Isotope Ratio Encoding of Sequence-Defined Oligomers

Márton Zwillinger, Lucile Fischer, Gergő Sályi, Soma Szabó, Márton Csékei, Ivan Huc, András Kotschy

Isotope ratio encoding of sequence-defined oligomers

Márton Zwillinger[1,2], Lucile Fischer[3], Gergő Sályi[1], Soma Szabó[1], Márton Csékei[1], Ivan Huc[4]*, András Kotschy[1]**

[1]Servier Research Institute of Medicinal Chemistry, H-1031 Budapest, Hungary

[2]Hevesy György PhD School of Chemistry, Eötvös Loránd University, H-1053 Budapest, Hungary

[3]CBMN UMR5248, University of Bordeaux-CNRS-IPB, F-33600 Pessac, France

[4]Department of Pharmacy and Center for Integrated Protein Science, D-81377 Munich, Germany

Encoding, Information storage, Oligomer, Stable isotope labelling, Isotope ratio encoding, Barcode, Deuterium labelling, Isotope pattern recognition

ABSTRACT: Information storage at the molecular level commonly entails encoding in the form of ordered sequences of different monomers and subsequent fragmentation and MS/MS analysis to read this information. Recent approaches also include the use of mixtures of distinct molecules non-covalently bonded to one another. Here, we present an alternate isotope ratio encoding approach utilizing deuterium-labelled monomers to produce hundreds of oligomers endowed with unique isotope distribution patterns. Mass spectrometric recognition of these patterns then allowed us to directly readout encoded information with high fidelity. Specifically, we show that all 256 tetramers comprised of four different monomers of identical constitution can be distinguished by their mass fingerprint using mono-, di-, tri- and tetradeuterated building blocks. The method is robust to experimental errors and does not require the most sophisticated MS instrumentation. Such isotope ratio encoded oligomers may serve as tags that carry information, but the method mainly opens up the capability to write information, e.g. about molecular identity, directly into a pure compound via its isotopologue distribution obviating the need for additional tagging, and avoiding the use of mixtures of different molecules.

INTRODUCTION

The encoding of information into chemical structures and the subsequent readout and transmission of this information bear relevance to multiple sub-fields of chemistry. In biopolymers, information is contained in the sequential organization of a defined set of monomers. Information can be stored under the form of genetic material, duplicated, and translated into a peptidic backbone. Protein folding itself can also be considered as a sort of translation of the purely sequential information of the primary structure into its three-dimensional functional expression. Furthermore, proteins engage in all kinds of recognition and chemical manipulations that constantly vehicle information through *e.g.* signaling cascades or allosteric transitions. The level of performance of biopolymers has quite understandably represented a huge source of inspiration for chemists to develop artificial systems capable of some sort of translation[1-4] or to transport the information associated with a chemical signal through concerted conformational changes.[5,6]

The amount of information that can be stored in sequences, even those written with limited alphabets – only four letters for nucleic acids – is enormous, so much so that DNA itself has been considered for digital data storage miniaturization.[7] Furthermore, advances in nucleic acid and protein sequencing technologies, some of which work at the single molecule level,[8] have promoted the use of biopolymers as information tags. For example, one bead-one compound (OBOC) chemical libraries[9-11] developed in the context of pharmaceutical research can be deconvoluted by labelling each bead with a chemical tag such as an oligonucleotide[12] or an α-peptide,[13] whose synthesis and analysis can be carried out with high fidelity (Figure 1A). DNA tags may also be directly covalently attached to the molecule they encode, *e.g.* in DNA-encoded libraries that exploit the fact that DNA can be amplified by PCR.[14-16] Alongside, interest has risen in polymer chemistry for sequence-defined polymers which can also be used as tags or simply store information.[17-19] Thus, considerable efforts are being devoted to the production of synthetic sequences beyond those of α-peptides and nucleotides, which may be resistant to more drastic conditions than biopolymers, and that may nevertheless be decoded through sequencing methods, primarily by mass spectrometric fragmentation and tandem analysis.[20-24] Mixtures of molecules,[25,26] including mixtures of peptides[27] or sequence-defined polymers[28-29] have recently been proposed for efficient data storage and readout. In some cases, a simple mass spectrum allows for information decoding.

Despite its demonstrated power, the use of chemical tags to encode multiple, possibly mixed molecules, *i.e.* to

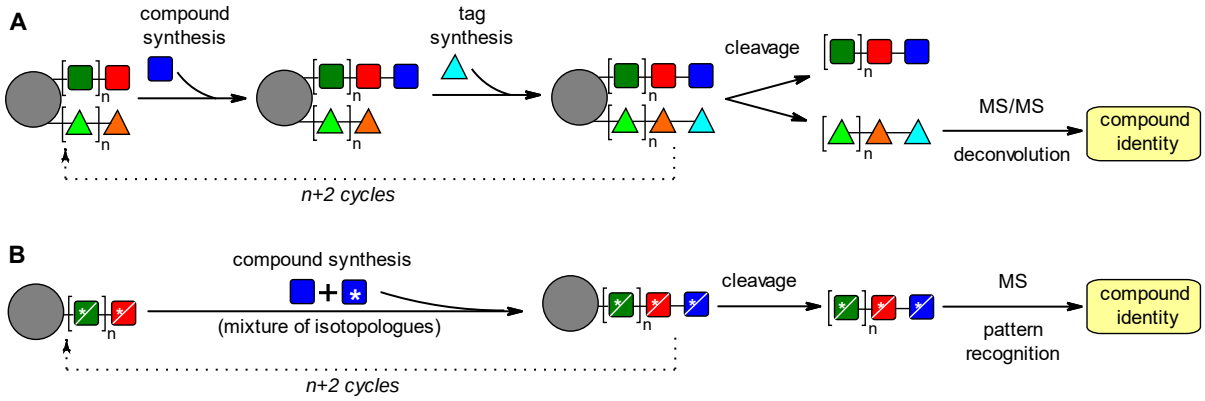


Figure 1. Comparison of chemical encoding and isotope ratio encoding. In OBOC synthesis, chemical encoding (A) entails the synthesis of the tag in addition to the synthesis of the corresponding compound and decoding via fragmentation and identification of the fragments by MS/MS analysis. Isotope ratio encoding (B) avoids the use of tags and allows for a direct readout of compound identity without fragmentation.

embed readable information about the identity of each molecule, bears inherent limitations in certain cases. For instance, the tags may themselves interfere with the interactions with the target. Furthermore, the tags and the molecules that are being labelled are independent chemical entities which must be constructed concomitantly (Figure 1A). This requires the use of orthogonal transformations and might compromise the chemical integrity of the molecules. As a complementary method for such difficult cases, we explore here the concept to include the code as part of the molecule itself and embed information about its nature in its isotope composition. Then compound growing steps are also encoding steps, and the mass spectrum of the final product can be read like a fingerprint, analogous to a bar code, from a single bead and a single measurement.

The concept of isotope encoding was actually presented 25 years ago as a means to improve the readability of peptidic tags.[30] However, this was before extremely sensitive and accurate mass spectrometers became routine laboratory instruments and before the development of statistical and ranking tools for analyzing MS fingerprints.[31] In a recent publication, Anslyn and co-workers also used isotope ratio encoding to label eight different sequences so that sequence information could be retrieved by iterative depolymerization and subsequent LC-MS analysis of the products even though the sequences were mixed.[28] Those isotope ratios needed to be easy to read as their role was to distinguish all depolymerized intermediates of each sequence. To the best of our knowledge, the concept was not implemented further to encode substantial amounts of information, such as the nature of the sequence itself, by mass fingerprints.

In this article, we present the construction and validation of an example isotope ratio encoding system relying on the recognition by a single MS measurement of isotopic fingerprints having yet unparalleled complexity. The method is based on the following, simple workflow: 1) Encoding design and optimization; 2) Labelled monomer synthesis; 3) Encoded oligomer synthesis; 4) MS analysis of oligomers; 5) MS pattern recognition, compound identification. We show that a set of building blocks labelled with either zero, one, two, three or four deuterium atoms allows for the reliable isotope ratio encoding of the 256 sequential combinations comprised of four building blocks. The encoding is efficient despite the building blocks being identical, except for their isotopic ratio.

RESULTS AND DISCUSSION

Molecular design and synthesis. Our proposed strategy is to encode information in isotope ratios and to decode this information using mass spectrometry (MS). Specifically, we propose to encode the identity of each molecular entity by a defined ratio of isotopologues – in our case, molecules differing only in their hydrogen-deuterium composition – beyond the isotopologues already present in natural abundance. The isotopologues should thus generate a series of MS signals of different intensities within a narrow mass range, a so-called MS fingerprint.

When encoding an oligomeric sequence, the mass range depends only on the atomic composition and is insensitive to the constitutional differences. As a consequence, the distinction of oligomers where the individual building blocks have the same atomic composition poses the highest challenge for isotope ratio encoding. Figure 2A shows the molecular formula we selected for encoding a sequence of 4 building blocks of the same atomic composition $C_{12}H_{12}N_2O$ connected by amide bonds, and an arbitrary selection of building blocks having that atomic composition that would be undistinguishable in their non-encoded form by mass spectrometry.

In theory, isotope ratio encoding is applicable to any given oligomer. From a practical perspective, the data-encoding molecular units should be amenable to selective isotope labelling, ensure chemical stability of the isotope code (*i.e.* resist H/D exchange reactions), and possess easily tunable physico-chemical properties, *e.g.* via some sort of functionalization, to accommodate the needs of a targeted application. We selected aromatic oligoamides of 8-amino-2-quinoline carboxylic acids (noted Q) with 4-aminoalkoxy side chains to validate our approach (Figure 2B). Such compounds[32-36] as well as other aromatic oligoamides[37-42] have been shown to adopt folded conformations and interact with proteins in a sidechain dependent manner and may be candidates for OBOC strategies.

Fmoc-protected Q monomers allow for quick access to different oligoamide combinations via solid phase synthesis (SPS).[43] We first developed efficient methods to install from one to four deuterium atoms on the quinoline ring. Thus FmocQ^{D0}(Boc)OH (**1**)

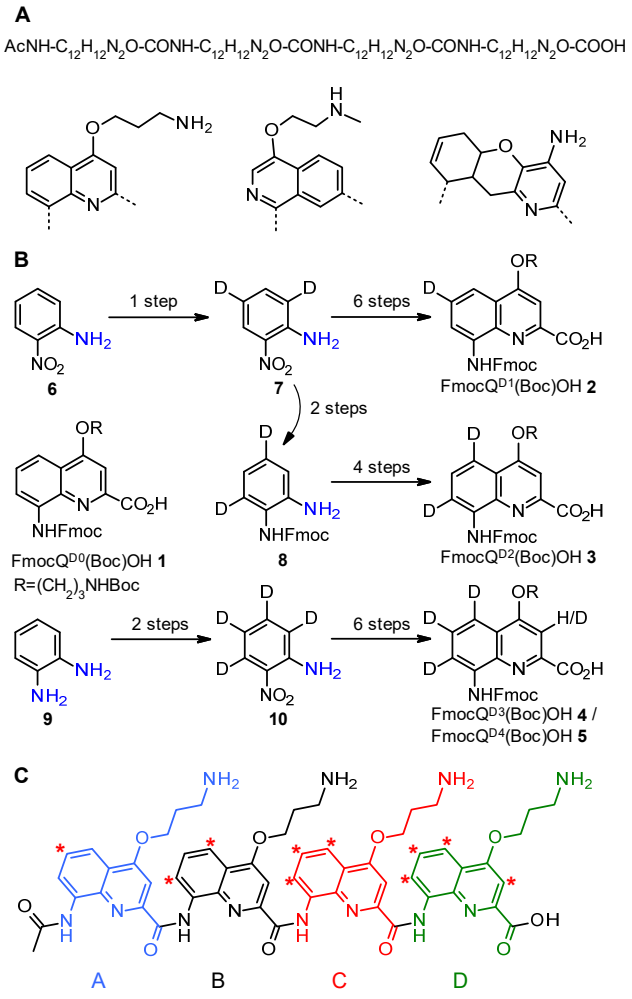


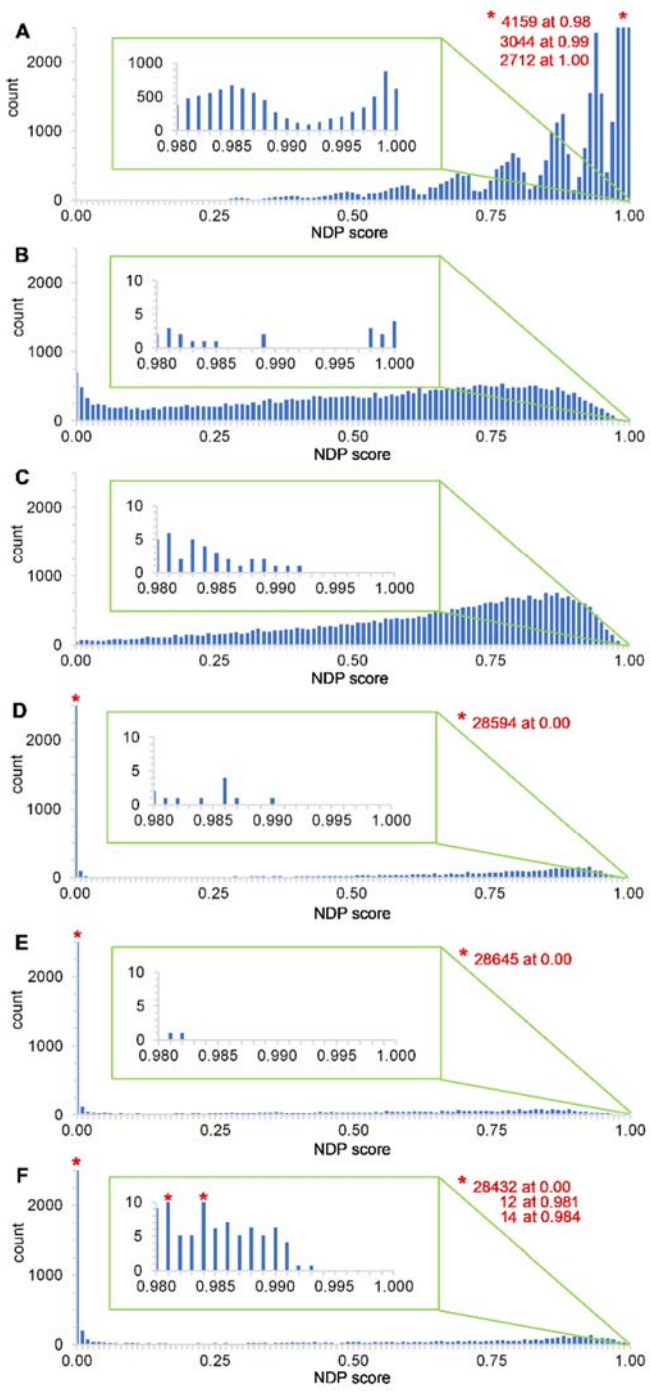
Figure 2. Deuterated monomers and oligomers. (A) The composition of the sequence used in the encoding studies and some examples of virtual building blocks that share the same atomic composition. (B) General formula of a non-deuterated Q monomer with a Boc protected 4-aminopropoxy side chain. FmocQ^{D0}(Boc)OH 1. Outline of the synthetic approach to the mono-, di-, tri-, and tetradeuterated analogues of 1 (2-5, see Figure S1 for details). (C) Generic acetylated tetrameric sequence actually synthesized for encoding. The letters A, B, C, and D correspond to monomer types having potentially different structures (as in (A)) but identical atomic composition. Red stars indicate possible deuteration sites.

and its isotopologues, FmocQ^{D1}(Boc)OH (**2**), FmocQ^{D2}(Boc)OH (**3**), FmocQ^{D3}(Boc)OH (**4**), and FmocQ^{D4}(Boc)OH (**5**) (Figure 2B) were prepared on the multigram scale with high isotopologue selectivity. To this end we followed a precursor labelling strategy, introducing deuteration in the first steps of the synthesis. Adapting the described synthesis of non-labelled monomers,[43-44] we used isotope labelled anilines bearing a nitro or protected amino group in the *ortho* position as source of deuterium atoms in the carbocycle. Selective electrophilic deuteration of 2-nitroaniline (**6**) in positions 4 and 6 afforded **7** with high selectivity, which was used for the preparation of **2**. To synthesize **3**, we had to shift deuteration from the future bridgehead position. Swapping of the amine protection of **7** gave **8** which bear the desired labelling pattern. Units **4** and **5** required the use of tetradeuterated nitroaniline. The direct exchange of all aromatic protons of 2-nitroaniline was not feasible. Instead, a more electron-rich precursor, *ortho*-phenylenediamine was perdeuterated and then oxidized to afford **10**. Using standard cyclization conditions, this intermediate was converted to **4** in six steps. Modification of the cyclization conditions by using deuterated solvents led to the incorporation of an additional deuterium atom in position 3 of the quinoline, thus affording **5** at the end of the reaction sequence. The deuterium content was closely monitored throughout the whole reaction sequence allowing to establish high and selective deuterium incorporation as well as its preservation throughout the synthesis. Furthermore, deuterium content was also stable to the conditions used for SPS, *i.e.* acid chloride activation, piperidine-mediated cleavage of Fmoc groups and TFA cleavage from the resin and Boc side chain deprotection (see the last section). The actual deuterium content of **1**-**5**, was assessed by MS. The deuteration rates were as follows: 90% tetradeuteration for **5** (10% of trideuteration); 98% trideuteration for **4** (2% of dideuteration); 96% dideuteration for **3** (4% of monodeuteration); and 99% of monodeuteration for **2** (Table S2). These measured values were used in subsequent calculations.

To test our method, we challenged the encoding of 4^4 (256) different tetrameric sequences that can be produced by using four monomers of the same atomic composition, hence all 256 combinations would have exactly the same mass without isotope labelling, so composition is not indicated by the mass itself. Figure 2C defines the acetylated tetrameric sequence studied (A, B, C and D; here the letter D stands for a monomer type - when referring to the deuterium symbol, a number follows: D_1, D_2…).

The density of the stored information in isotope encoding can be influenced by two factors: the number of MS peaks in the fingerprint used for coding a single unit (mass window), and the number of coding units combined to provide a code. By using D_0, D_1, D_2, D_3, and D_4 isotopologues in different ratios, we may encode any single unit over a mass range of up to five mass units. With each additional unit, we expand the mass window by 4 mass units and increase the coding capacity. Thus, the tetramers we set to encode should all appear in the 17 amu wide mass window stretching between the D_0 and D_{16} sequences' peaks.

Comparison of coding methods. We set to identify a suitable coding method of 256 data points over 1+16 mass units in a readable manner. Encoding a tetramer, *i.e.* defining which of the four A-D monomers is present at any of the four positions, requires to define sixteen distinct monomer isotopologue combinations (shown in the form of *code tables* in Figure 3). To make the actual implementation of the code practical, *i.e.* easy to realize experimentally, we considered monomers combining D_0 and only one of the D_1-D_4 isotopologues. One might argue that this practical constraint leads to suboptimal coding by not spreading the generated codes evenly across the D_0-D_{16} isotopologue space but, as we will see, the encoding of 256 elements is easily achievable this way. One then translates the code table into an isotopologue distribution by combining the isotope distribution of the four relevant codes for any tetramer. This means, for example, that the code CADB on Figure 3B translates to an acetylated tetramer that contains the following combination of building blocks: Ac-C(0.33D_3,0.77D_0)-A(0.66D_1,0.34D_0)-D(1.00D_4,0.00D_0)-B(0.33D_2,0.77D_0)-OH (acetylation is indicated in the composition table as a non-coding element). Thus, in practice the specified combinations of isotopologues must be used for each sequence elongation step, except for position 3 (D) which is encoded by a single isotopologue in CADB.



Code table

Monomer		A	B	C	D
Isotopologues		D_1 %	D_1 %	D_1 %	D_1 %
Mixing ratio at position	1	7	33	60	87
	2	13	40	67	93
	3	20	47	73	100
	4	27	54	80	0

Code table

Monomer		A	B	C	D
Isotopologues		D_1 %	D_2 %	D_3 %	D_4 %
Mixing ratio at position	1	100	0	33	66
	2	66	100	0	33
	3	33	66	100	0
	4	0	33	66	100

Code table

Monomer		A	B	C	D
Isotopologues		D_1 %	D_2 %	D_3 %	D_4 %
Mixing ratio at position	1	100	75	25	50
	2	75	100	50	25
	3	50	25	100	75
	4	25	50	75	100

Code table

Monomer		A	B	C	D
Isotopologues		D_1 %	D_2 %	D_3 %	D_4 %
Mixing ratio at position	1	100	75	25	50
	2	75	100	50	25
	3	50	25	100	75
	4	25	50	75	100

Code table

Monomer		A	B	C	D
Isotopologues		D_1 %	D_2 %	D_3 %	D_4 %
Mixing ratio at position	1	100	0	33	66
	2	66	100	0	33
	3	33	66	100	0
	4	0	33	66	100

Code table

Monomer		A	B	C	D
Isotopologues		D_1 %	D_2 %	D_3 %	D_4 %
Mixing ratio at position	1	100	0	33	66
	2	66	100	0	33
	3	33	66	100	0
	4	0	33	66	100

G

Composition 1

	C	H	N	O
A	13	13	3	2
B	13	13	3	2
C	13	13	3	2
D	13	13	3	2
non-coding	2	4	0	2

Composition 2

	C	H	N	O
A	13	13	3	2
B	14	15	3	2
C	15	17	3	2
D	16	19	3	2
non-coding	2	4	0	2

Composition 3

	C	H	N	O
A	13	13	3	2
B	14	15	3	2
C	15	17	3	2
D	16	19	3	2
non-coding	58	64	12	8

Figure 3. Comparison of different isotope ratio encoding methods. (A-F). Abundance of NDP scores when comparing pairwise all 256 tetrameric sequences comprised of units A-D. In each case, a table at right indicates which isotopologues are used and in which proportion (% with respect to D_0) to encode units A, B, C, or D when they are in position 1, 2, 3, or 4 of the sequence. In the diagrams, bar width is 0.01 NDP units, and 0.001 NDP units in the inset. (G) Molecular formulas of A-D used in the calculations. Composition 1 was used for calculations (A), (B) and (C). It consists of four monomers A-D with identical molecular formulas and a constant part (terminal acetyl and hydroxyl groups). Composition 2 was used in calculations (D) and (E). Here B-D each possesses from one to three additional CH_2 groups compared to A, respectively. Composition 3 was used in calculation (F). It is similar to composition 2 with the addition of a noncoding part much larger than simple terminal acetyl and hydroxyl groups.

To encode a combination of 4 building blocks over 4 positions, one can devise multiple coding methods (code tables). A fundamental challenge for any coding method is the unambiguous distinction of the different MS fingerprints. We assessed the efficiency of a given code table by calculating the 256 encoded MS fingerprints and making a pairwise comparison of their similarity. Recent developments in mass spectrometry (driven mostly by the widespread use of proteomics) resulted in the introduction of multiple statistical and ranking functions for this purpose.[31] We chose the "Normalized Dot Product (NDP)"

function. Besides being widely used in proteomics search algorithms,[45-47] NDP's use was also successfully extended to the comparison of the mass spectra of small molecules.[48-49] An NDP score can take up a value between 0 and 1. The closer the NDP value is to 1, the higher the similarity, and the more difficult it is to distinguish the pair. Thus, for each of the studied coding methods we used the NDP function for the pairwise comparison of each 256 MS fingerprints. That is 32640 different pairs in total ((256×255)/2).

This gave us a distribution of the MS fingerprint similarities as well as the similarity of the closest fingerprints, which should be the most difficult to distinguish experimentally. The so-obtained information allows for the comparison of different coding methods. To perform the large volume of calculations required for the above actions and to plot the results of statistical analysis, we developed a Microsoft Excel spreadsheet (see Macro S1.xlsm) supported with macros and a semi-automatic integration of the enviPat isotope fine structure calculator from Eawag.[50] This tool can handle encoded libraries of oligomers up to five units built up using combinations of D_0-D_5 monomers bearing a wide range of chemical and isotopic composition. It is universal and can be used for any encoding applications (see user's manual in the SI). Tools designed to handle larger libraries may in principle be built using the same approach.

We first considered a simple coding method in which the nature of each building block and its position in the sequence is encoded by combining the D_0 and D_1 isotopologues in a pre-defined ratio (Figure 3A). As shown in the code table in Figure 3A, this entails an incremental proportion of the D_1 isotopologue, 0%, 7%, 13%, 20%, 27% etc. for each of the sixteen distinct monomer isotopologue combinations. This encoding extends the mass window of the tetramers by only four mass units. As the NDP distribution in Figure 3A shows, this coding performs poorly and generates thousands of pairs of sequences of very similar isotopic distribution.

Next, we considered encoding the nature of each monomer by combining D_0 and a single $D_{1\text{-}4}$ isotopologue, *i.e.* D_0 and D_1 for A, D_0 and D_2 for B…, and the position of each monomer in the sequence by different $D_{1\text{-}4}/D_0$ ratios. For the ratio encoding, we first tested 0%-33%-66%-100% of deuterated monomer (Figure 3B). Here the overall mass window is expanded by 16 units. Coding was found to perform much better than encoding with D1 isotopologues only. Yet, encoding still generated about ten pairs of very similar sequences, with an NDP score very near to 1 (Figure 3B). Analysis of these cases revealed that allowing for both 0% and 100% D_0 at the different coding positions in conjunction with all building blocks having the same mass is responsible for the highly similar codes. Therefore, we adjusted the isotopologue ratios to 25%-50%-75%-100% of deuterated monomer (Figure 3C). This change successfully eliminated the highly similar code pairs from our model and provided an efficient coding, which was validated in subsequent proof-of-concept experiments.

Experimental validation of the coding method. Based on the model calculations, we selected the coding method depicted in Figure 3C for experimental implementation. In this set-up, the mass difference of the non-deuterated building block and the specifically deuterated isotopologue is unique for each A (Δ = 1 amu), B (Δ = 2 amu), C (Δ = 3 amu), and D (Δ = 4 amu). This means for example that the code BDCA translates to an acetylated tetramer that contains the following combination of building blocks: Ac-B(0.75D_2,0.25D_0)-D(0.25D_4,0.75D_0)-C(1.00D_3,0.00D_0)-A(0.25D_1,0.75D_0)-OH.

The proof-of-concept experiments required the synthesis and mass spectrometric characterization of representative encoded tetramers. Following the comparison of the 256 calculated isotope fingerprints arising from the use of the selected code (Figure 3C), we chose four pairs of sequences (**11**-**12**,**13**-**14**,**15**-**16**,**17**-**18**) with a very high NDP score (from 0.976 to 0.992), *i.e.* a priori difficult to distinguish. We also included codes **19** and **20** as a pair with average similarity (NDP score of 0.834). Figures 4A-4E give a visual impression of the actual similarity between the calculated MS fingerprints for each of the five pairs of sequence. We then set out to synthesize and subsequently analyze by MS the ten tetramers (**11**-**20**), including the most similar DCAB (**11**) / DCBB (**12**) pair.

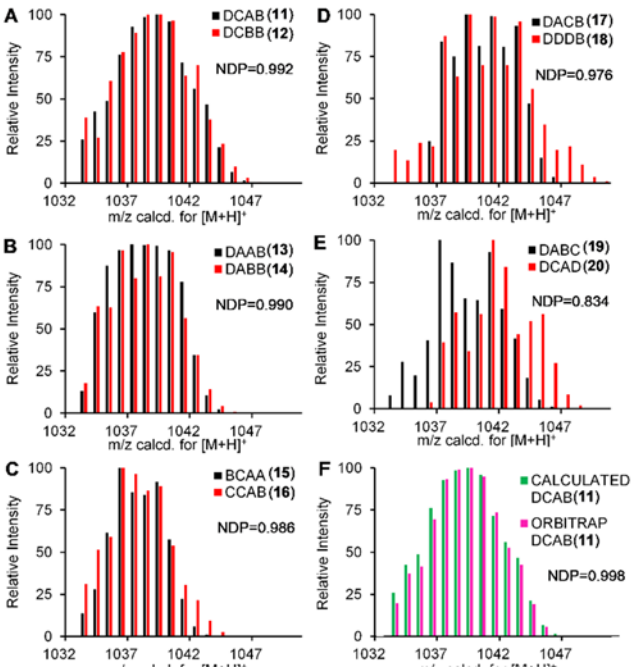


Figure 4. Comparison of the calculated MS fingerprints of the five synthesized sequence pairs (A-E). Comparison of the calculated and measured (*via* Orbitrap) spectra of sequence **11** DCAB (F).

Table 1. Deconvolution of the measured MS fingerprints measured on two different instruments (Orbitrap and TOF) into coding sequences.[1]

	Orbitrap		TOF	
Molecule analyzed	Most similar sequences[a]	NDP	Most similar sequences[a]	NDP

	Orbitrap		TOF	
Molecule analyzed	Most similar sequencesa	NDP	Most similar sequencesa	NDP

11	DCAB (**11**)	0.9985, 0.9979[b]	DCAB (**11**)	0.9926
	DCBB (**12**)	0.989	DCBB (**12**)	0.984
	CDDB	0.975	CADB	0.980
	DDAB	0.974	CDDB	0.975
	DBAB	0.973	BCAB	0.974
12	DCBB (**12**)	0.9955, 0.9899[c]	DCBB (**12**)	0.9958
	DCAB (**11**)	0.993	DCAB (**11**)	0.992
	DDAB	0.984	BCAB	0.977
	DABB (**14**)	0.979	DBAB	0.977
	BCAB	0.979	DDAB	0.973
13	DAAB (**13**)	0.9977	DAAB (**13**)	0.9942
	DABB (**14**)	0.987	BCAB	0.989
	BCAB	0.983	ACAB	0.986
	DDAB	0.982	DABB (**14**)	0.985
	DCAB (**1**)	0.976	BCBB	0.979
14	DABB (**14**)	0.9972	DABB (**14**)	0.9959
	DAAB (**13**)	0.990	DAAB (13)	0.986
	DDAB	0.980	BCAB	0.983
	BCAB	0.980	DDAB	0.981
	CCAB (**16**)	0.977	BCBB	0.973
15	BCAA (**15**)	0.9994	BCAA (**15**)	0.9936
	CCAB (**16**)	0.984	BAAB	0.989
	BCBA	0.981	ACAB	0.982
	ACAB	0.977	BABB	0.980
	BAAB	0.977	CBAB	0.979

16	CCAB (**16**)	0.9951	CCAB (**16**)	0.9956
	ACBB	0.982	BCAA (**15**)	0.989
	BCAA (**15**)	0.980	ACAB	0.984
	BDAB	0.978	BDAB	0.984
	ACAB	0.976	ACBB	0.983
17	DACB (**17**)	0.9933	DACB (**17**)	0.9949
	DDDB	0.974	ACDB	0.974
	CCDB	0.971	CCDB	0.969
	CDDB	0.970	BCDA	0.966
	ACDB	0.965	BCDB	0.961
18	DDDB (**18**)	0.9981	DDDB (**18**)	0.9917
	DACB (**17**)	0.982	DACB (**17**)	0.985
	BDDB	0.964	DCDB	0.972
	DCDB	0.964	BCDB	0.966
	CCDB	0.958	BDDB	0.962
19	DABC (**19**)	0.9990	DABC (**19**)	0.9945
	DAAC	0.977	DAAC	0.977
	DDAC	0.968	ACAC	0.971
	CCDA	0.967	DBAB	0.967
	DBAB	0.967	CCDA	0.965
20	DCAD (**20**)	0.9977	DCAD (**20**)	0.9921
	DADC	0.980	DADC	0.972
	DCBD	0.967	ACDC	0.964
	ACDC	0.964	DCBD	0.962
	CADC	0.964	BDCC	0.961

[a] The five sequences out of 256 codes that have the highest NDP score when compared to each measured MS fingerprint. [b] Measured form a single bead, average of 4 measurements. [c] Measured form a single bead, average of 3 measurements.

Acetylated tetramers **11**-**20** were prepared via microwave-assisted SPS on low loading Wang resin using established protocols.[46] After TFA-mediated cleavage from the resin and Boc group removal, the final products were well water-soluble due to the ammonium groups of the side chains and could be directly analysed by mass spectrometry. To test the robustness of the measurements, the MS fingerprints of the crude **11**-**20** were recorded on two different instruments, an Orbitrap and a TOF mass spectrometer (Tables S4 and S5). Both instruments possess a high-resolution mass analyzer but may be considered routine spectrometers in that they are present in most analytical facilities (see SI for details). The spectrometers should precisely deliver the isotopic distribution, but accurate m/z measurement is not required. Each of the measured MS fingerprints were compared to the 256 theoretical MS fingerprints by calculating the respective NDP values. For each sequence **11**-**20**, the five calculated MS fingerprints having the highest NDP score are listed in Table 1 for both the Orbitrap and TOF MS instruments. In all ten cases on both instruments, the sequence encoded by the analyzed molecule could be identified from its MS fingerprint: the most similar calculated fingerprint was that of the correct sequence, with an average best NDP score for the ten compounds of 0.997 for the Orbitrap and 0.994 for the TOF. A representative example of the similarity between calculated and measured spectra for a given sequence is shown in Figure 4F.

Thus, deconvolution proved to be very efficient and error free. Given that **11**-**18** are amongst the most difficult-to-distinguish sequences, *i.e.* there is at least one MS fingerprint among the 255 others that is very similar to their fingerprint, one can reasonably conclude that all 256 sequences were successfully and unambiguously encoded over a mass range of 1+16 units by the proposed encoding method.

The similarity of the measured and calculated MS fingerprints was always high, characterized by an NDP score above 0.992. In case of the more similar sequences, the second most similar MS fingerprint came close in NDP score: the differentiation between the first and second ranking hit was 0.003 (Orbitrap) or 0.004 (TOF) for the closest pair, followed by 0.005 in two instances. For example, due to their high similarity, **11** and **12** appear as each other's second most probable code. Moving to **13** and beyond, the similarity of the second-best guess deteriorates (NDP ≤ 0.990) which increases the confidence in the result of the deconvolution. We also note that the four lower ranked sequences vary from one instrument to the other when their similarity to the correct one is low. This indicates minute variations

in the measured isotope distribution between the TOF and the Orbitrap spectrometers. These slight variations were assigned to the presence of a low intensity secondary set of peaks, corresponding to $[2M+2H]^{2+}$ dimers, that overlap with the peaks of $[M+H]^{+}$. Dimer formation in the gas phase depends on experimental mass spectrometric conditions and may be suppressed by adjusting ionization parameters. Our experiments show efficient decoding without taking these into consideration, although this might be needed at higher proportions of dimer.

We also tested the robustness of our results with respect to several sources of experimental error. For example, the deuteration level of **2**-**5** that was used in the calculation may be subject to slight errors when measured by MS. Conversely, the isotopologue combinations were prepared by manually weighing samples of **1**-**5** on a microbalance and proportions might in practice slightly deviate from exactly 25/75, 50/50, or 75/25. Making a 1% error in the measurement of the deuteration rate was found to have negligible consequences. For instance, setting all deuteration rates 1% lower than their actual value, *i.e.* 98% monodeuteration for **2**, 95% dideuteration for **3**, 97% trideuteration for **4**, and 89% tetradeuteration for **5**, produced MS fingerprints that were extremely similar to the actual one. When comparing the theoretical fingerprint with that resulting from the error for each 256 sequences, the NDP score was 0.99947 in the worst case (for ABCD) and 0.99987 on average. If the error in measuring the deuteration rate is larger, encoding-decoding efficiency may slightly deteriorate. Thus, setting all deuteration rates 5% lower than their actual value (an error considered easy to avoid), *i.e.* 94% monodeuteration for **2**, 91% dideuteration for **3**, 93% trideuteration for **4**, and 85% tetradeuteration for **5**, produced MS fingerprints that differed somewhat from the actual one. The NDP was 0.98723 in the worst case (again for ABCD), and 0.99695 on average, which remains sufficient for discriminating most sequences. Of note, if these lower deuteration rates were actual, *e.g.* due to a less efficient deuteration chemistry, and if they were accurately measured, the reliability of encoding is negligibly altered (one NDP of 0.995 for the most similar fingerprints, *vs.* 0.992 for **11** and **12** with the level of deuteration we reached (see Figure S2). Deuteration thus needs not to be quantitative for efficient encoding.

Similarly, using slightly erroneous mixing ratios (*e.g.* a combination of 26/74, 76/24, 51/49) led to MS fingerprints essentially identical to those expected (NDP of 0.99948 in the worst case (for CDBA) and 0.99983 on average). Such experimental error would not impair sequence identification as achieved for **11**-**20**. If the experimental error of mixing ratio is larger (combinations of 30/70, 45/55, and 80/20 instead of the expected 25/75, 50/50 and 75/25) MS fingerprints deviate to the point that coding may be lost for certain sequences. The NDP is 0.98744 in the worst case (for CDBA) which is lower than the similarity between measured and calculated fingerprints required to identify **11**-**18** (but not **19** and **20**). Yet even with these large errors in preparing isotopologue combinations, the average NDP is 0.99582 which indicates that most fingerprints would still be correctly identified.

Assessment of isotope ratio encoding reproducibility and reliability. Indication that isotope labelling does not erode during the SPS procedures used for code generation came from the successful decoding of tetramers **11**-**20** from their mass fingerprints described above. It was nevertheless formally validated with $FmocQ^{D4}$(Boc)OH (**5**) which contains all possible sites of erosion. First, **5** was loaded on Wang resin and three SPS cycles were performed using the FmocQ(Boc)OH (**1**) monomer to elongate the chain. In every SPS cycle, a part of the resin was removed and the bound mono-, di-, tri- or tetramer was cleaved with TFA, which also removed the Boc protecting group, affording compounds **21**-**24** (Figure 5A). Their mass spectra were recorded, and the isotope labelling was compared to that of the initial monomer, showing no erosion of the deuteration within measurement error (Table S3). Since the first monomer is coupled to the resin using conditions slightly different from those for subsequent couplings, dimer **25** in which the N-terminal unit is deuterated (Figure 5B) was also prepared and analyzed. Changing the position of the labelled monomer did not result in erosion either (Table S3).

To assess the sensitivity and reproducibility of MS experiments, a series of repeated MS measurements were carried out on building block $FmocQ^{D3}$(Boc)OH (**4**) as well as on tetramer **11** (DCAB, Table 2). Repeated measurements of the MS fingerprint were followed by the calculation of the relative peak intensities in each series and the statistical analysis of the distribution of relative intensities for a given m/z value. From the statistical data confidence intervals of 95% and 99% were calculated for each peak. To calculate the inherent limitation of similarity comparison arising from the non-perfect reproducibility of the MS measurements, we calculated the similarity (NDP score) of two MS fingerprints: in the first the peak intensities were the average minus the respective confidence interval and in the second the average plus the confidence interval for each respective peak. For **4**, we obtained NDP scores of 1.0000 both at 95% and at 99% confidence, while for **11** we calculated NDP scores of 0.9997 and 0.9993, respectively. To our delight, single bead analyses of **11** gave results that were comparable to the ones obtained from bulk analysis. This finding supports the applicability of the method for tagging bead-based libraries.

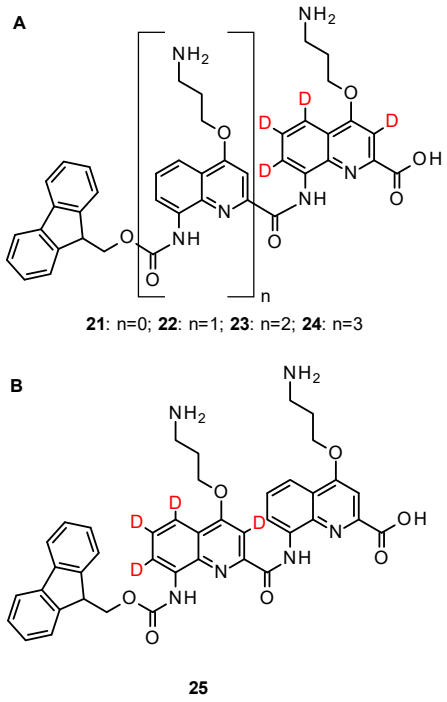


Figure 5. Oligomers for testing isotope labeling erosion during coupling and deprotection.

Table 2. Statistical analysis of the repeated measurements of MS fingerprints of a building block, and tetramer 11 (DCAB) in bulk and single bead-based analysis.

Cpd /n[a]		FmocQ^{D3}(Boc)OH (**4**) /10		DCAB /12 (**11**)		DCAB /4 [single bead] (**11**)	
		Average	SD[a]	Average	SD	Average	SD
Relative peak intensities	M[a]	1.84	0.10	20.09	0.74	20.61	0.25
	M+1	100.00	0.00	34.22	2.35	34.87	0.12
	M+2	35.71	0.41	48.63	1.63	48.83	0.35
	M+3	7.89	0.15	67.76	2.48	68.67	0.68
	M+4	1.41	0.04	88.85	1.15	89.10	0.41
	M+5	-	-	97.85	4.37	98.38	0.53
	M+6	-	-	100.00	0.00	100.00	0.00
	M+7	-	-	96.31	3.22	96.26	0.48
	M+8	-	-	76.03	3.17	75.89	0.34
	M+9	-	-	60.52	3.51	60.33	0.16
	M+10	-	-	45.08	0.69	44.16	0.14
	M+11	-	-	26.68	1.04	26.85	0.65
	M+12	-	-	9.48	1.96	10.24	0.09
	M+13	-	-	1.94	1.21	1.81	1.30
S95[a]	(++…/--…)	1.0000		0.9997		0.9998	
	(+-…/-+…)	1.0000		0.9991		0.9997	
S99[a]	(++…/--…)	1.0000		0.9993		0.9993	
	(+-…/-+…)	1.0000		0.9982		0.9990	

[a] n: number of measurements; M: peak with the lowest mass registered in the MS fingerprint; SD: standard deviation; S95: NDP value for the MS fingerprints at the two extrema of the 95% probability range; S99: NDP value for the MS fingerprints at the two extrema of the 99% probability range.

To further distort the shape of the MS fingerprints and decrease their similarity, we generated two further MS fingerprints. In the first, the peak intensities were the average minus the respective confidence interval for the M, M+2, and M+4 peaks and the average plus the relative confidence interval for the M+1 and M+3 peaks. In the second the peak intensities were the average plus the respective confidence interval for the M, M+2, and M+4 peaks and the average minus the relative confidence interval for the M+1 and M+3 peaks. The calculated NDP scores showed no difference for **4** (1.0000 both at 95% and 99% confidence), while for **11** the calculated similarity decreased manifesting in NDP scores of 0.9991 and 0.9982, respectively. These differences remain marginal and should not impact the distinguishing of MS fingerprints.

Finally, to assess the potential deuterium erosion on standing (durability), we repeated the MS analysis of monomers units **2-5** and coding sequences **11** and **12** after 22-55 months of standing. The measurements showed that the deuterium content remained essentially unchanged (Table S1, Table S2).

Scope and potential developments of the approach. The results above validate the concept of encoding information in the isotopologue ratio of a molecule. Unambiguous encoding could be implemented using one distinct isotopologue (*e.g.* +1, +2, +3, or +4 amu) mixed with the corresponding unlabeled building block for every unit introduced in the final molecule. For many simple building blocks, including *e.g.* amino acids, several isotopologues may be commercial or readily available with minimal synthetic effort. Also, the level and variety of deuteration may be reduced when the building blocks are not isomers but using only monodeuterated monomers is not sufficient for perfect encoding (see below).

The detection step requires minimal amounts of material since it can be performed even on a single bead. Moreover, sequencing techniques such as enzymatic digestion, iterative degradation or MS/MS fragmentation are not needed prior to analysis, as the mass fingerprints are obtained by a single MS measurement of oligomers. For the identification step, the method entails the comparison of the isotopologue distributions of compounds of interest with all distributions theoretically produced during the encoding. This is conveniently performed by spreadsheets such as the one we provide. To decode the sequence of a tetramer on one bead takes about 5 minutes of work per sample plus about 1.5 h for cleavage and evaporation. If one deals with hundreds of samples, analysis may be accelerated by parallel sample preparation and automation through coupling the digital output provided by essentially any mass spectrometer to the workflow. As a future development, using an appropriate cleavable linker may enable direct MALDI-MS analysis of bead-bound oligomers, obviating the need prior cleavage.

Isotope ratio encoding does not necessarily stand as a substitute to existing information encoding approaches. It rather appears to be a complement or supplement. For example, some sequences cannot be identified by MS/MS analysis of fragments. This is the case for Q_n oligomers: their fragmentation results in

side chain cleavage. It would also be the case when the successive reaction steps do not generate a linear sequence but a more complex structure. In addition, in the context of OBOC chemistry, the reaction conditions required for compound synthesis tags may be incompatible with the use of tags.

The concept of isotope ratio encoding may be advantageously extended to molecular mixture encoding. In isotope ratio encoding as defined in Figure 3C, each tetrameric sequence had the atomic composition 1 shown in Figure 3G: four monomers with a $C_{13}H_{13}N_3O_2$ molecular formula with its natural isotopic abundance, combined with a D_1-D_4 isotopologue, and a noncoding $C_2H_4O_2$ accounting for the terminal acetyl and OH groups. We also assessed the case when A, B, C, and D differ by $(CH_2)_n$ (n=0-3) units (composition 2 in Figure 3G). In this case, combinations represent different molecules. With an encoding of 25%-50%-75%-100% of deuterated monomer (Figure 3D), the distinguishability of the most similar MS fingerprints increased significantly. Very few pairs of sequences had an NDP above 0.98 and none at or above 0.99. Since the building blocks in this case have different molecular weights, we could further improve separation of the MS fingerprints by extending the isotopologue ratio to 0%-33%-66%-100%. As Figure 3E shows, with this coding, the most alike MS fingerprints have an NDP score of 0.982, the lowest obtained so far. Expanding this encoding method to the 3125 pentamers comprised of five monomers differing by one CH_2 unit and deuterated up to five times also gave very good results (Figure S3). Expectedly, with pentamers, the number of sequence pairs having a high NDP increases: 19 pairs are found to have an NDP>0.995. However, this number of possible ambiguities remains small compared to the total number of combinations. Furthermore, ambiguity never concerns more than two sequences at a time. Experimentally lifting an ambiguity would entail the synthesis and control of only two sequences.

We also assessed the potential limitation of using single isotope encoding of the 256 tetramers comprised of four monomers having distinct masses, as may occur in the context of an OBOC chemical library (Figure S4). Results were somewhat mixed. A total of 84 pairs (out of 32640) are undistinguishable (NDP of 1.00) because of a certain degree of degeneracy of the coding. In some cases, up to four sequences would have an identical isotope pattern. This may not be an ideal situation but it hints at the fact that labelling each monomer with a distinct level of deuteration is not needed when monomers have masses than are easy to distinguish.

Finally, we looked at the effect of appending a large $C_{58}H_{64}N_{12}O_8$ invariable non-coding sequence to the tetramers (composition 3 in Figure 3G). This scenario corresponds to a case where the coding tetramer represents only half of the molecule. For example, one could imagine a situation where four encoding units report on the nature and position of four other units that have not been labelled with deuterium – thus avoiding the need for labelling any monomer of interest. In essence, this amounts to include a tag within the molecular structure. Using the chemically different (homologous) building blocks, the different $D_{1\text{-}4}/D_0$ isotopologue pairs and the 0-100% coding range, we see an increased similarity between the most alike MS fingerprints (Figure 3F) with 6 pairs of sequences having a similarity score above 0.99, and a maximum value of 0.993. This increase is due to the natural isotope distribution of the large non-coding part leading to the rise of isotopologues whose molecular weight overlaps with the coding isotopologue's and increases ambiguity. Nevertheless, the distinction of the overall 256 sequences remains excellent.

CONCLUSION

We have demonstrated that high amounts of information may be reliably encoded over a relatively narrow mass window by isotope ratio encoding using combinations of isotope-labelled units. Decoding is conveniently performed by simple mass spectrometric measurements and similarity analysis. The mass analysis relies on molecular ions of the encoded compounds, thus the method offers a complementary solution in cases where fragmentation of the target compounds is unsuitable for MS/MS sequencing, but tagging should be avoided. The synthesis, encoding and mass spectrometric fingerprint readout that we used all involve standard laboratory transformations and instrumentation, making the method universally accessible and practical. Following this principle, the density of the encoded information – *i.e.* the number of different data points that can be written and read over a certain mass range – may be further increased if needed by expanding the variety of isotopologues. The fine-tuning of the encoding approach, *e.g.* using combinations of more than two isotopologues, might also further enhance accessible data density. Another advantage of isotope encoding is that it is independent of the chemical nature of the molecule, thus not limited to oligoamides. Multiple molecules constructed from several precursors, be they sequences or not, may be labelled in the same manner as the 256 tetramers considered here, using deuterium or isotopes of other atoms. The stability of the isotope label used is the only pre-requisite. The fact that we can encode a large number of data points over a narrow mass window enables us to limit the information code to a smaller portion of a (macro)molecule thus avoiding the need for isotope labelling of all building blocks. A major and yet unexploited advantage of isotope ratio encoding is that most of the relevant physical properties, including molecular interactions, remain practically unaltered in the process, which paves the way to applications in diverse areas such as biology and materials science. It is important to note that the chemical motif used for isotope ratio encoding can also be used as a tag if the access to the deuterated building blocks of the compound of interest is problematic.

The quinoline-based oligomers that served as a model system in our study have outstanding chemical stability as well as tunable physical properties and an ability to participate in molecular recognition processes. They may thus serve as persistent information tags for diverse applications in data storage and anti-counterfeiting, but the main interest of the method is precisely to avoid the need for tags through the labelling of the molecules of interest. For example, in the case of one bead-one compound libraries, reaction sequence information could be written directly into the library compounds allowing for fast synthesis and high throughput screening without concerns about the introduction of additional tags and their potential interference in the selection and identification processes.

ASSOCIATED CONTENT

Supporting Information. Details of computational spectrum prediction and similarity evaluation, experimental details, materials, methods, and characterization data, including copies of NMR spectra (Document S1.pdf). This material is available free of charge via the Internet at http://pubs.acs.org. Supplemental tool for calculations related to MS fingerprints (Microsoft Excel Macro S1.xlsm)

is available free of charge at https://drive.google.com/file/d/1Wjkja11cAtREa658TaZPP0MLt-MSlM5XD/view?usp=sharing, or directly from the authors.

AUTHOR INFORMATION

Corresponding Author

* Ivan Huc: ivan.huc@cup.lmu.de
** András Kotschy: andras.kotschy@servier.com

Author Contributions

The manuscript was written through contributions of all authors.

ACKNOWLEDGMENT

M. Z. thanks the financial support by the ÚNKP-16-2 ELTE/8315/117 (2016) grant from the New National Excellence Program of the Ministry of Human Capacities. This research has been implemented in the frame of project no. FIEK_16-1-2016-0005 „Development of molecular biomarker research and service center", with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the FIEK_16 funding scheme. This work has benefited from the facilities and expertise of the Biophysical and Structural Chemistry platform at IECB, CNRS UMS3033, INSERM US001, Bordeaux University, France.

REFERENCES

(1) Lewandowski, B.; De Bo, G.; Ward, J. W.; Papmeyer, M.; Kuschel, S.; Aldegunde, M. J.; Gramlich, P. M. E.; Heckmann, D.; Goldup, S. M.; D'Souza, D. M.; Fernandes, A. E.; Leigh, D. A. Sequence-specific peptide synthesis by an artificial small-molecule machine. *Science* **2013**, *339*, 189–193. DOI:10.1126/science.1229753.

(2) He, Y.; Liu, D. R. Autonomous multistep organic synthesis in a single isothermal solution mediated by a DNA walker. *Nat. Nanotechnol.* **2010**, *5*, 778–782. DOI:10.1038/nnano.2010.190.

(3) Gan, Q.; Wang, X.; Kauffmann, B.; Rosu, F.; Ferrand, Y.; Huc, I. Translation of rod-like template sequences into homochiral assemblies of stacked helical oligomers. *Nat. Nanotechnol.* **2017**, *12*, 447–452. DOI:10.1038/nnano.2017.15.

(4) McKee, M. L.; Milnes, P. J.; Bath, J.; Stulz, E.; Turberfield, A. J.; O'Reilly, R. K. Multistep DNA-templated reactions for the synthesis of functional sequence controlled oligomers. *Angew. Chem. Int. Ed.* **2010**, *49*,7948–7951. DOI:10.1002/anie.201002721.

(5) Brown, R. A.; Diemer, V.; Webb, S. J.; Clayden, J. End-to-end conformational communication through a synthetic purinergic receptor by ligand-induced helicity switching. *Nat. Chem.* **2013**, *5*, 853–860. DOI:10.1038/nchem.1747.

(6) Morris, D. T. J.; Wales, S. M.; Tilly, D.; Farrar, E. H. E.; Grayson, M. N.; Ward, J. W.; Clayden, J. A molecular communication channel consisting of a single reversible chain of hydrogen bonds in a conformationally flexible oligomer. *Chem.* **2021**, *7*, 2460–2472. DOI:10.1016/j.chempr.2021.06.022.

(7) Ceze, L.; Nivala, J.; Strauss, K. Molecular digital data storage using DNA. *Nat. Rev. Genet.* **2019**, *20*, 456–466. DOI:10.1038/s41576-019-0125-3.

(8) Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, B.; Bettmann, B.; Bibillo A.; Bjornson K.; Chaudhuri B.; Christians F.; Cicero R.; Clark S.; Dalal R.; Dewinter A.; Dixon J.; Foquet M.; Gaertner A.; Hardenbol P.; Heiner C.; Hester K.; Holden D.; Kearns G.; Kong X.; Kuse R.; Lacroix Y.; Lin S.; Lundquist P.; Ma C.; Marks P.; Maxham M.; Murphy D.; Park I.; Pham T.; Phillips M.; Roy J.; Sebra R.; Shen G.; Sorenson J.; Tomaney A.; Travers K.; Trulson M.; Vieceli J.; Wegener J.; Wu D.; Yang A.; Zaccarin D.; Zhao P.; Zhong F.; Korlach J.; Turner S. Real-time DNA sequencing from single polymerase molecules. *Science* **2009**, *323*, 133–138. DOI:10.1126/science.1162986.

(9) Lam, K. S.; Salmon, S. E.; Hersh, E. M.; Hruby, V. J.; Kazmierski, W. M.; Knapp, R. J. A new type of synthetic peptide library for identifying ligand-binding activity. *Nature* **1991**, *354*, 82–84. DOI:10.1038/354082a0.

(10) Houghten, R. A.; Pinilla, C.; Blondelle, S. E.; Appel, J. R.; Dooley, C. T.; Cuervo, J. H. Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature* **1991**, *354*, 84-86. DOI:10.1038/354084a0.

(11) Furka, A.; Sebestyén, F.; Asgedom, M.; Dibó, G. General method for rapid synthesis of multicomponent peptide mixtures. *Int. J. Pept. Protein Res.* **1991**, *37*, 487-493. DOI:10.1111/j.1399-3011.1991.tb00765.x.

(12) MacConnell, A. B.; McEnaney, P. J.; Cavett, V. J.; Paegel, B. M. DNA-Encoded Solid-Phase Synthesis: Encoding Language Design and Complex Oligomer Library Synthesis. ACS Comb. Sci. **2015**, *17*, 518–534. DOI:10.1021/acscombsci.5b00106.

(13) Nikolaiev, V.; Stierandová, A.; Krchnák, V.; Seligmann, B.; Lam, K. S.; Salmon, S. E.; Lebl, M. Peptide encoding for structure determination of nonsequenceable polymers within libraries synthesized and tested on solid-phase supports. *Pept. Res.* **1993**, *6*, 161-170. PMID: 8318748

(14) Brenner, S.; Lerner, R. A. Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 5381-5383. DOI:10.1073/pnas.89.12.5381.

(15) Kunig, V.; Potowski, M.; Gohla, A.; Brunschweiger, A. DNA-encoded libraries - an efficient small molecule discovery technology for the biomedical sciences. *Biol. Chem.* **2018**, *399*, 691–710. DOI:10.1515/hsz-2018-0119.

(16) Shi, Y.; Wu, Y. R.; Yu, J. Q.; Zhang, W. N.; Zhuang, C. L. DNA-encoded libraries (DELs): a review of on-DNA chemistries and their output. *RSC Adv.* **2021**, *11*, 2359–2376. DOI:10.1039/d0ra09889b.

(17) Colquhoun, H.; Lutz, J. F. Information-containing macromolecules. *Nat. Chem.* **2014**, *6*, 455–456. DOI:10.1038/nchem.1958.

(18) Meier, M.; Barner-Kowollik, C. A New Class of Materials: Sequence-Defined Macromolecules and Their Emerging Applications. *Adv. Mater.* **2019**, *31*, 1806027. DOI:10.1002/adma.201806027.

(19) Aksakal, R.; Mertens, C.; Soete, M.; Badi, N.; Du Prez, F. Applications of Discrete Synthetic Macromolecules in Life and Materials Science: Recent and Future Trends. *Adv. Sci.* **2021**, *8*, 2004038. DOI:10.1002/advs.202004038.

(20) Martens, S.; Landuyt, A.; Espeel, P.; Devreese, B.; Dawyndt, P.; Du Prez, F. Multifunctional sequence-defined macromolecules for chemical data storage. *Nat. Commun.* **2018**, *9*, 4451. DOI:10.1038/s41467-018-06926-3.

(21) Roy, R. K.; Meszynska, A.; Laure, C.; Charles, L.; Verchin, C.; Lutz, J. F. Design and synthesis of digitally encoded polymers that can be decoded and erased. *Nat. Commun.* **2015**, *6*, 7237. DOI:10.1038/ncomms8237.

(22) Frölich, M.; Hofheinz, D.; Meier, M. A. R. Reading mixtures of uniform sequence-defined macromolecules to increase data storage capacity. *Commun. Chem.* **2020**, 3184. DOI:10.1038/s42004-020-00431-9.

(23) Al Ouahabi, A.; Amalian, J. A.; Charles, L.; Lutz, J. F. Mass spectrometry sequencing of long digital polymers facilitated by programmed inter-byte fragmentation. *Nat. Commun.* **2017**, *8*, 967. DOI:10.1038/s41467-017-01104-3.

(24) Bathany, K.; Owens, N. W.; Guichard, G.; Schmitter, J. M. Sequencing of oligourea foldamers by tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 458–462. DOI:10.1007/s13361-012-0546-0.

(25) Arcadia, C. E.; Kennedy, E.; Geiser, J.; Dombroski, A.; Oakley, K.; Chen, S. L.; Sprague, L.; Ozmen, M.; Sello, J.; Weber, P. M.; Reda, S.; Rose, C.; Kim, E.; Rubenstein, B. M.; Rosenstein, J. K. Multicomponent molecular memory. *Nat. Comm.* **2020**, *11*, 691. DOI:10.1038/s41467-020-14455-1.

(26) Nagarkar, A. A.; Root, S. E.; Fink, M. J.; Ten, A. S.; Cafferty, B. J.; Richardson, D. S.; Mrksich, M.; Whitesides, G. M. Storing and Reading Information in Mixtures of Fluorescent Molecules. *ACS Cent. Sci.* **2021**, *7*, 1728-1735. DOI:10.1021/acscentsci.1c00728.

(27) Ng, C.; Tam, W. M.; Yin, H.; Wu, Q.; So, P. K.; Wong, M. Y.; Lau, F.; Yao, Z. P. Data storage using peptide sequences. *Nat. Comm.* **2021**, 12, 4242. DOI:10.1038/s41467-021-24496-9.

(28) Dahlhauser, S. D.; Wight, C. D.; Moor, S. R.; Scanga, R. A.; Ngo, P.; York, J. T.; Vera, M. S.; Blake, K. J.; Riddington, I. M.; Reuther, J. F.; Anslyn, E. V. Molecular Encryption and Steganography Using Mixtures of Simultaneous-ly Sequenced, Sequence-Defined Oligourethanes. *ACS Cent. Sci.* **2022**, 8, 1125–1133. DOI:10.1021/acscentsci.2c00460

(29) Frölich, M.; Hofheinz, D.; Meier, M. A. R. Reading mixtures of uniform sequence-defined macromolecules to increase data storage capacity. *Commun. Chem.* **2020**, *3*, 184. DOI:10.1038/s42004-020-00431-9.

(30) Geysen, H. M.; Wagner, C. D.; Bodnar, W. M.; Markworth, C. J.; Parke, G. J.; Schoenen, F. J.; Wagner, D. S.; Kinder, D. S. Isotope or mass encoding of combinatorial libraries. *Chem. Biol.* **1996**, *3*, 679-688. DOI:10.1016/s1074-5521(96)90136-2.

(31) Yilmaz, Ş.; Vandermarliere, E.; Martens, L. Methods to Calculate Spectrum Similarity. *Methods Mol. Biol.* **2017**, *1549*, 75–100. DOI:10.1007/978-1-4939-6740-7_7.

(32) Kumar, S.; Birol, M.; Schlamadinger, D. E.; Wojcik, S. P.; Rhoades, E.; Miranker, A. D. Foldamer-mediated manipulation of a pre-amyloid toxin. *Nat. Commun.* **2016**, *7*, 11412. DOI:10.1038/ncomms11412.

(33) Kumar, S.; Henning-Knechtel, A.; Chehade, I.; Magzoub, M.; Hamilton, A. D. Foldamer-Mediated Structural Rearrangement Attenuates Aβ Oligomerization and Cytotoxicity. *J. Am. Chem. Soc.* **2017**, *139*, 17098–17108. DOI:10.1021/jacs.7b08259.

(34) Ziach, K.; Chollet, C.; Parissi, V.; Prabhakaran, P.; Marchivie, M.; Corvaglia, V.; Bose, P. P.; Laxmi-Reddy, K.; Godde, F.; Schmitter, J. M.; Chaignepain S.; Pourquier P.; Huc I. Single helically folded aromatic oligoamides that mimic the charge surface of double-stranded B-DNA. *Nat. Chem.* **2018**, *10*, 511–518. DOI:10.1038/s41557-018-0018-7.

(35) Buratto, J.; Colombo, C.; Stupfel, M.; Dawson, S. J.; Dolain, C.; Langlois d'Estaintot, B.; Fischer, L.; Granier, T.; Laguerre, M.; Gallois, B.; Huc, I. Structure of a complex formed by a protein and a helical aromatic oligoamide foldamer at 2.1 Å resolution. *Angew. Chem. Int. Ed.* **2014**, *53*, 883–887. DOI:10.1002/anie.201309160.

(36) Reddy, P. S.; Langlois d'Estaintot, B.; Granier, T.; Mackereth, C. D.; Fischer, L.; Huc, I. Structure Elucidation of Helical Aromatic Foldamer-Protein Complexes with Large Contact Surface Areas. *Chemistry* **2019**, *25*, 11042–11047. DOI:10.1002/chem.201902942.

(37) Plante, J. P.; Burnley, T.; Malkova, B.; Webb, M. E.; Warriner, S. L.; Edwards, T. A.; Wilson, A. J. Oligobenzamide proteomimetic inhibitors of the p53-hDM2 protein-protein interaction. *Chem. Commun.* **2009**, 5091–5093. DOI:10.1039/b908207g.

(38) Barnard, A.; Long, K.; Yeo, D. J.; Miles, J. A.; Azzarito, V.; Burslem, G. M.; Prabhakaran, P.; Edwards, T. A.; Wilson, A. J. Orthogonal functionalisation of α-helix mimetics. *Org. Biomol. Chem.* **2014**, *12*, 6794–6799. DOI:10.1039/c4ob00915k.

(39) Jayatunga, M. K.; Thompson, S.; Hamilton, A. D. α-Helix mimetics: outwards and upwards. *Bioorg. Med. Chem. Lett.* **2014**, *24*, 717–724. DOI:10.1016/j.bmcl.2013.12.003.

(40) Saraogi, I.; Hebda, J. A.; Becerril, J.; Estroff, L. A.; Miranker, A. D.; Hamilton, A. D. Synthetic alpha-helix mimetics as agonists and antagonists of islet amyloid polypeptide aggregation. *Angew. Chem. Int. Ed.* **2010**, *49*, 736–739. DOI:10.1002/anie.200901694.

(41) Kumar, S.; Hamilton, A. D. α-Helix Mimetics as Modulators of Aβ Self-Assembly. *J. Am. Chem. Soc.* **2017**, *139*, 5744–5755. DOI:10.1021/jacs.6b09734.

(42) Flack, T.; Romain, C.; White, A.; Haycock, P. R.; Barnard, A. Design, Synthesis, and Conformational Analysis of Oligobenzanilides as Multifacial α-Helix Mimetics. *Org. Lett.* **2019**, *21*, 4433–4438. DOI:10.1021/acs.orglett.9b01115.

(43) Baptiste, B.; Douat-Casassus, C.; Laxmi-Reddy, K.; Godde, F.; Huc, I. Solid phase synthesis of aromatic oligoamides: application to helical water-soluble foldamers. *J. Org. Chem.* **2010**, *75*, 7175. DOI:10.1021/jo101360h.

(44) Gillies, E. R.; Deiss, F.; Staedel, C.; Schmitter, J. M.; Huc, I. Development and biological assessment of fully water-soluble helical aromatic amide foldamers. *Angew. Chem. Int. Ed.* **2007**, *46*, 4081-4084. DOI:10.1002/anie.200700301.

(45) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7*, 655–667. DOI:10.1002/pmic.200600625.

(46) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* **2006**, *5*, 1843–1849. DOI:10.1021/pr0602085.

(47) Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **2006**, *78*, 5678–5684. DOI:10.1021/ac060279n.

(48) Willard, M. A. B.; Smith, R. W.; McGuffin, V. L. Statistical approach to establish equivalence of unabbreviated mass spectra. *Rapid Commun. Mass Spectrom.* **2014**, *28*, 83–95. DOI:10.1002/rcm.6759.

(49) Willard, M. A. B.; McGuffin, V. L.; Smith, R. W. (2017). Statistical comparison of mass spectra for identification of amphetamine-type stimulants. *Forensic Sci. Int.* **2017**, *270*, 111–120. DOI:10.1016/j.forsciint.2016.11.013.

(50) Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H. Accelerated isotope fine structure calculation using pruned transition trees. *Anal. Chemi.* **2015**, *87*, 5738-5744. DOI:10.1021/acs.analchem.5b00941. URL: http://envipat.eawag.ch.

TOC graphic