Efficient molecular encoding in multifunctional self-immolative urethanes

Samuel D. Dahlhauser¹, Sarah R. Moor¹, Marissa S. Vera¹, Jordan T. York¹, Phuoc Ngo¹, Alexander J. Boley¹, Jaime N. Coronado¹, Zack B. Simpson², Eric V. Anslyn¹,³,*

¹University of Texas at Austin, Austin, TX 78712, USA
²Erisyon, Inc., 3202 Hemphill Park, Austin, TX 78705, USA
³Lead contact

SUMMARY

Molecular encoding in sequence-defined polymers shows promise as a new paradigm for data storage. Here, we report what is, to our knowledge, the first use of self-immolative oligourethanes for storing and reading encoded information. As a proof of principle, we describe how a text passage from Jane Austen’s Mansfield Park was encoded in sequence-defined oligourethanes and reconstructed via self-immolative sequencing. We develop Mol.E-coder, a software tool that uses a Huffman encoding scheme to convert the character table to hexadecimal. The oligourethanes are then generated by a high-throughput parallel synthesis. Sequencing of the oligourethanes by self-immolation is done concurrently in a parallel fashion, and the liquid chromatography-mass spectrometry (LC-MS) information decoded by our Mol.E-decoder software. The passage is capable of being reproduced wholly intact by a third-party, without any purifications or the use of tandem MS (MS/MS), despite multiple rounds of compression, encoding, and synthesis.

Graphical Abstract
Dense and cost-effective means for storing information for future use is needed as society continues to produce data exponentially. Abiotic polymers (plastics) are an exceptional platform for information storage because of their accessibility and limitless structural modifications. However, efficient, high-throughput means for “writing” on them and “reading” them are still needed. Herein, Dahlhauser et al. report the high-throughput synthesis (writing) and sequencing (reading) of urethanes.

INTRODUCTION

Synthetic sequence-defined polymers, such as peptides, peptoids, β-peptides, γ-peptides, polyureas, and polycarbamates, have seen significant advances in complexity, rivaling that of biopolymers. Much of the effort has been focused on deriving sequence-based functionality for purposes of self-assembly, supramolecular binding, and catalysis. However, information storage remains an underused function of these synthetic macromolecules. Reliable data storage is becoming one of the most significant technological challenges of the digital age. Encoding data at the molecular level could dramatically increase storage densities using earth-abundant elements (H, C, N, and O) and overcome some of the significant drawbacks encountered with conventional silicon-based data storage, such as durability and longevity. As an example, the well-known information storing biopolymer DNA is capable of storing vast amounts of information using a molecular code based on its four different monomers. Artificial DNA oligomers have been shown to store and retrieve large amounts of information. However, DNA suffers from issues regarding its
synthetic scalability and long-term stability.\textsuperscript{9} To address those short-comings, synthetic polymers/oligomers have been used for information storage,\textsuperscript{10-18} alongside multicomponent reactions,\textsuperscript{19-21} and small-molecule strategies.\textsuperscript{22-24} However, high-throughput, simple, and facile means for writing and reading in the synthetic macromolecules are still being sought.\textsuperscript{25}

Information storage within sequence-defined polymers can be achieved through a variety of creative strategies.\textsuperscript{8} Often, the sequence-defined polymers encode the information in polymeric strings of binary digits, encoded by using two variants of the monomers.\textsuperscript{10,12,26-29} Binary is one of the simplest encoding schemes and is the basis for most modern information technologies.\textsuperscript{30} A caveat to that simplicity is that longer strings of information are required to encode the information. For example, in the simple 256-character ASCII (American standard code for information interchange) table, it requires eight bits to encode one character ($2^8 = 256$). Thus, one character would be encoded by eight monomers along a polymer backbone. DNA has a theoretical maximum storage density of 2 bits/monomer (in practice, less), culminating in thousands of indexed strands varying in lengths from 100 to 1,000 monomers to achieve kilobyte storage capacities.\textsuperscript{7,29} For synthetic-sequence-defined polymers, those scales are not yet feasible. Laurent et al.\textsuperscript{31} succinctly describes the two parameters for enhancing storage capacity: (1) storage density (bits/monomer), and (2) chain length. To increase the density of synthetic digital polymers, the monomer pool (alphabet) used in the encoding must increase. Recently, numerous approaches to expand the number of functionalities along the oligomer backbone have been introduced.\textsuperscript{11,32,33} With a larger pool of building blocks, the density of information increases the base of the positional numeral system (i.e., base-2, 1 bit/monomer; versus base-8, 3 bits/monomer).\textsuperscript{34} Thus, more information can be encoded in shorter chain lengths, simplifying oligomer design and synthesis. To date, limited examples of multiply functionalized macromolecules for data storage have been reported.\textsuperscript{11,18,21,31}

The most significant challenge associated with an increased monomer pool is the increase in synthesis and sequencing complexity. Most synthetic sequence-defined oligomers are characterized by tandem mass spectrometry (MS/MS)\textsuperscript{10-12,14-17,26,31,34} or, more recently, with a mutated nanopore.\textsuperscript{35,36} In both cases, detectable variations in the monomers become difficult to elucidate and, as a result, are often limited to binary or two unique monomers per information-containing oligomer. In fact, careful examination of some of the MS/MS methods reported in previous encoding/decoding schemes reveal that the code can be difficult to decipher without prior knowledge of the structures of the molecules being used. To our knowledge, no examples of molecular encoding/decoding have been demonstrated in which the decoder does not know the structure of the molecules that hold the information or the information being encoded beforehand.

Previously, we reported a solid-phase synthesis and self-immolative sequencing paradigm.\textsuperscript{37} The strategy involves an intramolecular \textit{5-exo-trig} cyclization and elimination from the O-terminal of the oligourethanes, which iteratively removes each terminal monomer and allows the truncated oligomers to be characterized and profiled via liquid chromatography-mass spectrometry (LC-MS) (Scheme 1). As a powerful characterization tool, we aimed to develop this methodology for molecular information storage.
We report here a hexadecimal-based encoding schema using oligourethanes derived from β-amino alcohols. These particular oligourethanes can be sequenced chemically, sans purification, via the self-immolation LC-MS protocol, providing a robust and highly accurate readout. Further, parallel synthesis and sequencing methodologies were developed to afford rapid writing and reading. To demonstrate this paradigm, a passage from Jane Austen’s *Mansfield Park* was converted to a hexadecimal using our Mol.E-coder software and encoded in 18 oligourethanes. The encoded oligomers were deciphered by a third party, and the sequencing data were fed into our Mol.E-decoder software (see supplemental information), decoding the sequencing information to reveal the passage wholly intact, with 100% accuracy. We believe this to be the first use of a self-immolative sequencing platform for information encoding and decoding. The workflow of encoding via Huffman compression, parallel synthesis with a hexadecimal code (writing), parallel sequencing via LC-MS (reading), and final decoding represents one of the first platforms to effectively address both storage density and data compression in abiotic, digital oligomers, resulting in a simple and effective method for information-dense molecular storage.

RESULTS AND DISCUSSION

“Hello, World!”

To develop the self-immolative urethanes as a medium for information storage, we sought to first encode a small piece of information, in this case text, into a compressed-bit string, which could then be translated into a molecular form and, ultimately, read back. The first program written by aspiring programmers is often the well-established sanity test “Hello, World!” Huffman coding, a form of lossless data compression, was used to convert “Hello, World!” to binary (Figure 1). Huffman coding, like other entropy-encoding methods, represents the most common symbols (letters) with fewer bits, relative to the less-common symbols. Thus, encoding “Hello, World!” required only 42 bits, as opposed to the 104 bits required if a traditional ASCII table was used. From there, our bit string could be converted to any numerical-based system, such as octal or hexadecimal, as a means to further compact the information. Our encoding software, Mol.E-coder, was designed so that it could encode any and all unique symbols and is not limited to the English alphabet or ASCII-defined codes. See the supplemental experimental procedures for an explanation of the encoding algorithms, and GitHub for the encoding algorithm itself.

Molecular encoding by chemical synthesis

For this rather-short bit string, we chose to convert the information to an octal notation, which was then represented molecularly with eight unique monomers (Figure 2). Using the standard positional numeral system base conversion (as seen on an ASCII table), each octal character represents three binary digits. Thus, the 42-character bit string shrinks to a 14-character octal string “57451243036731.” Encoding that octal string at a molecular level requires that each symbol be represented by a discrete monomer along the oligomer backbone. Thus, a pool of eight unique monomers was synthesized by the reduction and deutero-reduction of commercially available canonical and non-canonical amino acids.
(Schemes S1 and S2; Figures S1 and S2). By simply deuterating the amino alcohols at the α-methylene, the mass of each monomer increases by 2 atomic mass units (AMU), effectively doubling the available monomer pool without changing the complexity of the synthesis or sequencing chemistry (a feature we take advantage of when encoding in hexadecimal). The monomers were then converted to the activated carbonate by reaction with 4-nitro-phenylchloroformate (Scheme S3). The amino alcohols used to encode the phrase are shown in Figure 2. Monomers were assigned their octal symbol according to their occurrence, so that cheaper monomers (e.g., alaninol) were assigned to the most frequently occurring symbols.

The 14-digit octal string was then written onto two oligomers (A1 and A2; Figures 3, S3, and S4), each carrying seven of the 14 digits. Previous examples have used mass tags to indicate the position of letters in words or the position of words in sentences to allow reconstruction of information in the correct order. We chose to use the physical location of the molecules in a 96-well plate to index our information, similar to how a mechanical hard-disk drive uses a physical location and a directory to store a computer’s data.

The two oligomers A1 and A2 (read from N–O termini: 4215475-Pheindex and 1376303-Pheindex, respectively) were successfully synthesized on the solid phase, in parallel, on a fritted 96-well plate, using our previously described methodology, which builds upon an original approach from Cho et al., with slight modifications to adapt it a well-plate format (Scheme S4; Notes S1 and S2). On each oligomer, the encoded information is preceded with a phenylalaninol as an indexing tool (Pheindex) to begin reading the mass spectra. This provides a reading frame for the mass spectra. A key feature is the ability to use all the created oligomers directly from the solid-phase resin without any purification. By implementing a reading frame with which to start calculating mass differences, any extraneous masses can be filtered out (Figure 3, for example). After the successful coupling of each monomer to the desired octamer, the N terminus was “capped” with 4-fluoro-7-nitrobenzofurazan (NBD-F) to act as a long-wavelength chromophore with which to monitor the self-immolation. From there, the oligomers were cleaved from the resin and directly sequenced.

Decoding the molecules

NBD-labeled oligomers A1 and A2 were sequenced in parallel in a well plate in DMSO with cesium carbonate (Cs2CO3) at 70°C and sampled for LC-MS at designated intervals (Figures 3, S27, and S35; Notes S4 and S5). As Scheme 1 shows, immolation removes each monomer from the O terminus, thus truncating the oligomers iteratively. The precursor masses of each iteration (8-mer, 7-mer, 6-mer, 5-mer, 4-mer, 3-mer, 2-mer, and 1-mer) were easily observed in the mass spectra (a spectra after 225 min of immolation is shown in Figure 3) and entered into a templated spreadsheet that is fed into our Mol.E-decoder software (Figures S28-S34 and S36-S42; see GitHub link). The mass differences between subsequent peaks were calculated by the algorithm and correlated to each monomer, which is associated with a specific octal symbol (Figure 2). The octal string is sorted according to the indices and converted back to binary by the same positional numeral-system base conversion. The bit string is finally decoded back through the Huffman algorithm to produce our original
information wholly intact with no errors (Figure 4). An in-depth analysis of the sequencing data for oligomers A1 and A2, as well as the data processing back to plain text, is given in the supplemental information (Figures S27-S43).

It is important to note that, from this sanity test, we found that the workflow, from information encoding to synthesis to sequencing and finally to decoding, integrated seamlessly. We observed that all eight monomers could be readily distinguished from one another, including the deuterated variant. Thus, mass differences among monomers can be as small as 2 AMU. The oligomers were sequenced without purification, meaning the noise accumulated after seven coupling and seven deprotection steps, one labeling step, and cleavage from the solid phase, was not significant. Further, by using the phenylalaninol reading frame, the 10% deletion observed in A1 (presumed to be from incomplete deprotection after the first coupling step) was inconsequential (Figure S43). Likewise, the overlapping peaks in the chromatograms of A2 were easily deconvoluted (Figures S37-S40).

**Encoding and decoding a passage from Jane Austen’s Mansfield Park**

With the success of our test, we sought to scale the information storage to explore its capacity and efficacy to store more-remarkable information. With seemingly infinite data to choose from, an apt but timeless quote from Jane Austen’s *Mansfield Park* was chosen: “If one scheme of happiness fails, human nature turns to another; if the first calculation is wrong, we make a second better: we find comfort somewhere.” The information, which is 153 characters long, including spaces, was converted to a 632-character bit string by the optimized Huffman encoding and converted to hexadecimal by a standard ASCII base conversion (Figure S44; supplemental experimental procedures). Remarkably, the resulting hexadecimal code was 158 characters long, only five more than the original information, which contained 26 unique characters, including punctuation.

To write our molecules in the hexadecimal numeral system, we increased our monomer pool to 16 by adding one more unique amino alcohol, as well as deuterating each one at the alpha methylene (Scheme S2; Figure S2; supplemental experimental procedures). Having now 16 unique monomers and the hexadecimal code, an encoding scheme was devised. With molecular information density being the goal, we chose to extend the oligomers to 10 in length. Thus, each oligomer would contain nine hex symbols and the index, resulting in seventeen 10-mer and a final 6-mer to encode all 158 characters. Figure 5 shows pictorially all of the information strings for the concept: oligourethanes, hex strings, bit strings, and the English-language translation.

The synthesis was performed as described earlier, by iterative coupling and deprotection in a fritted 96-well plate (oligomers G1–G12 and H1–H6, named after the wells in the plate; Note S3). All 18 oligomers were synthesized in parallel. Through the course of the 18 simultaneous syntheses, 17 of the 18 were successful with only minor deletions (discovered to be inadequate mixing of the wells during deprotection) (Figures S5-S22). Oligomer H6 was found to have suffered a significant deletion, and as such, it was resynthesized. With adequate mixing during the deprotection, no such deletions were again observed (Figure S22). All 18 oligomers were then simultaneously labeled with NBD-F and cleaved from the solid-phase resin with 0.1% trifluoroacetic acid in dichloromethane for 5 min. The
extended cleavage resulted in up to 50% trifluoroacetylation of the terminal alcohol (Figures S23-S26). However, the trifluoroacetyl ester is readily hydrolyzed under the sequencing conditions, and as such, the material can be carried through in whole without affecting the sequencing. As a control, but not for sequencing purposes, all 18 oligomers were confirmed by high-resolution mass spectroscopy (Table S1).

Oligomers G1–G12 and H1–H6 were sequenced concurrently via self-immolation in a 96-well plate in a 2:1 MeOH-to-water mixture with K₃PO₄ at 70°C in a heated shaker (Figures S45-S62). The reactions were monitored by LC-MS every 30 min for 2.5 h. Of the 176 masses (seventeen 10-mer, one 6-mer), 170 were observed clearly and distinctly. The precursor 10-mer for G8, G11, H2, and H5 did not ionize in significant quantities under the generalized low-resolution LC-MS conditions. Considering these were indexing masses, no encoded information was lost. Likewise, the 9-mer for G8 and H2 did not ionize significantly under the generalized low-resolution LC-MS conditions. Because of the chromatographic traces at 470 nm, the presence of the 10-mer and 9-mer in the samples was clearly observed, making it trivial to pinpoint the specific information that was missing (Figures S52, S55, S58, and S61). Knowing what information was missing, we easily addressed it by modifying the conditions and using a high-resolution instrument (Figure S63).

Considering future improvement of this sequencing methodology, some note-worthy trends regarding the effects of certain side chains on the sequencing reaction should be mentioned. As observed above, larger and more hydrophobic oligomers tend to ionize poorly. Removal of hydrophobic monomers like cyclohexylalaninol would help to address this problem. Further, oligomers in which valinol was at the O-terminal did not accumulate in significant quantities relative to other oligomers (Figure S64). Presumably, the isopropyl side chain of the valine-derived amino alcohol enforces greater steric compression by having a gem-dimethyl methine beta to the alcohol, rather than the methylene or methyl seen in each of the other side chains. Steric compression is a significant driving force in the 5-exo-trig cyclization, resulting in an enhanced rate of cyclization.45 Thus, oligomers with terminal valinols have a shorter lifetime in solution. Second, as expected, longer oligomers tend to have smaller changes in retention time per cyclization event, which can result in peak coelution. This is prominent when the terminal or cyclizing monomer is alaninol, likely because of its small side chain having little effect on the overall polarity of the macromolecule (Figure S65). Because the masses are observed, this is not seen as a detriment to the sequencing but is worthy of note for future designs.

The mass information from the LC-MS was fed into our Mol.E-decoder software (Figure S66). The algorithm assigned the hexadecimal symbol to the mass differences and sorted the hex string. The hex string was converted back to binary and decoded using the Huffman algorithm, returning the Jane Austen quote with no errors. As an important and unique verification of the sequencing methodology, and to truly explore its robustness, a third-party validation study was performed in which a co-author (J.N.C.), unaffiliated with the project, was given a set of instructions to “read” the molecule medium (see supplemental experimental procedures). After one pass, the participant was able to correctly decipher 156 of the 158 information-containing monomers (Figure S67). The two incorrect masses were
attributed to a reading error when interpreting the mass spectra (a deletion was incorrectly chosen; see Figure S68 and Note S6). The participant was given a slightly modified second set of instructions without being told the mistake made in the first attempt and was able to correctly decipher all 158 information-encoding monomers (supplemental experimental procedures; Figure S69). We were delighted at the success of the third-party validation study and believe it lends promise to the robustness of the sequencing platform as well as to our plans to automate the decoding of mass spectrometry data in future studies.

Herein, we have described a complete workflow for the encoding and decoding of information in self-immolative, macromolecular, sequence-defined oligourethanes. The data compression and encoding scheme use a highly configurable software, Mol.E-coder, which is presumably able to take any information, regardless of the language or symbol table it comprises, and compress that information into a universally recognized hexadecimal representation (available open source on GitHub). The “writing” process is incredibly robust, efficient, and amenable to high throughput. As proof of this principle, we synthesized 18 oligomers in parallel, demonstrating the ease with which this could be scaled. The molecular encoding uses a diverse and cheap feedstock of chemicals, being derived from commercially available amino acids. Further, the pool of monomers was effectively doubled by simple deutero-enrichment. The information is easily read using a simple sequencing process, requiring only base, heat, and an LC-MS system. The mass spectrometry data are fed into our Mol.E-decoder software, which then assigns the hex symbol to each monomer and decodes the information back to its original state (also available open source on GitHub). We invite other groups to modify and use the software for their own needs.

Although the field of abiotic sequence-defined polymers requires more advances to be able to rival the automation and parallel-sequencing capabilities inherent in the current use of nucleic acids, this work represents another significant advance toward such a goal. Information storage aside, demonstrated herein is the high-throughput synthesis of chiral abiotic oligomers as well as the high-throughput sequencing and characterization of complex macromolecules. Future research will look into the bottlenecks to scale these molecules for more-advanced information content, for example, error rates, length limitations on synthesis/sequencing, and the speed and throughput of the writing and reading. Further, we plan to use the rapid-characterization techniques presented here to explore their applications as sequence-defined oligomers, i.e., self-assembly, combinatorial chemistry, and catalysis.

**EXPERIMENTAL PROCEDURES**

**Resource availability**

**Lead contact**—Further information and requests for resources should be directed to, and will be fulfilled by, the lead contact, Eric V. Anslyn (anslyn@austin.utexas.edu).

**Materials availability**—Materials generated in this study are available by request from the lead contact.

**Data and code availability**—The authors declare that the data supporting the findings in the study are available within the article and the supplemental information. All other data are
available from the lead contact upon reasonable request. The software used in this study is open-source and available to the public at https://github.com/PhysicalOrganic/Mol.E-coder. This study did not generate any datasets.

**General procedure for reduction of Fmoc amino acids**

To a stirred solution of Fmoc-L-amino acid 1.0 equivalent in anhydrous THF (3.3 mL) was added 1,1′-carbonyldiimidazole (1.33 equivalent) at room temperature. The reaction was stirred for at least 10 min and then cooled to 0°C. Next, a solution of NaBH₄ (1.66 equivalent) in H₂O (1.66 mL or 0.6 M) was added. The solution was stirred for at least 30 min and up to 1.5 h. The reaction was quenched with the addition of 1 M HCl and extracted with EtOAc (3×). The combined organics were washed one time with brine, dried over Na₂SO₄, and concentrated under vacuum.

**General procedure for deutero-reduction of Fmoc amino acids**

To a stirred solution of Fmoc-L-amino acid 1.0 equivalent in anhydrous THF (3.3 mL) was added 1,1′-carbonyldiimidazole (1.33 equivalent) at room temperature. The reaction was stirred for at least 10 min and then cooled to 0°C. Next, a solution of NaBD₄ (1.66 equivalent) in D₂O (1.66 mL or 0.6 M) was added. The solution was stirred for at least 30 min and up to 1.5 h. The reaction was quenched by the addition of 1 M HCl and extracted with EtOAc (3×). The combined organics were washed one time with brine, dried over Na₂SO₄, and concentrated under vacuum.

**General procedure for synthesis of activated carbonates**

To a stirring solution of Fmoc-Amino alcohol (1.0 equivalent) in anhydrous dichloromethane (DCM; 0.2 M) was added pyridine (1.3 equivalent) drop-wise. Next, 4-nitro-phenyl chloroformate (1.5 equivalent) was added, and the reaction was left to stir overnight. Reaction was monitored by thin-layer chromatography (TLC) and, upon consumption of the starting material, was diluted excessively in DCM and transferred to a separatory funnel. The organic layer was washed with 1 M NaHSO₄ (3×), then 1 M Na₂CO₃ (5×, or until it stopped turning bright yellow), and finally brine. The organic layer was dried over Na₂SO₄, filtered, and concentrated in vacuo. The product was purified by silica gel chromatography.

**General procedure for parallel synthesis in a fritted deep-well 96-well plate coupling**

Phenylalaninol loaded (0.44 mmol/g, 200–400 mesh) 2-chlorotrypt polystyrene resin (15 mg, 0.0075 mmol) was added to 18 wells (G1–G12, H1–H6) of a 10-μm fritted deep-well 96-well plate (combinatorial microlute by Porvair Sciences). A 96-well microplate deep-well drain-cap mat (catalog no. 219005, Porvair) was used to seal the base of the filtration plate. The resin was suspended in 0.2 mL of anhydrous N-methyl-2-pyrrolidinone (NMP) and swollen. A stock solution of Hunig’s base (36 μL total, 2 μL or 0.01125 mmol per well) and hydroxybenzotriazole (110 mg total, 6.1 mg or 0.045 mmol per well) was made in NMP (1.8 mL total, 0.1 mL per well). The plate was agitated to mix the contents of the wells. Finally, the activated amino alcohols Fmoc-XXX-p-nitrophenyl carbonate (PNOC) (0.022 mmol per well) were dissolved in NMP (0.2 mL per well) in separate vials. Once dissolved, the monomers were transferred to their corresponding wells. The plate was sealed with Nunc Cell Rep Phys Sci. Author manuscript; available in PMC 2021 November 08.
aluminum sealing tape (no. 276014) and shaken overnight. Note: proper mixing of the wells is important for efficient coupling, and previous reports show the coupling to be complete in as little as 4 h.\textsuperscript{37}

Resin in all 18 wells was washed with NMP (5 × 2 mL), then DCM (5 × 2 mL), and finally Et\textsubscript{2}O (3 × 2 mL) by vacuum filtration through the fritted plate (Supelco PlatePrep 96-well vacuum manifold kit). Resin was dried overnight under vacuum. Test cleavages were effected with 0.1% trifluoroacetic acid (TFA) in DCM (5 × 1 mL) for 10 s each. Coupling efficiency was checked by LC-MS. Note: depending on the cleavage times and amounts of TFA used, the trifluoroacetic ester was observed by LC-MS. This ester was readily hydrolyzed by dissolving the sample in DCM and shaking with saturated NaHCO\textsubscript{3}.

### Deprotection

Resin loaded with terminal Fmoc-protected oligocarbamates (15 mg per well) was suspended in 20% piperidine in DMF (0.5 mL) and shaken for 2 h. Note: because of shorter reaction time, sufficient mixing of the well is required. Incomplete deprotections were observed with inadequate mixing. Cleavage of the dibenzofulvene-piperidine adduct can be quantified at 301 nM using Beer’s law. The resin was washed with DMF (5 × 2 mL), DCM (5 × 2 mL), and Et\textsubscript{2}O (3 × 2 mL) and dried overnight under vacuum.

### General procedure for parallel labeling of the terminal amine with NBD-fluoride

In the fritted 96-well plate, to each of the 18 wells containing 0.0075 mmol oligomer-loaded resin was added a solution of Hunig’s base (13 μL per well, 0.075 mmol), 4-fluoro-7-nitrobenzofurazan (6.9 mg or 0.0375 mmol per well) in anhydrous DMF (0.5 mL per well). The plate was sealed, and the reactions were left to shake overnight. Resin was washed with DMF (5 × 2 mL), DCM (10 × 2 mL or until no more yellow/green was observed in the wash), and Et\textsubscript{2}O (3 × 5 mL).

### General procedure for cleavage of oligomers from the resin

Cleavages were effected with 0.1% TFA in DCM (3 × 2 mL) for 5 min each. Resin was filtered for removal, and the cleaved product was concentrated in a clean 96-well plate. Note: depending on the cleavage times and amounts of TFA used, the trifluoroacetic ester of the terminal alcohol was observed by LC-MS. This ester was readily hydrolyzed in the proceeding sequencing methodology and, thus, ignored but is present in all initial LC-MS chromatograms.

### Sequencing of oligomers A1 and A2

The oligomers (measured to a final concentration between 1 and 2 mM) were dissolved in DMSO and added to a 96-well plate. Next, cesium carbonate was added to the solution. The final concentration of base was approximately 5 mM. The reaction was heated to 70°C and held at the temperature for 45 min. Reaction was sampled for LC-MS by taking 50 μL of the reaction and diluting it into 75 mL methanol. This was repeated at 105, 165, 225, and 300 min.
Sequencing of oligomers G1-G12 and H1-H6

The oligomers (measured to a final concentration between 1 and 2 mM) were dissolved in a 2:1 ratio of MeOH to H₂O and added to a 96-well plate. Next, potassium phosphate (K₃PO₄) was added to the solution. The final concentration of base was approximately 75 mM. The reaction was heated to 70°C and held at that temperature for 30 min. Reaction was sampled for LC-MS by taking 62.5 μL of the reaction and diluting it into 75 μL of 1:1 MeOH to H₂O. This was repeated at 60, 90, 120, and 150 min.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

With great appreciation, we acknowledge the financial support for this work from the Army Research Office (W911NF-17-1-0522), the Howard Hughes Medical Institute (GT10481), and the Welch Reagents Chair (F-0046) to E.V.A. We gratefully acknowledge Ian Riddington and Kristin Blake at the UT Mass Spectrometry Facility for their instrumental help throughout this project. Likewise, for the UT NMR facilities for Bruker AVANCE III 500: NIH grant no. 1 S10 OD021508-01.

REFERENCES

1. Solleder SC, Schneider RV, Wetzel KS, Boukis AC, and Meier MAR (2017). Recent progress in the design of monodisperse, sequence-defined macromolecules. Macromol. Rapid Commun 38, 1600711.
2. Hill DJ, Mio MJ, Prince RB, Hughes TS, and Moore JS (2001). A field guide to foldamers. Chem. Rev 101, 3893–4012. [PubMed: 11740924]
3. Martinek TA, and Fülöp F (2012). Peptidic foldamers: ramping up diversity. Chem. Soc. Rev 41, 687–702. [PubMed: 21769415]
4. Colquhoun H, and Lutz J-F (2014). Information-containing macromolecules. Nat. Chem 6, 455–456. [PubMed: 24848219]
5. Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, and Segal E (2012). Genomics: ENCODE explained. Nature 489, 52–55. [PubMed: 22955614]
6. Church GM, Gao Y, and Kosuri S (2012). Next-generation digital information storage in DNA. Science 337, 1628–1628. [PubMed: 22903519]
7. Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, and Birney E (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494, 77–80. [PubMed: 23354052]
8. Rutten MGTA, Vaandrager FW, Elemans JAAW, and Nolte RJM (2018). Encoding information into polymers. Nat. Rev. Chem 2, 365–381.
9. Röder B, Frühwirth K, Vogl C, Wagner M, and Rossmanith P (2010). Impact of long-term storage on stability of standard DNA for nucleic acid-based methods. J. Clin. Microbiol 48, 4260–4262. [PubMed: 20810770]
10. Roy RK, Meszynska A, Laure C, Charles L, Verchin C, and Lutz J-F (2015). Design and synthesis of digitally encoded polymers that can be decoded and erased. Nat. Commun 6, 7237. [PubMed: 26006165]
11. Martens S, Landuyt A, Espeel P, Devreeb B, Dawyndt P, and Du Prez F (2018). Multifunctional sequence-defined macromolecules for chemical data storage. Nat. Commun 9, 4451. [PubMed: 30367037]
12. Gunay US, Petit BE, Karamessini D, Al Ouahabi A, Amalian J-A, Chendo C, Bouquey M, Gigmes D, Charles L, and Lutz J-F (2016). Chemoselective synthesis of uniform sequence-coded polyurethanes and their use as molecular tags. Chem 1, 114–126.

Cell Rep Phys Sci. Author manuscript; available in PMC 2021 November 08.
13. Al Ouahabi A, Charles L, and Lutz J-F (2015). Synthesis of non-natural sequence-encoded polymers using phosphoramidite chemistry. J. Am. Chem. Soc 137, 5629–5635. [PubMed: 25851514]

14. Cavallo G, Al Ouahabi A, Oswald L, Charles L, and Lutz J-F (2016). Orthogonal synthesis of “easy-to-read” information-containing polymers using phosphoramidite and radical coupling steps. J. Am. Chem. Soc 138, 5629–5635. [PubMed: 27454229]

15. Ding K, Zhang Y, Huang Z, Liu B, Shi Q, Hu L, Zhou N, Zhang Z, and Zhu X (2019). Easily encodable/decodable digital polymers linked by dithiosuccinimide motif. Eur. Polym. J 119, 421–425.

16. Liu B, Shi Q, Hu L, Huang Z, Zhu X, and Zhang Z (2020). Engineering digital polymer based on thiol-maleimide Michael coupling toward effective writing and reading. Polym. Chem 11, 1702–1707.

17. Lee JM, Koo MB, Lee SW, Lee H, Kwon J, Shim YH, Kim SY, and Kim KT (2020). High-density information storage in an absolutely defined aperiodic sequence of monodisperse copolyester. Nat. Commun 11, 56. [PubMed: 31911612]

18. Leguizamon SC, and Scott TF (2020). Sequence-selective dynamic covalent assembly of information-bearing oligomers. Nat. Commun 11, 784. [PubMed: 32034159]

19. Arcadia CE, Kennedy E, Geiser J, Dombroski A, Oakley K, Chen S-L, Sprague L, Ozmen M, Sello J, Weber PM, et al. (2020). Multicomponent molecular memory. Nat. Commun 11, 691. [PubMed: 32019933]

20. Boukis AC, Reiter K, Frölich M, Hofheinz D, and Meier MAR (2018). Multicomponent reactions provide key molecules for secret communication. Nat. Commun 9, 1439. [PubMed: 29651145]

21. Boukis AC, and Meier MAR (2018). Data storage in sequence-defined macromolecules via multicomponent reactions. Eur. Polym. J 104, 32–38.

22. Cafferty BJ, Ten AS, Fink MJ, Morey S, Preston DJ, Mrksich M, and Whitesides GM (2019). Storage of information using small organic molecules. ACS Cent. Sci 5, 911–916. [PubMed: 31139727]

23. Sarkar T, Selvakumar K, Motiei L, and Margulies D (2016). Message in a molecule. Nat. Commun 7, 11374. [PubMed: 27138465]

24. La Clair JJ (2018). Encoding matter with regiospecific $^{12}$C/$^{13}$C isotopic labels. Chem. Commun. (Camb.) 54, 2611–2614. [PubMed: 29417122]

25. Mutlu H, and Lutz J-F (2014). Reading polymers: sequencing of natural and synthetic macromolecules. Angew. Chem. Int. Ed. Engl 53, 13010–13019. [PubMed: 25283068]

26. Charles L, Laure C, Lutz J-F, and Roy RK (2015). MS/MS sequencing of digitally encoded poly(alkoxyamine amide)s. Macromolecules 48, 4319–4328.

27. Karamessini D, Poyer S, Charles L, and Lutz J-F (2017). 2D sequence-coded oligourethane barcodes for plastic materials labeling. Macromol. Rapid Commun 38, 1700426.

28. Cavallo G, Poyer S, Amalian J-A, Dufour F, Burel A, Carapito C, Charles L, and Lutz J-F (2018). Cleavable binary dyads: simplifying data extraction and increasing storage density in digital polymers. Angew. Chem. Int. Ed. Engl 57, 6266–6269. [PubMed: 29633445]

29. Ceze L, Nivala J, and Strauss K (2019). Molecular digital data storage using DNA. Nat. Rev. Genet 20, 456–466. [PubMed: 31068682]

30. Leibniz GW (1703). Explication de l’Arithmetique Binaire (Jean Boudot).

31. Laurent E, Amalian J-A, Parmentier M, Oswald L, Al Ouahabi A, Dufour F, Launay K, Clément J-L, Gignes M, Delsuc M-A, et al. (2020). High-capacity digital polymers: storing images in single molecules. Macromolecules 53, 4022–4029.

32. Solleder SC, Zengel D, Wetzel KS, and Meier MAR (2016). A scalable and high-yield strategy for the synthesis of sequence-defined macromolecules. Angew. Chem. Int. Ed Engl 55, 1204–1207. [PubMed: 26663541]

33. Porel M, and Alabi CA (2014). Sequence-defined polymers via orthogonal allyl acrylamide building blocks. J. Am. Chem. Soc 136, 13162–13165. [PubMed: 25204618]

34. Lutz J-F (2015). Coding macromolecules: inputting information in polymers using monomer-based alphabets. Macromolecules 48, 4759–4767.
35. Cao C, Krapp LF, Al Ouahabi A, König NF, Cirauqui N, Radenovic A, Lutz J-F, and Peraro MD (2020). Aerolysin nanopores decode digital information stored in tailored macromolecular analytes. Sci. Adv 6, eabc2661. [PubMed: 33298438]

36. Boukhet M, König NF, Ouahabi AA, Baaken G, Lutz J-F, and Behrends JC (2017). Translocation of precision polymers through biological nanopores. Macromol. Rapid Commun 38, 1700680.

37. Dahlhauser SD, Escamilla PR, VandeWalle AN, York JT, Rapagnani RM, Shei JS, Glass SA, Coronado JN, Moor SR, Saunders DP, and Anslyn EV (2020). Sequencing of sequence-defined oligourethanes via controlled self-immolation. J. Am. Chem. Soc 142, 2744–2749. [PubMed: 31986251]

38. Mondal T, Charles L, and Lutz J-F (2020). Damage and repair in informational poly(N-substituted urethane)s. Angew. Chem. Int. Ed. Engl 59, 20390–20393. [PubMed: 32779792]

39. Langbridge JA (2014). Professional Embedded ARM Development, First Edition (Wrox Press).

40. Huffman DA (1952). A method for the construction of minimum-redundancy codes. Proceedings of the IRE 40, 1098–1101.

41. Morris Mano M, and Kime CR (2008). Logic and Computer Design Fundamentals, Fourth Edition (Pearson).

42. Hwang S-H, Blaskovich MA, and Kim H-O (2008). A convenient reduction of alpha-amino acids to 1,2-amino alcohols with retention of optical purity. Open Org. Chem. J 2, 107–109.

43. Al Ouahabi A, Amalian J-A, Charles L, and Lutz J-F (2017). Mass spectrometry sequencing of long digital polymers facilitated by programmed inter-byte fragmentation. Nat. Commun 8, 967. [PubMed: 29042552]

44. Cho CY, Moran EJ, Cherry SR, Stephans JC, Fodor SP, Adams CL, Sundaram A, Jacobs JW, and Schultz PG (1993). An unnatural biopolymer. Science 261, 1303–1305. [PubMed: 7689747]

45. Kirby AJ (1980). Effective molarities for intramolecular reactions. In Advances in Physical Organic Chemistry, Volume 17, Gold V and Bethell D, eds. Advances in Physical Organic Chemistry (Academic Press), pp. 183–278.
Highlights

An immolative self-sequencing platform to encode and decode molecular information
High-throughput synthesis of multifunctional oligomers comprising 16 unique monomers
High-throughput, parallelized sequencing methodology for faster information retrieval
Software processes and converts the mass spectrometry data back to original information
Figure 1. An optimized Huffman tree

An optimized Huffman tree was generated from the exact frequencies of the text symbols in the phrase “Hello, World!” The most common symbols are represented by the fewest binary digits (bits).
Figure 2. Amino alcohol monomers
The amino alcohols that comprised oligomers A1 and A2 and their assigned octal notations used to encode the information.
See also Schemes S1 and S2 and Figures S1 and S2.
Figure 3. Reading the sequence of the oligomer

Oligomer A1 is sequenced and read by single-quadrupole mass spectrometry after 225 min of immolation. Calculating the mass differences between the oligomer precursor masses (precursor masses shown in blue and circled in green) gives the molecular weight of the monomer (shown in red). The reading frame starts with the disappearance of the phenylalaninol index. The octal symbol is then correlated to the monomer (shown in black). See also Figures S27-S34.
Figure 4. Decoding scheme for “Hello, World!”
The calculated mass differences for the observed truncated oligomers are correlated to each monomer, which is correlated to its respective octal notation. The octal is converted back to binary, and the binary was decoded back to plain text.
See also Figure S42.
Figure 5. Information overlay

An overlay of the information in its various encodings, starting with G1 in the top left and proceeding sinistrodextrally to H6 in the bottom right: in red, the molecular form; in blue, the hexadecimal string; in black, the bit string; and in bold, the English text.
Scheme 1.
5-exo-trig cyclization for sequencing urethanes