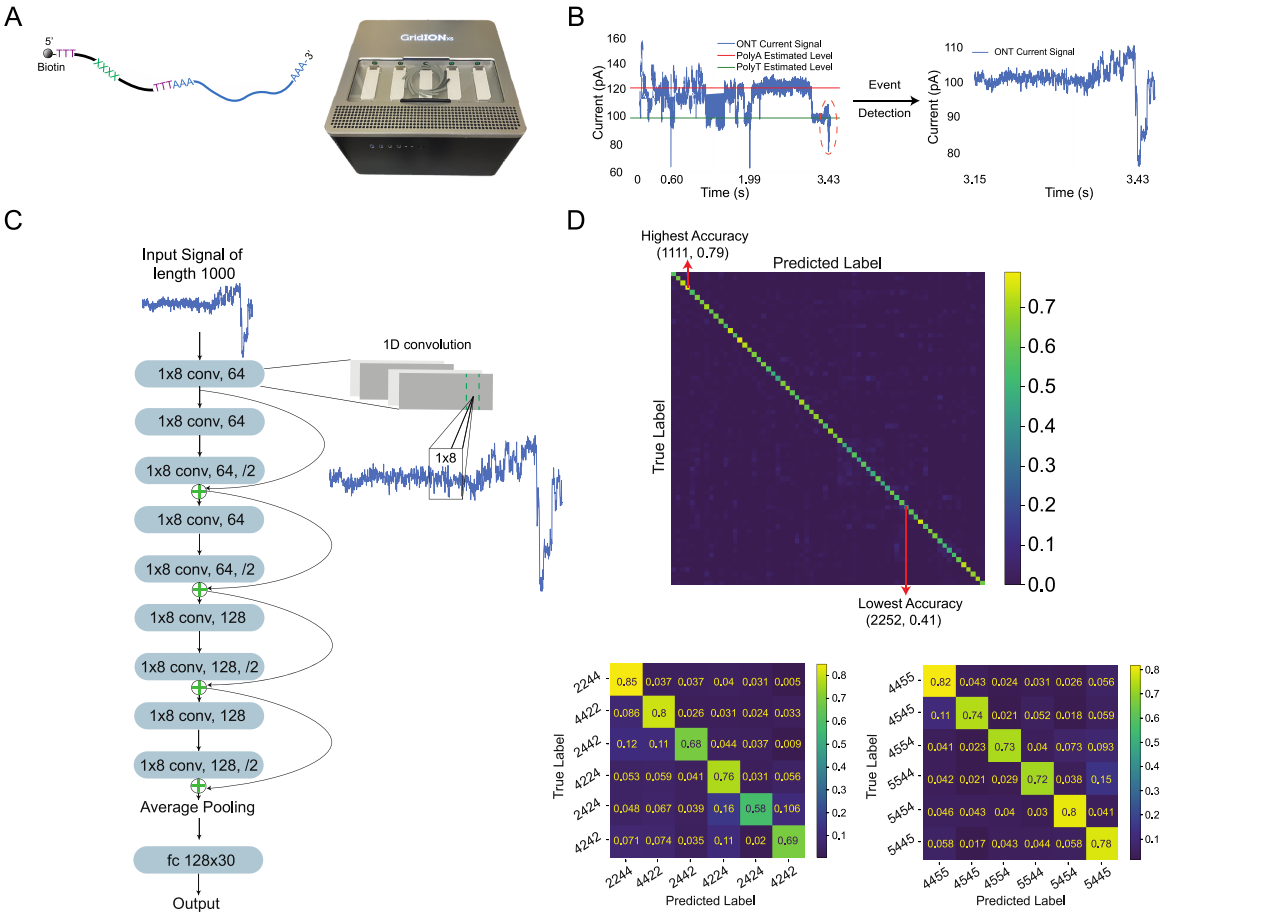


Figure 3. Sequencing oligos containing chemically modified nucleotides using ONT GridION. (A) Schematic of oligo design and a picture of the GridION sequencer used in our experiments. (B) (Left) Illustration of current levels of polyA and polyT regions, used in our custom level calibration scheme. Dashed orange circle indicates the region harboring the signals from chemically modified nucleotides. (Right) Region-of-interest in raw current signal obtained by identifying polyA-polyT patterns. (C) Neural network model used for classification. The 1D residual neural network architecture comprises nine 1D convolution blocks. For example, a 1D convolution block (1 × 8 conv, 64) indicates that the kernel size for the convolution is 1 × 8 and that the number of output channels is 64. Half-downsampling for each channel is denoted by (/2); averaging over all channels to arrive at a single vector is referred to as "average pooling"; the (fc 128 × 30) notation indicates a fully connected layer with the shape 128 × 30. (Right) Magnified view of the operation of 1D convolutional neural networks on time-series data. (D) (Top) Confusion matrix for 66 classes, all of which have roughly the same number of samples (subsampled to ∼3500 sample oligos in each class). Random guessing would lead to a classification accuracy of 1.52%, whereas the smallest accuracy from our model is 41% (tetramer 2252). For our model-based prediction, the mean classification accuracy is 60.28% ± 0.28% (39× larger than random guessing), and the highest observed accuracy is 79% (tetramer 1111). The exact number of samples in each class is listed in Table S4. (Bottom left) Confusion matrix for six selected classes using B2 and B4 (named as listed, subsampled to roughly 5000 samples per class). Random guessing leads to an accuracy of 16.67%, whereas our model-based prediction ensures an average classification accuracy of 72.25% ± 1.46%. (Bottom right) Confusion matrix for six selected classes using B4 and B5 (named as listed, subsampled to roughly 5000 samples per class). Random guessing leads to an accuracy of 16.67%, while our model-based prediction ensures an average accuracy of 77.84% ± 0.96%.

In theory, sequence context allows for high-resolution readout of arbitrary combinations and arrangements of natural and modified nucleotides (A, C, G, T, B1−B7). Although specific sets of tetramers might be confused during MspA reading, the method of shift reconciliation[41] allows for such sequences to be fully resolved using the information provided by different shifts of the tetramers within the constriction of the nanopore (Figure 2E). The concept of shift reconciliation is illustrated with the following example, where we consider a heterogeneous sequence of 23223. In terms of the corresponding residual current levels, the prefix tetramer 2322 is confusable with 2332 or 2323 at 150 mV. However, by shifting the sliding window one position to the right, we obtain the tetramer 3223 which is not confusable with any other block. Because the trimer prefix of 3223, 322, only matches the trimer suffix of only one of the tetramers 2322, 2332, 2322 (i.e., the first one), we unambiguously deduce that 2322 is the correct prefix tetramer.

Moving beyond tetramer detection via MspA, we demonstrate that commercially available nanopore-based sequencing technology (ONT GridION) can be used to classify/sequence oligos containing the proposed molecular alphabet. For GridION experiments, the same ssDNA oligos used in MspA experiments were extended at the 3′ terminus with a polyA tail of random length of >100 nts, which is used to increase the length of the oligos and guide them inside the pore (Figure 3A). We retrieved raw current signals from the GridION platform following a custom RNA sequencing protocol (methods section). We processed the raw current signals using deep learning techniques to discriminate and identify

https://doi.org/10.1021/acs.nanolett.1c04203
Nano Lett. 2022, 22, 1905−1914

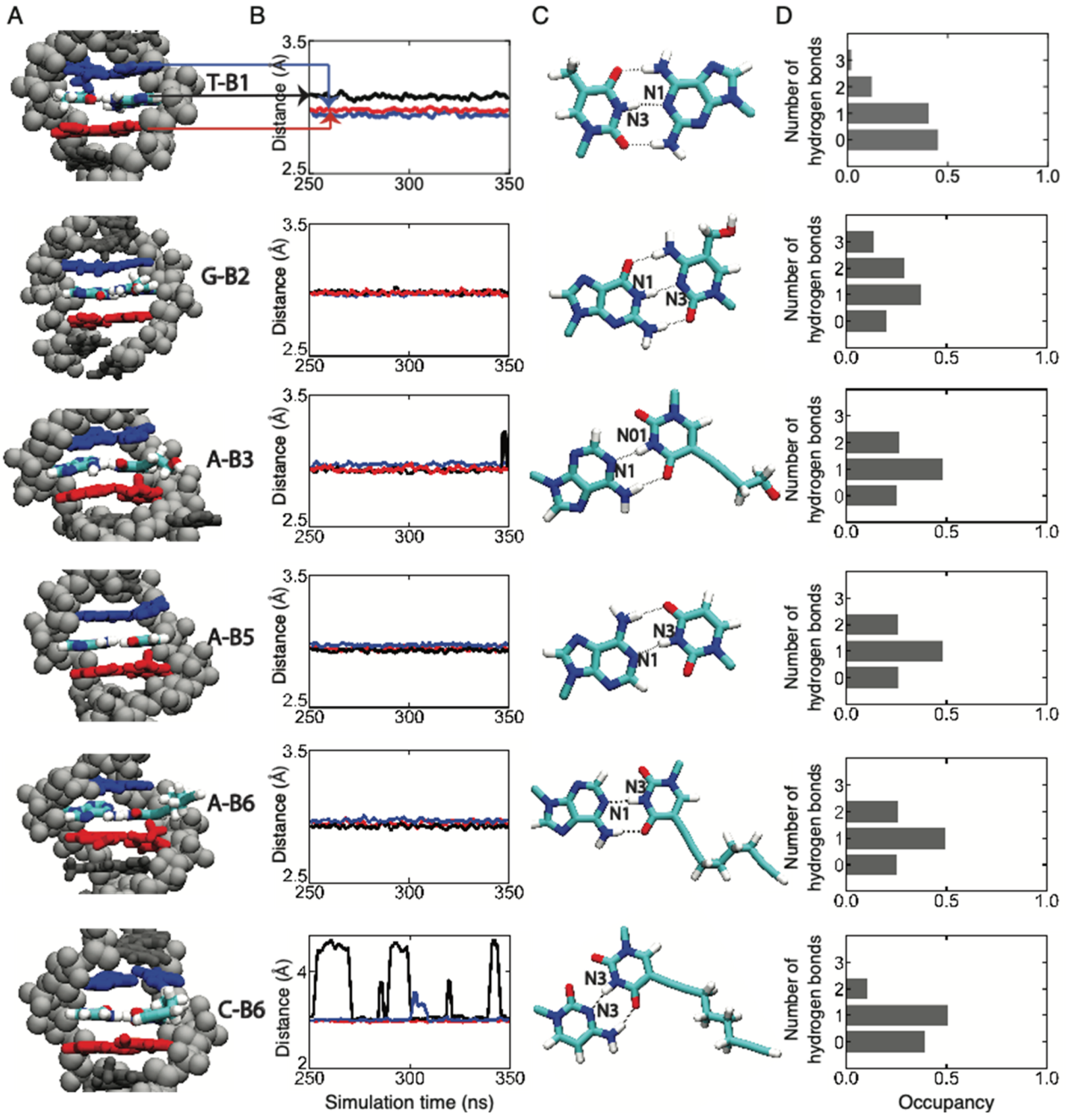


Figure 4. Stability of DNA duplexes containing chemically modified nucleotides. The backbone of the dodecamer is shown using silver spheres, whereas the bases are drawn as molecular bonds. Chemically modified bases and the natural bases that pair with them are colored according to the atom type (cyan for carbon, blue for nitrogen, and red for oxygen). Base pairs immediately adjacent to the modified base pair are colored in red or blue. (A) Microscopic configurations of modified base pairs (from top to bottom: B1−T, B2−G, A−B3, A−B5, A−B6, and C−B6). (B) Donor (N1)−acceptor (N3) distance (black) in the modified base pair (black) and in the adjacent base pairs (red and blue) during the last 100 ns of the 350 ns MD simulation. The arrows indicate the correspondence between the base pairs and the curves. The curves show a running average of the 10 ps sampled data with a 2 ns averaging window. (C) Microscopic configuration of modified base pairs. The black lines represent hydrogen bonds. The donor and the acceptor are labeled asides the atoms. (D) Probability of observing the specified number of hydrogen bonds within a modified base pair. The H-bonding probabilities were computed using the final 100 ns of a 350 ns all-atom MD simulation of a DNA dodecamer.

different combinations and orderings of the chemically modified nucleotides. As a first step, we isolated regions in the raw current signals corresponding to chemically modified nucleotides. For this purpose, we could not use the specialized software suite Tombo,[42] designed by ONT for identifying potentially modified nucleotides from nanopore sequencing data, as it requires basecalling, alignment, and further downstream processing. Accurate basecalling of chemically modified nucleotides is difficult to accomplish, which greatly complicates alignment and classification tasks for arbitrary subregions of the signal. Moreover, the most recent ONT basecaller, Bonito, based on convolutional neural networks, is trained and specialized to work for natural DNA only.[43] For these reasons, we developed an analysis framework that directly operates on raw current signals of the chemically modified nucleotides.

https://doi.org/10.1021/acs.nanolett.1c04203
Nano Lett. 2022, 22, 1905−1914

Analysis of raw current signals is challenging because nanopore current signals exhibit extreme variations known as level drifts (Figure S3). Level drifts arise because each membrane patch (recording channel) inside the device has its own electric circuit, and each pore has unique features. To address this challenge, we developed a two-step identification scheme depicted in Figure 3B. In the first step, we estimate the current level for the polyA region and subsequently use it for signal calibration. Similar calibration steps are standardly performed for nanopore sequencing of natural DNA, but they rely on adaptor-based calibrations since all analytes use identical adaptors with a well-defined sequence content. For actual level calibration, we used kernel density estimation of the signal level distribution,[44] followed by identification of the levels that have the two largest probabilities in the estimated distribution. This approach is justified because polyA regions constitute the longest signal component in our oligo sequences. Moreover, on average, polyT levels are expected to be lower than polyA levels, so readout regions that are trailed by nearly flat regions with a mean level value lower than that for the polyA tails are filtered using a finite state machine.[45] These regions are expected to bear signals from the chemically modified nucleotides. After extracting modification-bearing signals, raw current readouts are subsequently classified. For this task, we designed a 1D residual neural network model[46,47] (Figure 3C) containing 1D convolution layers (conv) that serve as feature extractors and one fully connected layer (fc) that serves as a classifier. The model is trained on oligo data corresponding to different combinations and orderings of chemically modified nucleotides, with each option supported by thousands of training samples (Table S4). Elements from each class are uniformly sampled at random in a balanced manner and split into training/validation/test sets with splitting percentages 60%/20%/20%, respectively.

Results from neural-network-guided identification tasks pertaining to five independent experimental runs are shown in Figure 3D. Confusion matrices are used to summarize the prediction accuracies, ranging between 0 and 1 (with 1 corresponding to perfectly accurate identification). Importantly, these results show that most tetramers are identified with high accuracy (i.e., the diagonal elements are significantly larger than the off-diagonal elements). The average classification accuracy for each model is provided in the caption of Figure 3D, along with the accuracy one would expect from random guessing. For example, we observed an accuracy of 0.85 for heterotetramers (2244, 2244), which is to be interpreted as an 85% success rate in correctly identifying the sequence 2244, or a 15% chance of misinterpreting 2244 as another combination or sequence order (Figure 3D). Overall, we performed a total of 13 different classification tasks, including one task for all classes (77 in total, from which only 66 were depicted due to small amounts of training data for the remaining 11 classes). We further included 12 tasks involving subsets of classes containing chemically modified nucleotides shown in Figure 1. For brevity, two results for 2X + 2Y classes and a summary of all results are shown in Figure 3D; the full set of results are shown in Figure S4.

Stable bonding of chemically modified nucleotides within a DNA double helix is important for DNA-based storage because it enables durable preservation of recorded information, as well as random access to the stored data by means of PCR reactions.[4] To better understand the interactions between chemically modified and natural nucleotides, we investigated the stability of modified DNA duplexes by carrying out all-atom molecular dynamics (MD) simulations of the Dickerson dodecamers[48] containing a pair of chemically modified nucleotides. Out of many possible variants, we chose to investigate the stability of B1−T, B2−G, B3−A, and B5−A base pairs, as suggested by prior publications[29−35] and Integrated DNA Technologies (IDT),[28] as well as the pairing of B4 and B6 with all four types of natural nucleotides. Each modified dodecamer was solvated in electrolyte solution and simulated for approximately 350 ns. Six modified natural base pairs (B1−T, B2−G, B3−A, B5−A, B6−A, and B6−C) were found to form stable hydrogen bond patterns within the duplex forming either two or three hydrogen bonds per base pairs (Figure 4). The average number of hydrogen bonds was found to be 0.71 for B1−T, 1.37 for B2−G, 1.01 for B3−A, 1.00 for B5−A, 1.00 for B6−A, and 0.70 for B6−C, which are results compatible with the numbers computed for the canonical base pairs (0.83 for A−T and 1.23 for C−G) using the same hydrogen bond criteria. In all other modified natural combinations, we observed local disruptions of the base pairing structure (Figures S5 and S6). In B4−A and B4−T pairs, the bases were observed to protrude out from the duplex without disrupting the hydrogen bonding of the surrounding base pairs. The B6−G pair formed a base stacking pattern, forcing the breakage of hydrogen bonds in the adjacent base pairs. Local unraveling of the duplex structure was observed in the systems containing B4−G, B4−C, and B6−T base pairs. On the basis of these results, we conclude that most of our chemically modified nucleotides introduce minor perturbations to the structure of the duplex except for B4, which does not fit well within the geometry of the classical DNA duplex but is not sufficient to produce a complete unraveling of the DNA duplex. However, we observed that an isolated B4−G base pair is able to maintain stable stacking interaction when simulated under conditions that mimic the presence of a longer DNA strand (Figure S6).

CONCLUSION

In closing, we report an expanded alphabet for DNA data storage compatible with nanopore sequencing technology. The unique feature of our approach is coupled, iterative selection and testing that involve determining suitability for forming stable duplex structures and nanopore sequencing. Overall, the described system enables the recording of digital data with increased storage density and more bits per synthesis cycle. In particular, our storage system enables a maximum recording density of $\log_2 11$ bits in each cycle, compared to $\log_2 4 = 2$ bits for natural DNA; this strategy also theoretically increases the rate (speed) of the recorder by $\frac{\log_2 11}{\log_2 4} = 1.73$-fold. Our extensive nanopore experiments provide strong evidence that many more chemically modified nucleotides can be used for molecular storage because many ionic current levels remain available; i.e., the ionic current spectrum is sparsely populated. In addition, our system allows for high-fidelity readouts and potentially enables PCR-based random-access for encodings restricted to duplex formation competent monomers. An illustrative, yet limited example of PCR-based random access is provided in Figure S7. Although not all pairings of chemical modifications may be suitable for amplification using natural enzymes, and some duplex formations may be unstable, the proposed system provides the first example of a coupled coding alphabet and channel selection and optimization paradigm. In

https://doi.org/10.1021/acs.nanolett.1c04203
Nano Lett. 2022, 22, 1905−1914

conclusion, this work demonstrates fundamentally new directions in molecular storage that hold the potential to advance the field of DNA-based data storage.

MATERIALS AND METHODS

Complete details of methods and materials used in this study are provided in the Supporting Information.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.nanolett.1c04203.

Details on materials and methods including the oligo design, MspA nanopore purification and experiments, ONT sequencing protocols, and MD simulations setup; additional data and notes regarding the MspA experiments, ONT readout processing workflow, and MD simulations; Figures S1−S7 and Tables S1−S5 containing supplementary results obtained in this work; Figures S8−S84 providing electrospray ionization mass spectrometry (ESI-MS) plots for all synthetic DNA oligos used in our experiments (PDF)

AUTHOR INFORMATION

Corresponding Authors

Aleksei Aksimentiev − *Center for Biophysics and Quantitative Biology, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States; Department of Physics, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-6042-8442

Min Chen − *Department of Chemistry, University of Massachusetts at Amherst, Amherst, Massachusetts 01003, United States*

Charles M. Schroeder − *Center for Biophysics and Quantitative Biology, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States; Beckman Institute for Advanced Science and Technology, Department of Materials Science and Engineering, and Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0001-6023-2274

Olgica Milenkovic − *Department of Electrical and Computer Engineering, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-1871-4912

Authors

S. Kasra Tabatabaei − *Center for Biophysics and Quantitative Biology, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States; Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0001-9369-5873

Bach Pham − *Department of Chemistry, University of Massachusetts at Amherst, Amherst, Massachusetts 01003, United States*

Chao Pan − *Department of Electrical and Computer Engineering, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-9275-7072

Jingqian Liu − *Center for Biophysics and Quantitative Biology, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States*

Shubham Chandak − *Department of Electrical Engineering, Stanford University, Stanford, California 94305, United States*

Spencer A. Shorkey − *Department of Chemistry, University of Massachusetts at Amherst, Amherst, Massachusetts 01003, United States;* orcid.org/0000-0002-7502-6980

Alvaro G. Hernandez − *Roy J. Carver Biotechnology Center, University of Illinois at Urbana—Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-6320-1550

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.nanolett.1c04203

Author Contributions

∞S.K.T., B.P., C.P., J.L., and O.M. contributed equally. O.M. and C.M.S. conceived the idea of using the expanded DNA alphabet for data storage. S.K.T., O.M., C.M.S., M.C., and B.P. designed the ssDNA oligos containing the modified nucleotides. M.C., B.P., and S.A.S. designed and performed the MspA nanopore experiments. A.G.H. designed and performed the ONT sequencing experiments. C.P., O.M., S.C., and A.G.H. performed the ONT raw output data analysis. J.L. and A.A. designed and performed the MD simulations. All authors contributed toward the system development and participated in the writing of the manuscript.

Notes

The authors declare the following competing financial interest(s): University of Illinois is filing a patent based on this work on behalf of the authors.

ACKNOWLEDGMENTS

The work was funded by the NSF+SRC SemiSynBio program under Agreement Number 1807526 and NSF Grants 1618366 and 2008125. A.A. and M.C. acknowledge support from NHGRI/NIH via Grant R21-HG011741 and NIH Grant R01 GM115442 (M.C.). The supercomputer time was provided by the University of Illinois at the Blue Waters Petascale System. The authors gratefully acknowledge many useful discussions with Profs. Jean-Pierre Leburton and Xiuling Li, as well as Nagendra Athreya and Apratim Khadelwal and Dr. Bo Li.

REFERENCES

(1) Church, G. M.; Gao, Y.; Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science* **2012**, *337* (6102), 1628−1628.

(2) Goldman, N.; Bertone, P.; Chen, S.; Dessimoz, C.; LeProust, E. M.; Sipos, B. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **2013**, *494* (7435), 77−80.

(3) Yazdi, S. M. H. T.; Gabrys, R.; Milenkovic, O. Portable and Error-Free DNA-Based Data Storage. *Sci. Rep* **2017**, *7* (1), 5011.

(4) Tabatabaei Yazdi, S. M. H.; Yuan, Y.; Ma, J.; Zhao, H.; Milenkovic, O. A Rewritable, Random-Access DNA-Based Storage System. *Sci. Rep* **2015**, *5* (1), 14138.

(5) Grass, R. N.; Heckel, R.; Puddu, M.; Paunescu, D.; Stark, W. J. Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angew. Chem. Int. Ed* **2015**, *54* (8), 2552−5.

(6) Zhirnov, V.; Zadegan, R. M.; Sandhu, G. S.; Church, G. M.; Hughes, W. L. Nucleic acid memory. *Nat. Mater.* **2016**, *15* (4), 366−70.

(7) Milenkovic, O.; Gabrys, R.; Kiah, H. M.; Tabatabaei Yazdi, S. M. H. Exabytes in a Test Tube. *IEEE Spectrum* **2018**, *55* (5), 40−5.
(8) Yazdi, S. M. H. T.; Kiah, H. M.; Garcia-Ruiz, E.; Ma, J.; Zhao, H.; Milenkovic, O. DNA-Based Storage: Trends and Methods. *IEEE Trans Mol. Biol. Multi-Scale Commun.* **2015**, *1* (3), 230−48.
(9) Erlich, Y.; Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **2017**, *355* (6328), 950−4.
(10) Organick, L.; Ang, S. D.; Chen, Y.-J.; Lopez, R.; Yekhanin, S.; Makarychev, K.; et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **2018**, *36* (3), 242−8.
(11) Biolytic DNA RNA High Throughput Oligo Synthesizer. https://www.biolytic.com/t-dna-rna-oligo-synthesizer-768xlc.aspx (accessed February 9, 2022).
(12) IDT oPools Oligo Pools. https://www.idtdna.com/pages/products/custom-dna-rna/dna-oligos/custom-dna-oligos/opools-oligo-pools (accessed February 9, 2022).
(13) Fan, J.; Han, F.; Liu, H. Challenges of Big Data analysis. *Nat. Sci. Rev.* **2014**, *1* (2), 293−314.
(14) Hoshika, S.; Leal, N. A.; Kim, M.-J.; Kim, M.-S.; Karalkar, N. B.; Kim, H.-J.; et al. Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science* **2019**, *363* (6429), 884−7.
(15) Cao, C.; Krapp, L. F.; Al Ouahabi, A.; König, N. F.; Cirauqui, N.; Radenovic, A.; et al. Aerolysin nanopores decode digital information stored in tailored macromolecular analytes. *Sci. Adv.* **2020**, *6* (50), No. eabc2661.
(16) Ledbetter, M. P.; Craig, J. M.; Karadeema, R. J.; Noakes, M. T.; Kim, H. C.; Abell, S. J.; et al. Nanopore Sequencing of an Expanded Genetic Alphabet Reveals High-Fidelity Replication of a Predominantly Hydrophobic Unnatural Base Pair. *J. Am. Chem. Soc.* **2020**, *142* (5), 2110−4.
(17) Craig, J. M.; Laszlo, A. H.; Derrington, I. M.; Ross, B. C.; Brinkerhoff, H.; Nova, I. C.; et al. Direct Detection of Unnatural DNA Nucleotides dNaM and d5SICS using the MspA Nanopore. Romesberg F, editor. *PLoS One* **2015**, *10* (11), No. e0143253.
(18) Schreiber, J.; Wescoe, Z. L.; Abu-Shumays, R.; Vivian, J. T.; Baatar, B.; Karplus, K.; et al. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (47), 18910−5.
(19) Butler, T. Z.; Pavlenok, M.; Derrington, I. M.; Niederweis, M.; Gundlach, J. H. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (52), 20647−52.
(20) Derrington, I. M.; Butler, T. Z.; Collins, M. D.; Manrao, E.; Pavlenok, M.; Niederweis, M.; et al. Nanopore DNA sequencing with MspA. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (37), 16060−5.
(21) Cao, J.; Jia, W.; Zhang, J.; Xu, X.; Yan, S.; Wang, Y.; et al. Giant single molecule chemistry events observed from a tetrachloroaurate-(III) embedded Mycobacterium smegmatis porin A nanopore. *Nat. Commun.* **2019**, *10* (1), 5668.
(22) Wang, Y.; Zhao, Y.; Bollas, A.; Wang, Y.; Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **2021**, *39* (11), 1348−65.
(23) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33−8.
(24) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31* (4), 671−90.
(25) Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; MacKerell, A. D. Optimization of the CHARMM Additive Force Field for DNA: Improved Treatment of the BI/BII Conformational Equilibrium. *J. Chem. Theory Comput.* **2012**, *8* (1), 348−62.
(26) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **2020**, *153* (4), 044130.
(27) Stoddart, D.; Heron, A. J.; Mikhailova, E.; Maglia, G.; Bayley, H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (19), 7702−7707.
(28) Integrated DNA Technologies. Modified bases modifications. https://www.idtdna.com/site/Catalog/Modifications/Category/7 (accessed February 9, 2022).
(29) Kirnos, M. D.; Khudyakov, I. Y.; Alexandrushkina, N. I.; Vanyushin, B. F. 2-Aminoadenine is an adenine substituting for a base in S-2L cyanophage DNA. *Nature* **1977**, *270* (5635), 369−70.
(30) Khudyakov, I. Ya.; Kirnos, M.D.; Alexandrushkina, N.I.; Vanyushin, B.F. Cyanophage S-2L contains DNA with 2,6-diaminopurine substituted for adenine. *Virology* **1978**, *88* (1), 8−18.
(31) Szulik, M. W.; Pallan, P. S.; Nocek, B.; Voehler, M.; Banerjee, S.; Brooks, S.; et al. Differential Stabilities and Sequence-Dependent Base Pair Opening Dynamics of Watson−Crick Base Pairs with 5-Hydroxymethylcytosine, 5-Formylcytosine, or 5-Carboxylcytosine. *Biochemistry* **2015**, *54* (5), 1294−305.
(32) Gallego, J.; Loakes, D. Solution structure and dynamics of DNA duplexes containing the universal base analogues 5-nitroindole and 5-nitroindole 3-carboxamide. *Nucleic Acids Res.* **2007**, *35* (9), 2904−12.
(33) Krokan, H. E.; Drabløs, F.; Slupphaug, G. Uracil in DNA − occurrence, consequences and repair. *Oncogene* **2002**, *21* (58), 8935−48.
(34) Seela, F.; Sirivolu, V. R. DNA Containing Side Chains with Terminal Triple Bonds: Base-Pair Stability and Functionalization of Alkynylated Pyrimidines and 7-Deazapurines. *Chem. Biodiversity* **2006**, *3* (5), 509−14.
(35) Seela, F.; Sirivolu, V. R. Nucleosides and Oligonucleotides with Diynyl Side Chains: Base Pairing and Functionalization of 2′-Deoxyuridine Derivatives by the Copper(I)-Catalyzed Alkyne-Azide 'Click' Cycloaddition. *Helv. Chim. Acta* **2007**, *90* (3), 535−52.
(36) Manrao, E. A.; Derrington, I. M.; Pavlenok, M.; Niederweis, M.; Gundlach, J. H. Nucleotide discrimination with DNA immobilized in the MSPA nanopore. *PLoS One* **2011**, *6* (10), No. e25723.
(37) Laszlo, A. H.; Derrington, I. M.; Gundlach, J. H. Subangstrom Measurements of Enzyme Function Using a Biological Nanopore, SPRNT. *Methods Enzymol.* **2017**, *582*, 387−414.
(38) Manrao, E. A.; Derrington, I. M.; Laszlo, A. H.; Langford, K. W.; Hopper, M. K.; Gillgren, N.; et al. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.* **2012**, *30* (4), 349−53.
(39) Bhattacharya, S.; Yoo, J.; Aksimentiev, A. Water Mediates Recognition of DNA Sequence *via* Ionic Current Blockade in a Biological Nanopore. *ACS Nano* **2016**, *10* (4), 4644−51.
(40) Noakes, M. T.; Brinkerhoff, H.; Laszlo, A. H.; Derrington, I. M.; Langford, K. W.; Mount, J. W.; et al. Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage. *Nat. Biotechnol.* **2019**, *37* (6), 651−6.
(41) Timp, W.; Comer, J.; Aksimentiev, A. DNA Base-Calling from a Nanopore Using a Viterbi Algorithm. *Biophys. J.* **2012**, *102* (10), L37−9.
(42) Stoiber, M.; Quick, J.; Egan, R.; Eun Lee, J.; Celniker, S.; Neely, R. K.; et al. *De novo* Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv* **2016**, 094672 (accessed August 26, 2021).
(43) Bonito; A PyTorch Basecaller for Oxford Nanopore Reads. https://github.com/nanoporetech/bonito (accessed February 9, 2022).
(44) Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*, 1st ed.; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, 2015; DOI: 10.1002/9781118575574 (accessed June 22, 2021).
(45) Gill, A. *Introduction to the Theory of Finite-State Machines*; McGraw-Hill: New York, 1962.
(46) Hong, S.; Xu, Y.; Khare, A.; Priambada, S.; Maher, K.; Aljiffry, A.; et al. HOLMES: Health OnLine Model Ensemble Serving for Deep Learning Models in Intensive Care Units. *Proceedings*, 26th ACM SIGKDD International Conference on Knowledge Discovery &

https://doi.org/10.1021/acs.nanolett.1c04203
Nano Lett. 2022, 22, 1905−1914

Data Mining, Virtual Event, CA, USA; ACM, 2020; pp 1614−24, DOI: 10.1145/3394486.3403212 (accessed July 21, 2021).
(47) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *Proceedings*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA; IEEE, 2016; pp 770−8, http://ieeexplore.ieee.org/document/7780459/ (accessed July 21, 2021).
(48) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; et al. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78* (4), 2179−83.

https://doi.org/10.1021/acs.nanolett.1c04203
Nano Lett. 2022, 22, 1905−1914