Mass spectrometry sequencing of long digital polymers facilitated by programmed inter-byte fragmentation
Abdelaziz Al Ouahabi, Jean-Arthur Amalian, Laurence Charles, Jean-Francois Lutz

To cite this version:
Abdelaziz Al Ouahabi, Jean-Arthur Amalian, Laurence Charles, Jean-Francois Lutz. Mass spectrometry sequencing of long digital polymers facilitated by programmed inter-byte fragmentation. Nature Communications, 2017, 8 (1), pp.967. 10.1038/s41467-017-01104-3. hal-01774259

HAL Id: hal-01774259
https://amu.hal.science/hal-01774259v1
Submitted on 3 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Mass spectrometry sequencing of long digital polymers facilitated by programmed inter-byte fragmentation

Abdelaziz Al Ouahabi, Jean-Arthur Amalian, Laurence Charles & Jean-François Lutz

In the context of data storage miniaturization, it was recently shown that digital information can be stored in the monomer sequences of non-natural macromolecules. However, the sequencing of such digital polymers is currently limited to short chains. Here, we report that intact multi-byte digital polymers can be sequenced in a moderate resolution mass spectrometer and that full sequence coverage can be attained without requiring pre-analysis digestion or the help of sequence databases. In order to do so, the polymers are designed to undergo controlled fragmentations in collision-induced dissociation conditions. Each byte of the sequence is labeled by an identification tag and a weak alkoxyamine group is placed between 2 bytes. As a consequence of this design, the NO-C bonds break first upon collisional activation, thus leading to a pattern of mass tag-shifted intact bytes. Afterwards, each byte is individually sequenced in pseudo-MS³ conditions and the whole sequence is found.

1 Université de Strasbourg, CNRS, Institut Charles Sadron UPR22, 23 rue du Loess, 67034 Strasbourg, France. 2 Aix-Marseille Université, CNRS, UMR 7273, Institute of Radical Chemistry, 13397 Marseille, France. Correspondence and requests for materials should be addressed to L.C. (email: laurence.charles@univ-amu.fr) or to J.-F.L. (email: jflutz@unistra.fr)
It has been demonstrated in recent years that digital information can be stored in biological and synthetic macromolecules. In such digital polymers, the monomer units that constitute the chains are used as molecular bits and assembled through controlled synthesis into readable digital sequences. For example, it has been reported that ordered oligonucleotide sequences enable storage of several kilobytes of data in DNA chains. Alternatively, our group has demonstrated that binary messages can also be stored in different types of synthetic macromolecules. Overall, molecular encryption in polymers opens up exciting avenues for massive information storage as well as long-term data storage. Importantly, the use of digitally encoded polymers allows room temperature storage and thus activated relatively easily. MS/MS sequencing of long polymer chains enables storage of several kilobytes of data in DNA synthetic macromolecules. As schematized in Fig. 1d, inter-byte NO-C bonds shall break first during the first activation stage and lead to a MS2 spectrum containing all intact individual bytes. Then, subjecting each byte to a second activation stage should yield MS3 spectra that would allow an easier sequencing task, thanks to the small size and charge state of dissociating species. However, since bytes may be isobaric (i.e., contain the same number of 0 and 1 units), each byte shall be first labeled with a tag, which allows its identification in MS2 and permits to know its location in the initial sequence. Although a wide variety of molecules could be potentially used as byte tags, natural (noted A, T, G, C) and non-natural (noted B, I, F) nucleotides were selected in the present work (Fig. 1c), because the corresponding phosphoramidite monomers are commercially available. As shown in Fig. 1c, the molecular structure of each byte tag was selected in order to create an unequivocal mass and isotopic signature for each byte. In particular, two simple criteria shall be fulfilled: the molar mass of a given byte tag shall not be a multiple of 28, which is the mass difference between a 0 and 1 synthons, and the mass difference between 2-byte tags shall not be a multiple of 28 and shall also not be smaller than 3 Da since triply charged species (vista infra) are studied in MS3. The byte tag sequence, T, C, A, G, B, I, F, no tag, was chosen to label the bytes and reading was arbitrarily started from the non-marked byte (i.e., opposite to the sense of synthesis). This means, for example, that the first, penultimate, and last bytes of a sequence always contain no tag, a C-tag, and a T-tag, respectively. Hence, after performing MS, MS2, and MS3 steps, the whole sequence of the original polymer can be comprehensively reconstructed.

In order to verify the feasibility of this concept, model polymers containing only 2 bytes of information separated by one single cleavable NO-C site were first studied (Supplementary Table 1, Entries 3, 4). In particular, two different alkoxamines containing either a mono- or a di-methylated carbon were considered. These cleavable groups were incorporated in between the bytes during polymer synthesis using phosphoramidite monomers a1 and a2 (Supplementary Fig. 3). It was found that the mono-methylated alkoxamine a1 is not optimal because the energy threshold for the NO-C bond homolysis appears to be close to that of inner-byte phosphate bond cleavage, thus resulting in a polluted MS2 spectrum (Supplementary Fig. 4a). On the other hand, the more labile NO-C linkage in dimethylated a2 leads to preferential fragmentation in MS2 conditions, where phosphate bonds do not break (Supplementary Fig. 4b). Yet, it should be remarked that a2 leads to some slight in-source byte fragmentation in the initial MS analysis of the complete polymer. However, these ions can be easily tracked and disclosed based on their much lower charge state compared to the intact macromolecule. Thus, alkoxamine a2 was selected for the synthesis of digital poly(phosphodiester)s containing 4, 5, 6, or 8 bytes of information (Supplementary Table 1). All these polymers were synthesized by automated phosphoramidite chemistry. For each monomer attachment, cycles involving dimethoxytrityl (DMT) deprotection of the reactive alcohol sites; phosphoramidite coupling; oxidation of the formed phosphite into a phosphate; and capping of the unreacted alcohol sites by reaction of long sequence-defined polymers. After synthesis, the DMT-terminated digital sequences were cleaved from the support and purified on a
reverse-phase cartridge. This purification process allows separation of the desired DMT-terminated sequences from failure sequences capped by acetic anhydride. Afterwards, the terminal DMT group is removed. As shown in Supplementary Table 1, all polymers were obtained in good yields after reverse-phase column purification.

**Mass spectrometry sequencing of the multi-byte polymers.** Figure 2 shows the three-stage analysis of a polymer containing 4 bytes of information (Supplementary Table 1, Entry 5). This polymer was first analyzed by negative mode electrospray ionization (ESI)-MS, which revealed a dominant charge state. It should be specified that MS^2 experiments were not performed here for sequencing purposes, and traditionally defined as the production of fragments that differ in mass by a single building unit and hence allow the original chain to be reconstructed. Instead, activation of precursor ions aims here at producing fragments that all contain a single byte (to be further sequenced in MS^3). However, since dissociation of precursor with n bytes proceeds by competitive NO-C bond cleavages, primary product ions contain from 1 byte (either the first or the last one) to n−1...
Fig. 2 Sequencing of a 4-byte digital polymer that contains the ASCII-encoded word Byte. a High-resolution electrospray mass spectrum (MS1) obtained in the negative ion mode for a 4-byte digital polymer (Supplementary Table 1, Entry 5). The upper numbers represent the different charge states observed for the polymer. Dark grey diamonds indicate in-source fragments (see Supplementary Table 2 for a detailed interpretation of each peak) and dark grey squares designate different charge states of two different synthesis impurities. b MS2 spectrum (0.56 eV, center of mass frame) obtained by collision-induced dissociation of the [M-12H]12− precursor ion, which carries on average three charges per byte. In this case, the MS/MS single-byte, double-byte, and triple-byte fragments are predominantly observed as trianions, hexa-anions, and nona-anions, respectively. Other charge states can be observed to a minor extent and are denoted by dark grey circles (see Supplementary Table 2 for a detailed interpretation of each peak). c Molecular sequencing (pseudo-MS3) of a byte fragment obtained by collision-induced dissociation (0.56 eV, center of mass frame) of the precursor trianion [M-3H]3−. For clarity, only the sequencing of byte 4 is shown as an example in this figure. The sequencing of bytes 1–3 is shown in Supplementary Fig. 6.

bytes. Collision energy was hence raised to promote consecutive dissociations of large primary product ions, in order to form secondary fragments that each contains one inner-chain byte. Energy has, however, to remain below dissociation threshold of phosphate bonds to prevent inner-byte fragmentation. In such conditions, a very clear bytes pattern can be observed in the resulting MS2 spectrum (Fig. 2b). Each byte appears predominantly as a trianion (while remaining as minor signals at −2 or −4 charge states) and the byte tags lead to unambiguous mass shifts that allow identification of their initial location in the chain (Supplementary Table 3). Importantly, the concept also works for polymers containing similar or isomeric bytes. For instance, Supplementary Fig. 5 shows the sequencing of a polymer containing four times the same byte (Supplementary Table 1, Entry 6). Even in such a case, the byte tags allow unequivocal detection of each byte in MS2. Fragments resulting from partially cleaved chains, hence containing either two or three bytes and, respectively, detected at −6 and −9 charge states, were also observed in MS/MS spectra, as exemplified in Fig. 2b. After MS2 byte cleavage, each byte was individually sequenced by MS3. In order to take advantage of the resolving capabilities offered by orthogonal acceleration time-of-flight (oa-TOF) mass analyzers to safely assign fragments, the Q-oa-TOF instrument used here for MS and MS2 stages was also employed to perform pseudo-MS3 experiments. Typically, deprotonated polymers were first activated in the instrument interface by raising the cone (or skimmer) voltage to perform in-source CID; then, so-released byte fragments were mass selected in the quadrupole for further excitation in the collision cell and fragment measurement in the oa-TOF. Figure 2c shows, for example, the sequencing of the
Tagged last byte of polymer 5 and the sequencing of bytes 1–3 is shown in Supplementary Fig. 6. Hence, the complete digital sequence of the polymer was easily deciphered. In order to evidence the universality of this concept, polymers with other 4-byte sequences (Supplementary Table 1, Entries 6, 7) were analyzed (Supplementary Figs. 7–11).

Supplementary Figs. 12–20 show the data obtained for 5- and 6-byte polymers (Supplementary Table 1, Entries 8–10) and confirm that messages encoded in longer polymers can also be easily deciphered using the proposed multi-step sequencing. It is, however, interesting to point out that 5-byte polymers were only labeled with natural byte tags A, T, G, C, whereas the 6-byte polymer also contains a non-natural byte tag B. The latter contains a bromine atom and, therefore, does not only allow byte identification by mass but also by isotopic pattern. Ultimately, an 8-byte polymer (Supplementary Table 1, Entry 11) was synthesized and analyzed. Figure 3 shows the MS2 spectrum obtained from the precursor anion containing 24 negative charges (i.e., three charges in average per byte). After promoting multiple in-chain NO-C fragmentations, a clear pattern of byte trianions was measured in MS2, thus allowing individual byte sequencing by pseudo-MS3 (Supplementary Figs. 21–25).

Discussion

The development of practical polymer-based digital memories requires libraries of macromolecules that contain at least a few bytes of data in each chain. Sequence-coded polymer chains containing more than a hundred bits can be synthesized using automated solid-phase chemistry or inkjet technologies. However, MS sequencing of such long chains is very challenging, in particular, when targeting full sequence coverage as mandatory for digital polymers. Indeed, assuming an efficient collisional activation, the total ion current is spread over a large number of fragments. Moreover, these fragments are most often produced at different charge states because they are formed from highly charged macromolecules generated by ESI, a soft technique used to ensure their structural integrity. For instance, when full sequence coverage is required, it is known that MS/MS is most efficient for the sequencing of peptides containing less than 20 residues and that the analysis of longer proteins requires enzymatic digestion and HPLC separation of complex mixtures. Similarly, MS/MS sequencing of intact nucleic acids is not trivial. Such a situation can hardly be improved, as the molecular structure of biopolymers is set by biology.

In contrast, as demonstrated in this article, synthetic polymer chemistry allows design of digital polymers that may undergo controlled fragmentations in MS/MS conditions. Indeed, the experimental data highlighted in this paper indicate that intact long sequence-coded chains can be fully sequenced in a routine mass spectrometer operating at low collision energy, without requiring digestion, purification, or separation steps prior to analysis. The use of databases is also not mandatory in this approach to decipher the coded sequences. To attain such a MS readability, the molecular structure of the polymers was carefully engineered and, in particular, two key features were implemented: the use of cleavable inter-bytes spacer that promotes programmed MS2 fragmentation and the use of mass tags that allow identification of byte original location. The former feature allows the actual MS sequencing task to be limited to very short (8 bits) oligo(phosphodiester)s, while the latter one permits reliable reconstruction of the byte sequence. It should be noted that the reported concept is not limited to poly(phosphodiester)s and could be extended to other types of digital polymers. For instance, alkoxyamine groups are most probably not the only type of interbyte links that can be used in this approach. Depending on the type of digital polymer that shall be deciphered by MS, other cleavable linkers may be imagined. In fact, the general rule is that...
the inter-byte linkers require less energy for being broken in CID conditions than the intra-byte bonds that connect the bits. Furthermore, byte tags shall not necessarily be nucleotides. As mentioned in the results section, these markers have been selected in the present work because their phosphoramidite derivatives are commercially available. Yet, other byte tags can be envisioned. Here, the general rule is that the molar mass of a byte tag and the mass difference between two byte tags shall not be a multiple of the mass difference between 2 bits. In addition, markers with markedly different isotopic signatures can also be imagined.

Overall, this work opens up interesting perspectives for the design of polymer-based molecular memories. For such technological applications, it is important to specify that the synthesis of very-long-digital polymers (i.e., linear chains containing several hundreds of coded bits) is not an objective. Indeed, polymer-based memory devices will most probably rely on libraries of coded chains, as already done in the field of DNA storage.

Methods

Materials. 2-cyanoethyl diisopropyl-chlorophosphoramidite (95%, Alfa Aesar), 4,4′-dimethoxytriphenylmethyl chloride (97.0%, Sigma-Aldrich), N,N-diisopropyl-pyridinylamine (DPEPA, 99%, Alfa Aesar), triethylamine (99%, Alfa Aesar), 1,1,4,7,7-pentamethyldiethylenetriamine (PMDTA, 98%, Alfa Aesar), 3-amino-1-propanol (98%, Alfa Aesar), anhydrous methylamine (AMA, 98%, Sigma-Aldrich), anhydrous hydroxide solution (20.0–30.0% NH4, Sigma-Aldrich), 2-bromopropanol bromide (97%, Sigma-Aldrich), α-bromoisobutyl bromide (98%, Sigma-Aldrich), 4-hydroxy-TEMPO (97%, Sigma-Aldrich), and copper(I) bromide (9.6 g, 44.4 mmol), 3-aminopropan-1-ol (3.3 g, 44.0 mmol) and triethylamine (210.0130; found 210.0145. 2-bromo-N-(3-bis(4-methoxyphenyl)phenyl)alkyl)phosphorodiamidate (b2). The compound b2 was isolated in 90% yield following the same procedure for b1 using d2 (2.42 g, 10.8 mmol); hydroxy-TEMPO (1.72 g, 10 mmol), CuCl (2.0 g, 20 mmol) and PMDTA (3.46 g, 20 mmol). TLC (ethyl acetate): Rf = 0.48; 1H NMR (400 MHz, CDCl3, δ ppm): 6.79 (1H, NH), 3.95 (1H, H-CH2), 3.61 (1H, J = 4.9 Hz, 2H, HO-CH2), 3.42 (2H, -NH-CH2-), 3.19 (3H, 1H, -OH), 1.83–1.87 (2H, -N-(C(CH3)2)2-CH2-), 1.71 (2H, -N-(C(CH3)2)2-CH2-), 1.11–1.18 (6H, Jf = 29.6 Hz, 12H, (-N-(C(CH3)2)2-CH2-)). 13C NMR (75 MHz, CDCl3, δ ppm): 176.2, 83.0, 59.1, 49.1, 35.6, 21.6 ppm. HRMS (m/z) [M + H]+ calcld. for C42H35NO4, 676.8791; found 676.8756. 2-(4-hydroxy-2,6,6-tetramethylpiperidin-1-yl)-2-propanol (b2). The compound b2 was isolated in 90% yield following the same procedure for b1 using d2 (2.42 g, 10.8 mmol); hydroxy-TEMPO (1.72 g, 10 mmol), CuCl (2.0 g, 20 mmol) and PMDTA (3.46 g, 20 mmol). TLC (ethyl acetate): Rf = 0.48; 1H NMR (400 MHz, CDCl3, δ ppm): 6.79 (1H, NH), 3.95 (1H, H-CH2), 3.61 (1H, J = 4.9 Hz, 2H, HO-CH2), 3.42 (2H, -NH-CH2-), 3.19 (3H, 1H, -OH), 1.83–1.87 (2H, -N-(C(CH3)2)2-CH2-), 1.71 (2H, -N-(C(CH3)2)2-CH2-), 1.11–1.18 (6H, Jf = 29.6 Hz, 12H, (-N-(C(CH3)2)2-CH2-)). 13C NMR (75 MHz, CDCl3, δ ppm): 176.2, 83.0, 59.1, 49.1, 35.6, 21.6 ppm. HRMS (m/z) [M + H]+ calcld. for C42H35NO4, 676.8791; found 676.8756.
The DMT-protected polymers were cleaved from the solid support using a solution of the targeted sequence-coded polymers were cleaved on the column and washed or from the activated precursor, as illustrated in Supplementary Fig. 36. Other controllability of the carbon-centered radical in the—termination of all byte fragments—containing (i.e., wiz−2-cyanoethyl—containing (i.e., wiz

| Reference | Title | Publisher | Volume | Page | Year |
|-----------|-------|-----------|--------|------|------|
| 1. | Zhivnin, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. | Nat. Mater. | 15 | 366–370 | 2016 |
| 2. | Colquhoun, H. & Lutz, J.-F. Information-containing macromolecules. | Nat. Chem. | 6 | 453–456 | 2014 |
| 3. | Lutz, J.-F., Ouchi, M., Liu, D. R. & Sawamoto, M. Sequence-controlled polymers. | Science | 341 | 1238149 | 2013 |
| 4. | Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. | Science | 357 | 1628–1628 | 2017 |
| 5. | Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. | Nature | 494 | 77–80 | 2013 |
| 6. | Mayer, C., McNroy, G. R., Murat, P., Van Delft, P. & Balasubramanian, S. An epigenetics-inspired DNA-based data storage system. | Angew. Chem. Int. Ed. | 55 | 11144–11148 | 2016 |
| 7. | Ehrlich, Y. & Zielinski, D. DNA fountain enables a robust and efficient storage mechanism. | Science | 355 | 950 | 2017 |
| 8. | Al Ouahabi, A., Charles, L. & Lutz, J.-F. Non-linear synthesis. | Angew. Chem. Int. Ed. | 137 | 5629–5635 | 2017 |
| 9. | Roy, R. K. et al. Design and synthesis of digitally encoded polymers that can be decoded and erased. | Nat. Commun. | 6 | 7237 | 2015 |
| 10. | Gorny, U. S. et al. Chemoselective synthesis of uniform sequence-coded polyurethanes and their use as molecular tags. | Chem. | 1 | 114–126 | 2016 |
| 11. | Grass, R. W., Heckl, P., Puddu, M., Paunesku, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. | Angew. Chem. Int. Ed. | 54 | 2552–2555 | 2015 |
| 12. | Shendure, J. & Ji, H. Next-generation DNA sequencing. | Nat. Biotechnol. | 26 | 1135–1145 | 2008 |
| 13. | Apte, C. et al. The potential and challenges of nanoparticle sequencing. | Nat. Biotechnol. | 26 | 1146–1153 | 2008 |
| 14. | Mutlu, H. & Lutz, J.-F. Reading polymers: sequencing of natural and synthetic macromolecules. | Angew. Chem. Int. Ed. | 53 | 13010–13019 | 2014 |
| 15. | Wesdemiotis, C. Multidimensional mass spectrometry of synthetic polymers and advanced materials. | Angew. Chem. Int. Ed. | 56 | 1452–1464 | 2017 |
| 16. | Lutz, J.-F. Coding macromolecules: inputting information in polymers using monomer-based alphabets. | Macromolecules | 48 | 4759–4767 | 2015 |
| 17. | Cavallo, G., Al Ouahabi, A., Oswald, L., Charles, L. & Lutz, J.-F. Orthogonal synthesis through “easy-to-read” information-containing polymers using phosphoramidite and radical coupling steps. | J. Am. Chem. Soc. | 138 | 9417–9420 | 2016 |
| 18. | Roy, R. K., Laure, C., Fischer-Krauser, D., Charles, L. & Lutz, J.-F. Convergent synthesis of digitally encoded poly(alkoxysilane amide). | Chem. Commun. | 51 | 15677–15680 | 2015 |
| 19. | Al Ouahabi, A., Kotera, M., Charles, L. & Lutz, J.-F. Synthesis of monodisperse sequence-coded polymers with lengths above DP100. | ACS Macro Lett. | 4 | 1077–1080 | 2015 |
| 20. | Gurfinkel, M. et al. Caruthers, M. H. Gene synthesis machines: DNA chemistry and its uses. | Science | 285 | 281–285 | 1985 |
| 21. | Edwardson, T. G. W., Carneiro, K. M. M., Serpell, C. J. & Sleiman, H. F. An easy-to-read information-containing polymers using phosphoramidite and radical coupling steps. | Science | 306 | 9417–9420 | 2016 |
| 22. | Pan, X. et al. Automated synthesis of well-designed polypeptides and biohybrids by atomic transfer radical polymerization using a DNA synthesizer. | Angew. Chem. Int. Ed. | 56 | 2740–2743 | 2017 |
| 23. | LeProust, E. M. et al. Synthesis of high-quality libraries of long (150mers) oligonucleotides by a novel depurination controlled process. | Nucleic Acids Res. | 38 | 2522–2540 | 2010 |
Acknowledgements
These results were obtained with the support of the French National Research Agency (ANR project 00111001, grant numbers ANR-16-CE29-0004-01 and ANR-16-CE29-0004-02). J.-F.L. also thanks the excellence network Chemistry of Complex Systems (LabEx CSC) and the CNRS for financial support.

Author contributions
J.-F.L. wrote the paper. J.-F.L. and L.C. conceived the idea, designed the experiments and supervised research. A.A.O. synthesized all monomers and polymers. J.-A.A. and L.C. performed mass spectrometry measurements. All authors analyzed the data, provided figures, and commented on the manuscript.

Additional information
Supplementary Information accompanies this paper at doi:10.1038/s41467-017-01104-3.

Competing interests: The authors declare no competing financial 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) 2017