Multifunctional sequence-defined macromolecules for chemical data storage

Steven Martens\textsuperscript{1}, Annelies Landuyt\textsuperscript{1}, Pieter Espeel\textsuperscript{1}, Bart Devreese\textsuperscript{2}, Peter Dawyndt\textsuperscript{3} & Filip Du Prez\textsuperscript{1}

Sequence-defined macromolecules consist of a defined chain length (single mass), end-groups, composition and topology and prove promising in application fields such as anti-counterfeiting, biological mimicking and data storage. Here we show the potential use of multifunctional sequence-defined macromolecules as a storage medium. As a proof-of-principle, we describe how short text fragments (human-readable data) and QR codes (machine-readable data) are encoded as a collection of oligomers and how the original data can be reconstructed. The amide-urethane containing oligomers are generated using an automated protecting-group free, two-step iterative protocol based on thiolactone chemistry. Tandem mass spectrometry techniques have been explored to provide detailed analysis of the oligomer sequences. We have developed the generic software tools Chemcoder for encoding/decoding binary data as a collection of multifunctional macromolecules and Chemreader for reconstructing oligomer sequences from mass spectra to automate the process of chemical writing and reading.
Reliable data storage, already in the zettabyte ($10^{21}$) range and increasing, is one of the largest technological challenges of the digital age. While current conventional storage devices are still able to cope with this increasing demand, they occupy large floor areas in data centres, depend on ever rarer elements and require a great deal of energy, a resource already being stretched to the edge of current capacity. Encoding data at the (macro)molecular level could overcome these drawbacks, because physical maintenance is negligible, storage densities can be dramatically increased and sources of elements (C, H, N, O) as constituents of the information-containing macromolecules are highly abundant.

DNA, carrier of genetic information and arguably nature’s largest biopolymer, has already been used as a macromolecular carrier of information, able to archive, manage (DNA hard disk) and encrypt data, easily retrieved by well-established read-out tools. Moreover, immense storage densities can be achieved, i.e. $10^8$ times more information per mm$^3$ than in hard disk or flash memories. For example, 200 megabytes of data, including a high-definition music video and 100 books, were recently stored on DNA that contained more than 1.5 billion base pairs. Although the encoded information can be copied by DNA replication, susceptible to errors, DNA holds serious practical issues related to long-term stability and synthetic scalability. Indeed, DNA is sensitive to both hydrolysis and other degradation reactions, such as deamination and dimerization. These issues could be overcome with synthetic sequence-defined polymers if the backbone and side chains are chosen wisely. The structure of DNA is also quite complex, and the four-letter nucleotide code that makes up its ‘alphabet’ is limited compared to the vast diversity of synthetic building blocks. Another important issue hampering the large-scale use of DNA is the limited availability, the latter mostly connected with the scarceness of biologically available phosphorus in nature. Although recent research indicates that DNA is more stable than flash memory and that the amount of silicon might not be able to cope with the production of chips, it should be emphasized that, compared to phosphorus, silicon is 300 times more easily available on earth and can be retrieved from more accessible minerals. While producing DNA on a large enough scale is not possible on earth and can be retrieved from more accessible minerals, the amount of silicon might not be able to cope with the vast diversity of synthetic building blocks.

For peptide chemistry and synthetic sequences, examples can already be found of computational algorithms that revolutionized sequence-order reading, database building and de novo identification. Our research was inspired by the controlled fragmentation of the prepared sequence-defined oligomers and the time-consuming interpretation of the MS/MS-spectra to develop an automated sequencing tool, called Chemreader. The algorithm is first tested and improved by decoding a sentence written on oligo(amide-urethane)s. As can be seen in Fig. 1, the sequence can be fully read from left to right and vice versa (Fig.1). To write both the QR code, to oligomers and vice versa (Fig.1). To write both the QR code, oligomers can be made by using thiolactone (Tla) chemistry. Although a range of chemical functionalities could easily be inserted with both approaches, the one using acrylates, instead of amines, to introduce side-chain functionalities was more advantageous. It resulted in longer high-purity sequences and made use of an automated protocol (Fig. 1). Two different linkers, a Tla-containing alcohol and an acid, were used to connect the thiolactone moiety and solid support. For peptide chemistry and synthetic sequences, examples can already be found of computational algorithms that revolutionized sequence-order reading, database building and de novo identification.

Results

Read-out of the oligomers. In order to determine on the one hand the fragmentation behaviour of the oligomers and on the other hand the most suitable mass technique for the oligo(amide-urethane)s analysis (electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI) tandem mass spectrometry), a pentamer Z5 was first prepared starting from an acid linker and analysed (Supplementary Tables 1, 2 and Supplementary Figures 14–17). The fragments generated in the collision cell of these mass spectrometers mainly resulted from a controlled fragmentation on the urethane bond. As can be seen in Fig. 2, the sequence can be fully read from left to right and vice versa. In terms of potential mass range analysis for longer
sequences, we decided to continue with MALDI-TOF/TOF MS/MS. Although both positive and negative ion mode proved to work in the past for a variety of sequence defined oligomers, only positive mode was used here because the signal-to-noise ratio of MALDI-TOF MS signals is typically better in this mode. While it has been shown that the same synthetic platform also allows for modification of the oligomer backbone, the level of functionalization has been restricted to the side-chain in order to reduce the complexity of the resulting mass spectra (vide infra).

Following this initial study, six hexamers that were previously prepared on an automated synthesizer starting from an alcohol linker (H1-H6) were also deciphered with MALDI-MS/MS (Supplementary Table 3–8 and Supplementary Figures 1, 18–35). These hexamers were built with benzyl-, butyl- or tetrahydrofuranyl acrylate and contain repetitions in their sequences. Although H2 and H5 have the same mass, the order of their sequence could be easily determined and thus they could be differentiated unambiguously. For sequence H5, a more detailed analysis of the MS/MS spectrum was performed (Supplementary Figure 32).

Development of Chemreader algorithm. Once we had proven the easy read-out of these sequences, we explored their potential to store data and developed an algorithm (Chemreader) that automates the read-out process. Pentamer Z5 was used for the initial development of Chemreader. The algorithm uses both the masses of the collection of functionalities and the length of the monomer sequence as input parameters. In a first step, the program generates all possible fragments that could possibly be formed. Subsequently, it searches for matching masses that are obtained after MS/MS analysis. Finally, fragments are combined to reconstruct the original sequence. If we inspect pentamer Z5 in more detail (Fig. 2), fragmentation on the urethane bonds leads to the fragments necessary to perform the automated sequence analysis with the Chemreader algorithm. In all cases, both the start-containing fragment (left fragment with the acid linker) and the stop-containing fragment (right fragment with the thiolactone ring) are present in the spectrum. Presence of these two fragments makes it easy for the algorithm to unambiguously translate the MS/MS spectrum into the exact pentamer structure. The Chemreader algorithm has linear time complexity in the length of the polymers and the number of building blocks (octamers with a 20-character alphabet are resolved in the order of milliseconds on a standard laptop). A more detailed explanation of the algorithm can be found in the Supplementary Methods (Supplementary Figures 2, 3).

Writing and reading human-readable data. Next, we attempted to write the question TO WRITE OR NOT TO WRITE ON Oligos? on short oligomers. For this, the eight different words are converted into individual oligomers, using acrylates as a chemical alphabet to represent the individual characters. Comparable to previous research in which mass tags were added to oligomers to indicate the position of a letter in a word, the position of the words in the sentence has been encoded to enable the reconstruction of the words in the correct order. As a result, the sentence is actually encoded as 1TO 2WRITE 3OR 4NOT 5TO 6WRITE 7ON 8OLIGOS? The sentence was written twice using the two different linkers, showing the versatility of the α-end groups used for writing the oligomers (Supplementary Table 10–25 and Supplementary Figures 36–83). The acrylates (19 in total, each with a different mass) correspond to the different letters, numbers and the question mark in the sentence (Supplementary Table 9, Supplementary Figures 4–13 and Fig. 3).

Decoding the sentence requires knowledge about the alphabet (acrylates used), the number of words and the length of each word. Each word can be analysed separately. Given this information, Chemreader can reconstruct the original sentence. Only for the word BOLIGOS?, one peak corresponding to the smallest fragment was absent. However, due to the redundancy in
Fig. 2 Determining the sequence order. Tandem mass analysis (MALDI-MS/MS) of a pentamer Z5 with five different functionalities. In blue the read-out is highlighted from right to left, in purple from left to right. The coloured arrows indicate the mass difference between two mass fragments and the functionality that is responsible for this difference.

Fig. 3 Writing a sentence with sequences. The first two words of the question '1TO 2WRITE 3OR 4NOT 5TO 6WRITE 7ON 8OLIGOS?' in their chemical form. The different functionalities (in blue), introduced via acrylates in the chemical protocol, express a different letter or number.
overlapping fragments and the left-right and right-left reconstruction of the data, the octamer could be correctly translated. While encoding a human-readable sentence in sequence-defined polymers provided a first proof-of-principle to demonstrate the power of the Chemreader algorithm, the applied encoding scheme is not scalable to larger text fragments due to variable-length position encoding and to larger alphabet sizes (e.g. ASCII or Unicode) as separate acrylates are needed for all characters in the alphabet.

**Writing and reading of machine-readable data.** A second and more ambitious challenge was the synthesis and analysis of different oligomers to encode a $33 \times 33$ QR code, corresponding to a square grid containing 1089 pixels. With the ever-increasing use of smartphones, QR codes have become a simple way of communicating short messages. In producing a sample QR code that encodes the URL of the Wikipedia page of August Kekulé, we took advantage of the redundancy built into these codes—for error correction purposes—to embed a visual representation of the benzene ring. Kekulé was the first to understand the structure of benzene and made a proposal for its structure (1865) during his stay at Ghent University (1858–1867).

The black and white dots in a QR code represent bits (0 and 1) in the binary numeral system. As such, a QR code is nothing more than a two-dimensional bit string. To achieve the goal of encoding the QR code in sequence-defined polymers, the bit string was converted into a sequence of functionalities. To automate the process of encoding and decoding bit strings as collections of oligomers, a software tool called Chemcoder was developed.

The general outline of the Chemcoder algorithm is schematically represented in Fig. 4. The encoding of a QR code bit string is done in a series of steps. In a first step, the bit string is converted into a sequence of so-called flags (=side-chain functionalities). As this sequence of flags is too long to be encoded in a single oligomer, it is split into short fixed-length fragments. To give the last fragment the same length as the other fragments, it occasionally has to be filled with a non-coding spacer region (black; Fig. 4) as well as the total length of the original bit string (blue; Fig. 4). Decoding can only be done if the sequence of all fragments, they are dereplicated, sorted, trimmed and glued together. Finally, the sequence of flags is converted into the bit string that reconstructs the original QR code.

![Encode](image1)

**Fig. 4** Encoding and decoding of the QR code. Encoding scheme (left). The bit string representing the QR code is first translated into a pentadecimal numeral system (base-20). The sequence of 'flags' is then cut into smaller pieces. In a final step, the position of each fragment (purple) and the length of the bit string (blue) is added. The last fragment may be filled with a non-coding spacer (black); Decoding scheme (right). After determination of the sequence of all fragments, they are dereplicated, sorted, trimmed and glued together. Finally, the sequence of flags is converted into the bit string that reconstructs the original QR code.

As every oligomer had to be analysed separately, a future challenge would be to combine techniques for the analysis of much more complex samples, in order to guarantee a high data density. An example, well known in the context of peptide analysis, consists of the coupling of liquid chromatography to a tandem ESI-MS/MS equipment to separate different oligomers in the LC dimension and determine the sequence in the tandem MS dimension.

The storage capacity of sequence-defined oligomers based on thiolactone chemistry was explored. It is possible for such oligomers to directly contain digital information in a useful way (QR code) while a controlled fragmentation on the urethane bond allowed for an easy read-out of the oligomers. An algorithm, called Chemreader, was developed to facilitate the read-out of these sequences, which allows one to read the information stored within sequence-defined structures in a fast and automated way on a standard laptop. The Chemreader algorithm contributes to...
solving the sequence-reading bottleneck of sequence-defined polymers. In order to test the Chemreader algorithm, a sentence in natural language was first successfully written and read, followed by the more ambitious challenge of encoding a 33 × 33 QR code in 71 different, analysed oligomers. Besides, we developed the software tool Chemcoder to quickly encode binary data as a compact collection of oligomer fragments, and vice versa. Both algorithms are extremely fast and highly configurable for application on other sets of sequence-defined polymers. A reference implementation is available open source on GitHub (see section Additional Information for URL). We invite other groups to apply them on their own data sets or make any modifications for their own needs.

As the results obtained prove the possibilities for using these mono-disperse, multi-functional oligomers in the field of data storage, this study is another indication for the long-term potential that sequence-defined polymers hold to real-world applications and thus provides further validation for this rapidly developing branch of macromolecular chemistry. Undoubtedly, this will spark further research on the analysis and applicability of sequence-defined polymers worldwide. One of the main research challenges remains the further exploration of non-destructive techniques for the read-out of the sequence order in complex mixtures.

**Materials**

DMSO-d₆ ([2206-27-1], ≥99.8%), and CHCl₃-d ([865-49-6], ≥99.8%) were purchased from Euriso-top. Acryloyl chloride ([814-68-6], 96%) was purchased from acer GmbH. Acetoniitrile ([75-05-8], HPLC grade), 1,4-Dioxane ([123-91-1], HPLC grade), and Triethylamine ([121-44-8], 99%) were purchased from Acrors Organics. DL-Homocysteine thiocacetoic hydrochloride ([6038-19-3], 99%) was purchased from Haishang industry (Jinan City, China). Magnesium sulphate hydrate ([22189-08-8], ≥99%), Potassium carbonate ([584-08-7], ≥99%) and Sodium bicarbonate ([144-55-8], ≥99.5%) were purchased from Carl Roth. Trifluoroacetic acid ([77-05-1], Petroleum ether (ether n-pentane, ≥95%)) and 2-Chlorotrityl chloride resin ([40274-68-0], 100-200 mesh, 1% DVB, 1.6 mmol/g) were purchased from Iris Biotech GmbH. Acetyl chloride ([73-36-5], ≥99%), Bromoacetyl bromide ([598-21-0], ≥98%), Butyl acrylate ([141-32-2], ≥99%), Chloroform ([865-49-6], ≥99%), Citronellol ([106-22-9], ≥99.5%), Dichloromethane ([75-09-2], ≥99%), Diethyltereth ([60-29-7], ≥99.9%), N,N-Diisopropylthiophenylethylamine (DIPA), ([7087-68-5], 99%), N,N-Dimethylformamide ([68-12-2], anhydrous, 99.8%), Ethanolamine ([141-43-5], ≥99%), Ethyl acrylate ([140-88-5], ≥99%), Glutaric anhydride ([108-55-4], 95%), 1-Heptanol ([111-70-6], ≥98%), 2-Hydroxyethyl acrylate ([8181-61-9], ≥96%), Isobornyl acrylate ([5588-33-5], technical grade), 2-Mercaptobenzoic acid ([160-24-2], ≥99%), Methanol ([64-17-5], ≥99.9%), Methyl acrylate ([96-33-3], ≥99%), Phenothiazine ([92-84-2], ≥99%), 1-Propanol ([71-23-8], 99%), Propargyl acrylate ([10447-47-1], ≥99%), Pyridine ([110-86-1], ≥99%), Tetrhydrofuran ([109-99-9], ≥99%) were purchased from Sigma-Aldrich and used without purification, except isobornyl acrylate which was distilled. Benzyl acrylate ([2495-35-4], >97%), 2-Cyanethyl Acrylate ([106-71-8], >95%), Cyclohexyl Acrylate ([3066-71-5], ≥98%), Dibutyltiran dilaurate ([77-58-7], ≥98%), N,N-Diethylacrylamide ([2675-94-7], ≥98%), 2-(Dimethylamino)ethyl acrylate ([2439-35-2], ≥98%), 2-Ethoxethanol ([110-80-5], ≥99%), 2-(2-Ethoxyethoxy) ethyl acrylate ([7328-17-8], ≥98%), 2-Ethylhyl Acrylate ([103-11-7], ≥99%), Isobornyl acrylate ([14245-24-7], 99%), Methanol ([67-56-1], ≥99.9%), Methacrylate ([96-33-3], ≥99%), Phenothiazine ([92-84-2], ≥99%), 1-Propanol ([71-23-8], 99%), Propargyl acrylate ([10447-47-1], ≥99%), Pyridine ([110-86-1], ≥99%), Tetrhydrufurfuryl acrylate ([1299-48-6]) was purchased from Polysciences and used without purification. Hydrochloric acid 36% (HCl, [7642-73-2]) was purchased from Chem-Lab and used without purification. Solvents (CHCl₃, CH₂Cl₂, DPEA and pyridine) for the preparation of functionalized thiolactone or the immobilization of functionalized thiolactone links were distilled from CaH₂ prior to use. Silicagel (ROCI, SI 1721, 60, 40-63 µm) was used to perform preparative column chromatography, eluting with technical solvents. The collected fractions were analysed by thin layer chromatography (TLC plates, Macherey-Nagel, SIL G-25 UV254). The a-isocyanato-γ-thiolactone or the di-functionalization of γ-thiolactone or the di-functionalization of γ-thiolactone was used to perform preparative column chromatography, eluting with technical solvents. The collected fractions were analysed by thin layer chromatography (TLC plates, Machereynagel, SIL G-25 UV254). The a-isocyanato-γ-thiolactone, the di-functionalization of γ-thiolactone and 3,7-dimethyl-6-1-yl (acrylonitrile) acrylate were synthesized according to literature procedures.

**Materials**

For the identification of the chemical identity of the synthesized polymers, the following solvents were purchased from Sigma-Aldrich and used without purification: Acetonitrile ([75-05-8], 99%), 1-Propanol ([71-23-8], ≥98%), 2-Ethoxethanol ([110-80-5], ≥99%), 2-(2-Ethoxyethoxy) ethyl acrylate ([7328-17-8], ≥98%), 2-Ethylhyl Acrylate ([103-11-7], ≥99%), Isobornyl acrylate ([14245-24-7], 99%), Methanol ([67-56-1], ≥99.9%), Methacrylate ([96-33-3], ≥99%), Phenothiazine ([92-84-2], ≥99%), 1-Propanol ([71-23-8], 99%), Propargyl acrylate ([10447-47-1], ≥99%), Pyridine ([110-86-1], ≥99%), Tetrhydrufurfuryl acrylate ([1299-48-6]) was purchased from Polysciences and used without purification. Hydrochloric acid 36% (HCl, [7642-73-2]) was purchased from Chem-Lab and used without purification. Solvents (CHCl₃, CH₂Cl₂, DPEA and pyridine) for the preparation of functionalized thiolactone or the immobilization of functionalized thiolactone links were distilled from CaH₂ prior to use. Silicagel (ROCI, SI 1721, 60, 40-63 µm) was used to perform preparative column chromatography, eluting with technical solvents. The collected fractions were analysed by thin layer chromatography (TLC plates, Macherey-Nagel, SIL G-25 UV254). The a-isocyanato-γ-thiolactone, the di-functionalization of γ-thiolactone and 3,7-dimethyl-6-1-yl (acrylonitrile) acrylate were synthesized according to literature procedures.

**Experimental procedures**

Detailed experimental procedures are described in the Supplementary Methods and are accompanied with reaction schemes when appropriate.

**Code availability**

Both the algorithms can be found at https://github.com/chemstore/chemreader. The individual algorithms can be found at https://github.com/chemstore/chemcoder and https://github.com/chemstore/chemreader.

**Data availability**

All relevant data are available within the paper and its Supplementary Information. The algorithms can be found at https://github.com/chemstore. All other data are available from the authors upon request.

Received: 4 June 2018 Accepted: 3 October 2018

**References**

Colquhoun, H. & Lutz, J.-F. Information-containing macromolecules. Nat. Chem. 6, 455–456 (2014).

Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthetic DNA. Nature 507, 731–735 (2014).

Zhnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. Nat. Mater. 15, 366–370 (2016).

Bornholt, J. et al. A DNA-based archival storage system. SIGPLAN Not. 51, 637–649 (2016).
9. Shendure, J. & Ji, H. Next-generation DNA sequencing.

8. Shiu, H. J., Ng, K. L., Fang, J. F., Lee, R. C. T. & Huang, C. H. Data hiding technologies based on DNA sequences. Inf. Sci. (Ny.) 180, 2196–2208 (2010).

7. Kawane, K., Motani, K. & Nagata, S. DNA degradation and its defects. Cold Spring Harb. Perspect. Biol. 6, 1–14 (2014).

6. Childers, D. L., Corman, J., Edwards, M. & Elser, J. J. Sustainability challenges in the 21st century. Nat. Sci. 8, 3469 (2016).

5. Cavallo, G., Al Ouahabi, A., Oswald, L., Charles, L. & Lutz, J.-F. Orthogonal sequences and functionality. Angew. Chem. Int. Ed. 53, 8290–8296 (2017).

4. Boukis, A. C. & Meier, M. A. R. Data storage in sequence-defined macromolecules via hetero bifunctional synthons. Chem. Commun. 51, 1799–1802 (2015).

3. Burel, A., Carapito, C., Lutz, J.-F. & Charles, L. MS-DECORDER: millisecond sequencing of coded polymers. Macromolecules 50, 8290–8296 (2017).

2. Martens, S. et al. Automated synthesis of monodisperse oligomers, featuring sequence control and tailored functionalization. J. Am. Chem. Soc. 138, 14182–14185 (2016).

1. Burel, A., Carapito, C., Lutz, J.-F. & Charles, L. MS-DECORDER: millisecond sequencing of coded polymers. Macromolecules 50, 8290–8296 (2017).
**Additional information**

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-06926-3.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018