nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-24496-9 OPEN

Check for updates

Data storage using peptide sequences

Cheuk Chi A. Ng[1,2], Wai Man Tam[3], Haidi Yin[1,2], Qian Wu[1,2], Pui-Kin So[4], Melody Yee-Man Wong[5], Francis C. M. Lau[3]✉ & Zhong-Ping Yao[1,2]✉

Humankind is generating digital data at an exponential rate. These data are typically stored using electronic, magnetic or optical devices, which require large physical spaces and cannot last for a very long time. Here we report the use of peptide sequences for data storage, which can be durable and of high storage density. With the selection of suitable constitutive amino acids, designs of address codes and error-correction schemes to protect the order and integrity of the stored data, optimization of the analytical protocol and development of a software to effectively recover peptide sequences from the tandem mass spectra, we demonstrated the feasibility of this method by successfully storing and retrieving a text file and the music file Silent Night with 40 and 511 18-mer peptides respectively. This method for the first time links data storage with the peptide synthesis industry and proteomics techniques, and is expected to stimulate the development of relevant fields.

[1] State Key Laboratory of Chemical Biology and Drug Discovery, Research Institute for Future Food and Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China. [2] State Key Laboratory of Chinese Medicine and Molecular Pharmacology (Incubation) and Shenzhen Key Laboratory of Food Biological Safety Control, The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China. [3] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China. [4] University Research Facility in Life Sciences, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China. [5] University Research Facility in Chemical and Environmental Analysis, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China. ✉email: francis-cm.lau@polyu.edu.hk; zhongping.yao@polyu.edu.hk

From the beginning of civilization, the media for storing data have been continuously evolving from such as stone tablets, animal bones, and bamboo tablets to paper, with improvements on data density over time. Since the invention of electronics in the last century, the percentage of data stored in digital form has been increasing rapidly to almost 100% recently[1]. Moreover, the amount of data generated has been increasing exponentially, from several ZB in 2008 to expected 74 ZB in 2021, causing a much increased demand for data storage correspondingly[2]. Most of the digital data are stored in physical media such as hard drives. In addition, many of the data are rarely accessed and are archived on reels of magnetic tapes. However, the physical thickness of the tapes and the size of magnetic domains limit the maximum data density, which is expected to reach a plateau soon. Furthermore, data in old tapes need to be copied onto new tapes regularly, as the magnetic tapes can normally last for 10 to 20 years only. This process is time-consuming and expensive. Hence, next-generation media that can store digital data with a much higher data density and durability are needed.

One of the emerging technologies to fulfill this need is storing digital data in molecules. A widely reported technique is data storage with deoxyribonucleic acid (DNA), where the capability of DNA data storage had advanced from several bytes decades ago[3] to hundreds-MB-scale recently[4–6]. While early examples did not achieve complete data recovery[7], the data integrity has been improving by incorporating error-correction schemes in DNA data storage, from simple repetitions[8] towards more complex and efficient schemes such as Reed–Solomon (RS) code[9] and fountain code[10]. DNA could offer much higher data density than magnetic tapes[11] and store information for thousands of years[9].

In addition to DNA, other molecules were investigated for storing digital data. For example, Roy et al. encoded data in synthetic polymers such as poly(alkoxyamine amide)s[12], and Huang et al. transformed the data into binary trees and encoded the transformed data in dendrimers[13]. Very recently, Cafferty et al. encoded data in organic molecules using a set of molecules with different masses representing the 0 and 1 in digital data and read out using matrix-assisted laser desorption/ionization mass spectrometry[14].

Here we report the use of peptide sequences for digital data storage, a method that has not been reported before[15]. Compared to DNA and other types of polymers, peptides offer several advantages for data storage. Firstly, in DNA, typically only four natural nucleotides are used as monomers due to the requirement of enzyme recognition for PCR amplification and high-throughput sequencing. In synthetic peptides, a much greater variety of monomers (amino acids) can be incorporated because enzyme recognition is not mandatory in the synthesis and sequencing of peptides. In addition to the 20 natural amino acids, many unnatural amino acids can be used. The increased set of possible monomers and lower masses than those of nucleotides could in principle allow peptides to have a higher density than DNA for data storage. Secondly, peptides can be more stable than DNA. It has been shown that after millions of years, peptides or proteins could still be detected and sequenced but DNA had already degraded[16,17]. Comparing to DNA, peptides cannot be amplified with techniques such as polymerase chain reaction (PCR). However, using tandem mass spectrometry (MS/MS)-based techniques, peptides can be detected and sequenced with good sensitivity and direct data readout without PCR-like preprocessing[18–20]. Moreover, the field of proteomics has been developing rapidly with constantly improving methods, hardware and software to allow sequencing of thousands of peptides within a very short time;[21,22] the peptide synthesis industry has been established, and the price for peptide synthesis continues to decrease. Thus comparing to other polymers or small molecules, peptides could better leverage the established methodologies and industry for design and sequencing.

Results and discussion

We have developed a method for data storage using peptide sequences, with the precise ordering of amino acids encoding the order of digital bits. As shown in Fig. 1, in our method, amino acids are assigned as sequences of digital bits (Table 1). Raw data are first encoded as long strings of 0 s and 1 s, which correspond to sequences of amino acids, i.e., peptides, according to the assignments. The peptides are synthesized and hence the data are stored. To retrieve the data, the peptides are sequenced, and the obtained sequences are converted into bits of 0 and 1, which are then decoded as the raw data. The peptides can be commercially synthesized, and MS/MS is the state-of-the-art technique for peptide sequencing. Peptides must not be too long in order to ensure effective synthesis and sequencing. Therefore, the encoded strings are broken into smaller parts, and an address indicator is added into each part to ensure all the parts will be in their original order when they are read back. In this way, raw data will be stored in a mixture of peptides, which can be separated and sequenced using liquid chromatography coupled with MS/MS (LC-MS/MS). The keys of this method are the successful synthesis, detection and sequencing of all the peptides, which have been achieved by selecting suitable amino acids to comprise the peptides, designing

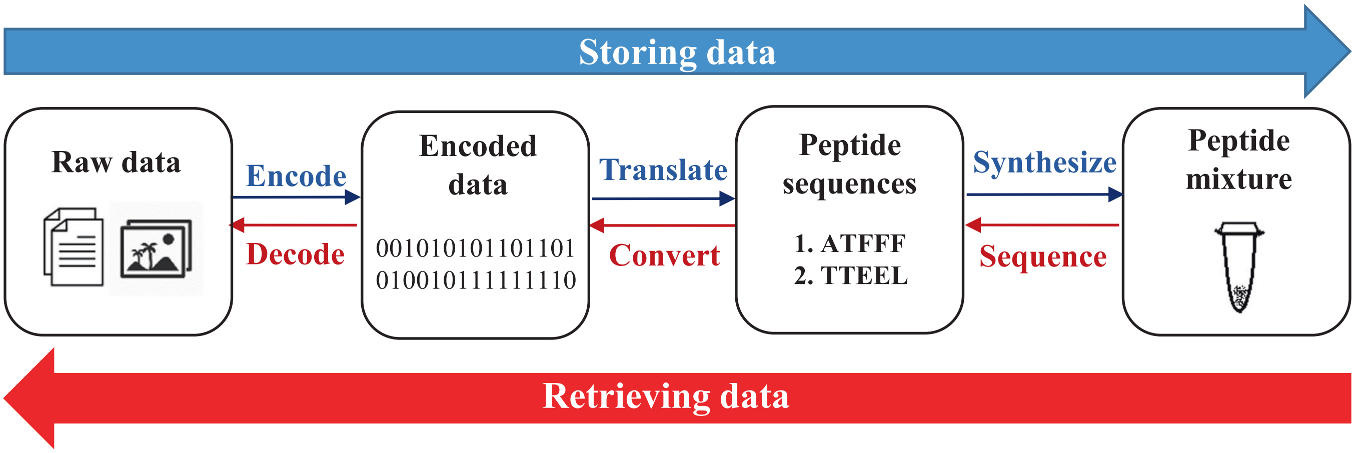
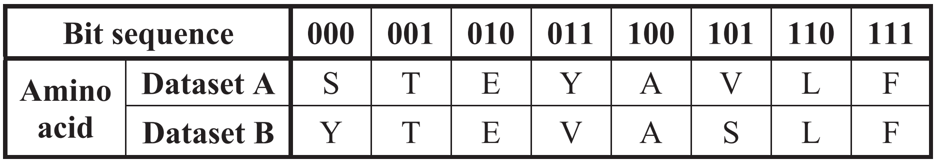


Fig. 1 Overview of the process of storing and retrieving data into and from peptides. The direction in blue represents the data storing process, while the direction in red represents the data retrieving process.

Table 1 The one-to-one mapping of bit sequences to amino acids.



Bit sequence		000	001	010	011	100	101	110	111
Amino acid	Dataset A	S	T	E	Y	A	V	L	F
	Dataset B	Y	T	E	V	A	S	L	F

suitable error-correction coding schemes, optimizing the protocol for LC-MS/MS analysis, and developing a software to effectively recover peptide sequences from the MS/MS spectra. These efforts are illustrated below, with more details available in Methods and Supplementary Information sections.

In proteomics studies, thousands of proteins could be reliably identified in one LC-MS/MS analysis, even with low sequencing coverage of the peptides, since the peptides are originated from proteins with sequences available in databases for searching[21,22]. However, such a strategy cannot be used for sequencing data-bearing peptides, which requires nearly all the amino acids of each peptide to be correctly sequenced in order to recover all encoded information. De novo peptide sequencing, a technique based on high accuracy MS/MS[23,24] and widely used for sequencing of monoclonal antibodies in industry[25], was thus used to sequence the encoded peptides. Fortunately, different from proteomic peptides with totally random and unknown sequences, the peptide sequences used for data storage can be designed beforehand according to some rules such that the sequencing accuracy is optimized.

For the peptide design, we considered several parameters with an aim to increasing the success rate of complete sequencing. The first parameter is the peptide length. Shorter peptides are easier to be synthesized and sequenced with fewer missed fragmentation, while longer peptides could store more data per peptide, reducing the number of peptides required as well as the number of addresses and error correction overhead for the same amount of data. To balance these factors, the peptide length was fixed to 18-mer long in this study. The second parameter is the choice and positioning of amino acids. Among the 20 natural amino acids, proline (P) was eliminated as peptides containing P are difficult to synthesize[26]. Histidine (H), lysine (K), and arginine (R) were not used in the middle or at the N-terminus as they caused sharp decrease in peak intensity[27,28]. Methionine (M) and cysteine (C) were eliminated because they were prone to oxidization and formation of disulfide bridges, respectively. Asparagine (N) and glutamine (Q) were eliminated as they were prone to amine loss during fragmentation in MS/MS[27]. Isoleucine (I) was eliminated as it is isobaric with leucine (L). From the 11 remaining amino acids, eight amino acids, i.e., alanine (A), valine (V), leucine (L), serine (S), threonine (T), phenylalanine (F), tyrosine (Y) and glutamic acid (E), were selected to comprise the data storage peptides with 3 bits per amino acid (Table 1). (Note that it requires 16 amino acids in order to encode 4 bits per amino acid.) C-terminal arginine has been found to promote the signal intensity of the y-ion series and suppress the b-ion series[28]. As this would simplify spectral analysis, a non-data-bearing amino acid, R, was placed at the C-terminus for each peptide. As the first and second amino acids from the N-terminus were rarely fragmented in MS/MS[29], another non-data-bearing amino acid, F, was fixed as the first amino acid from the N-terminus, so that the mass of the second amino acid could still be calculated when the first and second amino acids failed to be fragmented. Another reason for fixing F at the N-terminus was to balance the hydrophobicity of peptides, as F was hydrophobic, while R that was fixed at the C-terminus was hydrophilic. Peptides with medium hydrophobicity could facilitate peptide synthesis, as the solubility of hydrophobic peptides is low, and very hydrophilic peptides are difficult to be purified by HPLC. These choices would produce peptides that could be easily synthesized and chemically stable. They could also generate MS/MS spectra that easily allow correct peptide sequence recovery.

To further protect data integrity, error-correction schemes[30] were incorporated during encoding, such that when peptides were not synthesized, detected, or sequenced well, the missing data could still be inferred from the appended redundant data[9,10]. In this study, we designed a concatenated error-correction code, assuming that 10% of amino acids were missing or incorrect during storage and retrieval, and the orders of the second and third amino acids counted from both N-terminus (Table 2, symbols S_1 and S_2) and C-terminus (Table 2, symbols S_{15} and S_{16}) might be ambiguous because gap masses due to fragmentation were more common on these sites. The error of 10% missing amino acids was protected using an advanced low-density parity-check (LDPC)[31] or RS[32] code, while the error of ambiguous order of specific amino acids was protected by two bits, with each bit protecting the order of the second and third amino acids on each end of the peptide. From the design of the peptide structure, each amino acid represented one symbol, which contains 3 bits of information (Table 1, S1 and S2).

To retrieve the data from the stored peptides, the peptide mixtures were separated with LC and then subjected to fragmentation to produce MS/MS spectra that allowed recovery of the amino acid order based on the mass differences between the fragment ions (Fig. 2b, c). Currently available proteomics and de novo sequencing software[21–25,29] were found not to work well for the sequence recovery since they were not developed for the specific peptides used in this project. An in-house software using a highest-intensity-tag-based method (Fig. 2d), tailored to the arrangements of amino acids of the designed peptides, was thus developed to recover peptide sequences from the MS/MS spectra (Table S3) in this project. The peptide sequences were then grouped by scoring and finally decoded to recover the original data.

As proofs of concept, two datasets were stored and retrieved in this study. Dataset A was an 848 bits long BIG5-formatted text for “The Hong Kong Polytechnic University, 80th anniversary.” in both Chinese and English and the motto of The Hong Kong Polytechnic University in Chinese (Fig. 2a), while dataset B was 13,752 bits long, containing the music Silent Night in MIDI (Supplementary Audio 1) format and its title in ASCII format. Dataset A was encoded (Table 1 and S1) and translated into 40 18-mer peptides (Table S4), which were synthesized for data storage. For data retrieval, the peptide mixture was analyzed using LC-MS/MS (Fig. 2b), and the acquired MS/MS spectra (Fig. 2c) were processed with the in-house software for recovery of the

Table 2 Structure of sequences with 16 3-bit symbols, where each 3-bit symbol can be translated to one amino acid according to Table 1.

Symbol	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}
Sequence for dataset A	Add	Add	$Q_{1,2}$	0	0	0	$Q_{15,16}$	c	c	c	c	c	c	c	c	c
			c	c	c	c	c	c	c	c	c	c	c	c	c	c
			c	c	c	c	c	c	c	c	c	c	c	c	c	c
Sequence for dataset B	Add	Add	Add	$Q_{1,2}$	c	c	c	c	c	c	c	c	c	c	c	c
				$Q_{2,3}$	c	c	c	c	c	c	c	c	c	c	c	c
				$Q_{15,16}$	c	c	c	c	c	c	c	c	c	c	c	c

The first 2 or 3 symbols are used to assign the address ("Add", light red). The bit $Q_{i,j}$ is used to record the order of S_i and S_j (white). The other symbols are used to store (i) the coded bits c (green) including the information and the parity bits; and (ii) some zero bits (blue) to ensure that at least three symbols are hydrophilic amino acids.

peptide sequences, which were converted back to sequences of bits according to the previous assignments (Table 1) and then decoded back to the original raw data. The results showed that the sequences of all 40 peptides were correctly obtained, allowing complete retrieval of the original data. Similar procedures (Fig. 3 and Table 1 and S2) were employed for storage and retrieval of dataset B, which required 511 18-mer peptides for data storage. The results showed that 93.7% (7659/8176) of the amino acids were correctly recovered (Table S5). After the error-correction decoding procedure that could recover a maximum of 10% of incorrect or lost amino acids, the original music and title were fully retrieved.

In DNA data storage that used four nucleotides as monomers, each nucleotide represented 2 bits, while the use of eight amino acids as monomers in our method enabled each amino acid to represent 3 bits. Together with the lower masses of amino acids, in principle, the storage density of our method could be 3.72 times of the DNA method, i.e., storage of the same data using a lower amount (lighter) of peptides than DNA. The storage density of our method can still be further improved with use of 16 or more amino acids. Practically, the retrievable data density was 1.7×10^{10} bits/g and 2.6×10^{9} bits/g for datasets A and B, respectively, which were about nine orders of magnitudes lower than those of the DNA method[11]. The major reason for this is that DNA can be amplified by PCR prior to sequencing while peptides cannot, therefore the number of molecules required to retrieve data for the DNA method can be far fewer than the peptide method. The peptide-based data density can be significantly improved with optimized peptide sequencing, since picomole amounts of peptides were used for analysis in these proof-of-concept studies while peptide detection and sequencing at attomole[18–20], yoctomole[33], or even single molecule[34–36] scales have been reported.

In summary, we demonstrated that it was feasible to store data using peptide sequences and to retrieve the data using LC-MS/MS analysis. This method offers a new possibility for data storage with potentially high storage density and durability. Peptide synthesis industry and proteomics techniques have been developed to the stages that can allow the use of peptides for data storage. Our method for the first time connects these fields together and can promote the development of these and other relevant fields. Currently, peptide synthesis and sequencing are still relatively expensive and time-consuming in practice, and scaling-up significantly would require further developments in these fields. As the stored data become much larger, much more peptides would be required to encode the data, leading to much more complicated peptide sequencing that would challenge the analytical capabilities of current LC-MS/MS techniques, and new analytical techniques and strategies would be needed to solve the problems. However, with the improved techniques and reduced time and costs of the peptide synthesis and sequencing, which have been happening in the past decades, peptide data storage may become practically available in the future, especially in critical applications that demands minimum weight and long duration for stable storage of very big data.

Methods

Materials. Peptides (lyophilized, as trifluoroacetate salts, >50% purity) were synthesized by Genscript Inc. (Nanjing, China) and GL Biochem (Shanghai, China). The peptides were dissolved in dimethyl sulfoxide (10 µg/mL), mixed together for each dataset, and diluted with 50% acetonitrile with a 1:1 ratio before analysis. Methanol and acetonitrile (HPLC grade) were from Duksan (South Korea). Formic acid (99–100%) was from VWR (France). Water was purified by MilliQ system.

LC-MS/MS analysis of the peptide mixtures. The step-by-step protocol used in this work is available on Protocol Exchange[37]. The peptide mixtures were separated using a Waters Acquity UPLC system with a C18 column (Agilent AdvanceBio Peptide Map, 2.1 × 150 mm, 2.7 µm particle size, 120 Å pore size). Mobile phase A was 0.2% formic acid in water and B was 0.2% formic acid in acetonitrile. The flow rate was 0.3 mL/min and the temperature was 55 °C. The gradient changed from 10% B to 18% B at 0 to 2 min, from 18% B to 22% B at 2 to 8 min, from 22% B to 34% B at 8 to 48 min, from 34% B to 40% B at 48 to 64 min, from 40% B to 55% B at 64 to 75 min, from 55% B to 80% B at 75 to 78 min, and remained at 80% B from 78 to 83 min.

MS/MS analysis was performed using an Orbitrap Fusion Lumos mass spectrometer (ThermoFisher Scientific, San Jose, CA) operated in positive ion mode. The spray voltage for electrospray ionization was +3600 V, and both ion transfer tube temperature and vaporizer temperature were 280 °C. In each cycle, a MS1 scan with m/z from 900 to 1400 Da was performed with a resolution of 30 K. Ions were selected for MS/MS with quadruple, using advanced peak determination (APD) with default charge of +2, top-speed mode with 3 s cycles, mass tolerance of 25 ppm, dynamic exclusion window of 4 s, and isolation window width of 1.6 or 0.7 Da. High-energy collision dissociation (HCD) at 28% of normalized collision energy with stepped collision energy of 5% was used for the fragmentation. MS/MS spectra were obtained with m/z from 240 to 2450 Da and a resolution of 15 K.

Error-correction code design. The structures of the sequences used for encoding, sequencing, and decoding are shown in Table 2 for datasets A and B.

For dataset A, the first two symbols S_1 and S_2 are used to assign the address (orange). The order-checking bits $Q_{1,2}$ and $Q_{15,16}$ are used to record the order of S_1 and S_2 and the order of S_{15} and S_{16}, respectively (white). Note that 3 zero bits (blue) are filled in the first bits of S_4–$S_{6,}$ which can ensure that at least three symbols are hydrophilic amino acids. The other symbols are used to store the coded bits c including the information and the parity bits (green) of the error-correction codes. For dataset B, due to more information bits and longer address, three address symbols and three order-checking bits are used. Such structure can be easily modified for longer sequence with longer address.

As shown in Table S4, a 40 × 16 block is constructed for error correction and peptide sequencing of dataset A. Each row had 16 symbols (i.e., S_1, S_2, …, S_{16}) to

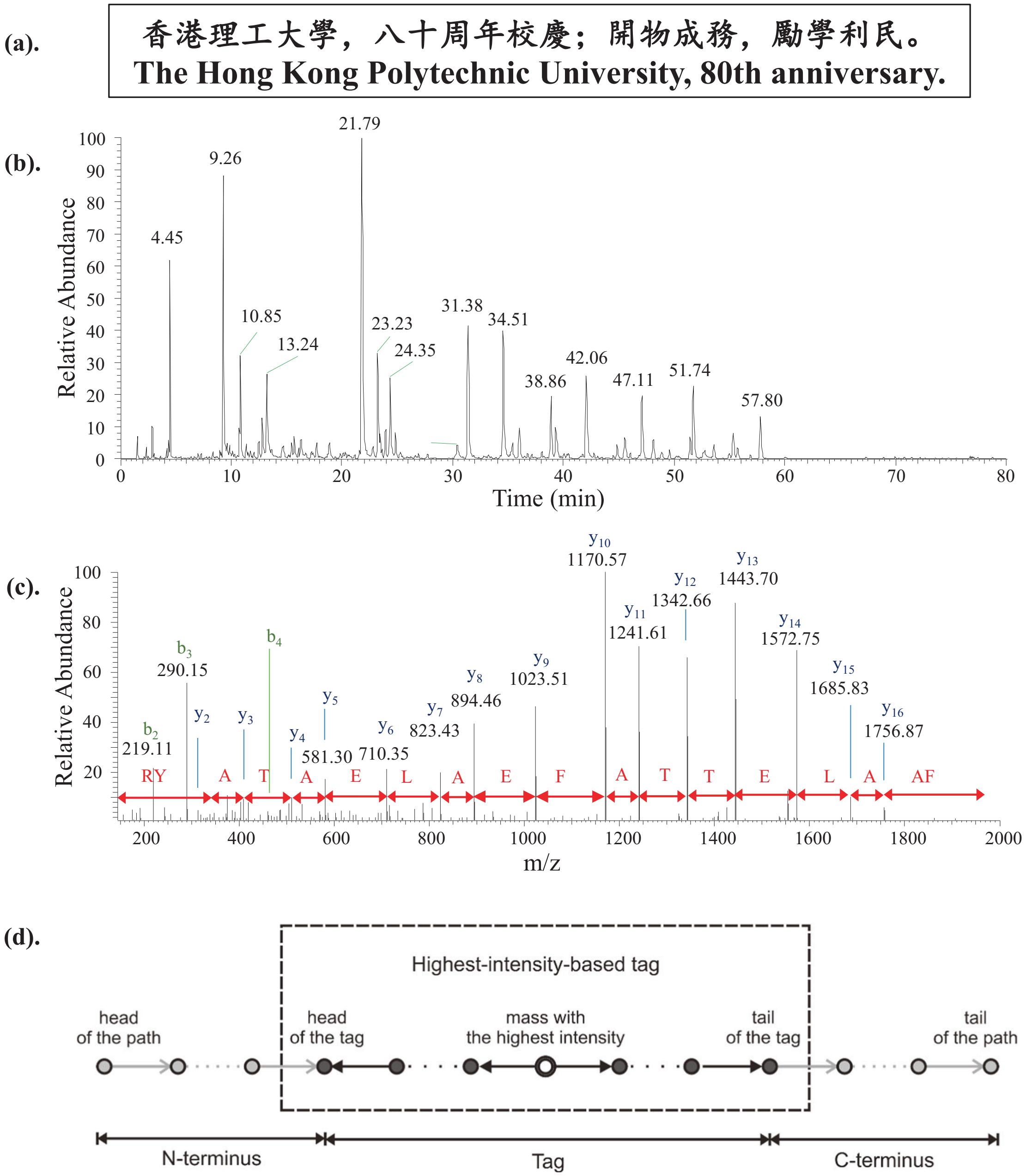


Fig. 2 Overview of data retrieval from dataset A. a The message of dataset A; **b** The chromatogram for analysis of the 40 peptides for dataset A; **c** A typical MS/MS spectrum for analysis of peptides in dataset A, and the sequence of one of the data-bearing peptide read out from the spectrum; **d** The highest-intensity-tag-based sequencing method used in the sequence recovery.

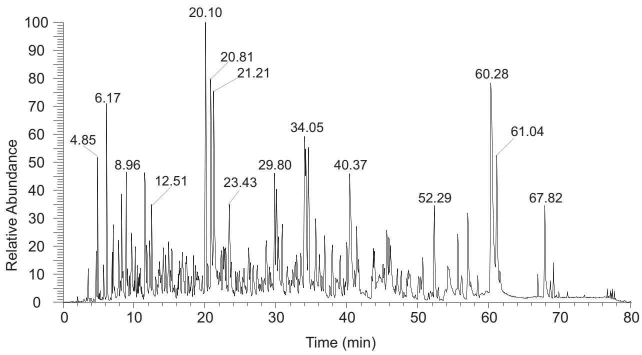


Fig. 3 The chromatogram for analysis of the peptide mixture encoding dataset B.

represent a 16-mer data-bearing peptide sequence. In the design of encoding scheme for dataset A, the first two symbols in each sequence were used to store the address, and the remaining 14 symbols were used to store information. Hence a total of 560 symbols are available in the 40 peptide sequences (total 560 × 3 = 1680 bits, see Table S2 for the 40 peptide sequences). Then 850 information bits (i.e., b_1, b_2, …, b_{850}) were filled in the data block according to the following arrangements (Table S4):

Bits b_1–b_{400} were filled in the second and the third bits of Symbols S_3–S_7 of the peptide sequences Seq #1 to Seq #40;
Bits b_{401}–b_{760} were filled in Symbols S_{14}–S_{16} of the peptide sequences Seq #1 to Seq #40;
Bits b_{761}–b_{850} were filled in Symbol S_{13} of the peptide sequences Seq #11 to Seq #40.

Furthermore, the first bits of Symbols S_4–S_6 (represented by b_{851}–b_{970}) were filled with “0” bits. The purpose is to ensure that a minimum of three symbols in each sequence having values 0, 1, 2, 3, which will be represented by hydrophilic amino acids S, T, E, Y, as mentioned in the design. There were also order-checking bits Q and redundant bits $P_i^{(j)}$ ($i = 1, 2, …, n$; $j = 1, 2,$ and 3) derived by LDPC codes in the peptide sequences (where n is the number of parity bits in each LDPC code). The order-checking bit $Q_{i,j}$ is "1" if symbol S_i is larger than S_j. Otherwise, the order-checking bit is "0". Thus, the maximum overall code rate R of the block was $850/(14 \times 3 \times 40) = 0.506$. As dataset A only consisted of 848 bits, 2 zeros were appended to the end of the data in order to fully fill the block.

Based on the results of dataset A, we further made these assumptions of possible errors when designing the error-correction scheme for dataset B: (i) 10% of the three-symbol sequences $\{S_5S_6S_7\}$ and $\{S_8S_9S_{10}\}$ cannot be recovered correctly; and (ii) 15% of the three-symbol sequences $\{S_{11}S_{12}S_{13}\}$ and $\{S_{14}S_{15}S_{16}\}$ cannot be recovered correctly. Based on these assumptions, we proposed another error-correction method based on the RS code[32] that used: (i) three order-checking bits for each peptide sequence; and four RS codes to recover the original data even when any arbitrary 10% three-symbol sequences $\{S_5S_6S_7\}$, any arbitrary 10% three-symbol sequences $\{S_8S_9S_{10}\}$, any arbitrary 15% three-symbol sequences $\{S_{11}S_{12}S_{13}\}$, and any arbitrary 15% three-symbol sequences $\{S_{14}S_{15}S_{16}\}$, cannot be recovered correctly.

When this scheme was used on dataset B, a 511×16 block of symbols was constructed (Table S5), which comprises $511 \times 16 \times 3 = 24528$ bits. The three-symbol sets $A_{i,1}$ $A_{i,2}$ $A_{i,3}$ ($i = 1, 2, \ldots, 511$) were used for addressing, with Symbols S_1 to S_3 having the values of 000, 001, 002, …, 775, 776. The three bits of Symbol S_4 were the three order-checking bits used to protect the order of Symbols S_1 and S_2, the order of Symbols S_2 and S_3, and the order of Symbols S_{15} and S_{16}, respectively. Then there were $511 \times 12 \times 3 = 18396$ bit positions in Symbols S_5 to S_{16} of the block to store the information and parity bits for the RS codes. Due to the different protection requirements for the partial sequences $\{S_5S_6S_7S_8S_9S_{10}\}$ and $\{S_{11}S_{12}S_{13}S_{14}S_{15}S_{16}\}$, two (511, 409) RS codes were used for partial sequences $\{S_5S_6S_7\}$ (RS1) and $\{S_8S_9S_{10}\}$ (RS2), another two (511, 357) RS codes were used for partial sequences $\{S_{11}S_{12}S_{13}\}$ (RS3) and $\{S_{14}S_{15}S_{16}\}$ (RS4). Each symbol in RS code comprised 9 bits. The (511, 409) RS and (511, 357) RS codes could correct up to 51 and 77 9-bit symbol errors, respectively (Table S5). Moreover, the numbers of total information bits and total parity bits of all four RS codes were given by $(409 + 357) \times 2 \times 9 = 13788$ and $(102 + 154) \times 2 \times 9 = 4608$, respectively. The maximum overall code rate R of the block is given by $13788/(511 \times 16 \times 3) = 0.562$. As dataset B only had 13752 bits, zeros were appended to the end such that the block could be fully filled. This code rate is comparable to that of DNA (varied from 0.17 for repetition encoding[8] to 0.785 for fountain encoding[10], assuming a maximum capacity of 2 bits per nucleotide). Improvement of code rate is possible if longer peptides could be used, if the peptide design could be improved to reduce error, and if the coding scheme could be more focused on the error-prone amino acid positions to reduce unnecessary redundancy.

Sequence recovery. An in-house software was developed for recovery of peptide sequences from the MS/MS spectra. To reduce the amount of false positives, during spectral analysis, the maximum error for each peak was set to 25 ppm (in line with the experimental parameters), and the masses were corrected to at least five decimal places.

The spectra were first extracted from the Thermos RAW file using MSConvert[38]. Then, the spectra were passed onto a preprocessing unit, which included deconvolution to obtain a list of masses and charges of isotopic clusters with one or more peaks each, and identifying the monoisotopic mass and charge of parent ion. As only 2+ precursor ions were selected for MS/MS during experiments, most fragments would be predominately 1+ unless their masses were close to precursor mass. Also, it was predicted that in most peptides, the strongest peak in the isotopic cluster would be M + 1 rather than M if the formula mass were above ca. 1800[39], where M was the monoisotopic mass. Therefore, it was assumed that if a certain isotopic cluster only consisted of one peak, that peak was singly charged and the monoisotopic mass would be corrected based on the deconvoluted mass accordingly. MS/MS spectra with less than 12 peaks with the most intense peak lower than 30,000 counts were filtered out at this stage to speed up analysis.

After that, de novo sequencing based on the graph model is utilized for determining the peptide sequences from the preprocessed spectra[23]. In the graph model, the MS/MS spectrum is represented by a directed acyclic graph (DAG). The peaks of the spectrum can be taken as vertices, while an edge is added between two vertices when the mass gap between two peaks is equal to the mass of an amino acid. The objective is to find the longest path in the graph starting from the head vertex to the tail vertex. The sequence identification was started in the middle part of the MS/MS spectrum. The sequence tagging method first infers a partial sequence called tag, and then finds the whole sequence that can match the tag. The tags containing the amino acid with the highest intensity were first obtained, which were called highest-intensity-based tags (Fig. 4). To generate valid sequence candidates, both ends of the tag would be extended and connected to the N- and C-termini by searching the sequences containing amino acids with matching gap masses. The scores for the sequence candidates based on the following five factors: the length of consecutive amino acids retrieved, the number of amino acids retrieved, match error, intensity, and the number of occurrences for different ion types with different offsets. The higher the score is, the more likely that the sequence is correct.

Details of the highest-intensity-tag based sequencing method. In the highest-intensity-tag based sequencing method (Fig. 4), the sequences were estimated in a manner that first inferred a partial sequence with a small amount of reliable information and then found the missing part of the sequence with less reliable data or the raw data. The *m/z* value with the first highest intensity was first recognized to further infer the tag or the path. Although using short tags (e.g., with three amino acids) such as GutenTag[40], DirecTag[41], and NovoHCD[42] could avoid introducing the wrong amino acids, the number of candidate tags would be relatively larger and harder to infer the sequences due to insufficient information provided by the tag. Therefore, in this project, the length of tag was variable and could be up to the length of the peptide, which helped to reduce the search space. When a tag contained wrong amino acids, it could not be extended well towards N-terminus and C-terminus. In this case, the length of the tag was shortened by adaptively reducing the number of the higher-intensity data points used for the tag-finding algorithm. In addition, the vertex with the highest intensity may not definitely present in the correct path due to the uncertainty of the data. When valid paths could not be found, it may be possible to infer the tag with the second or even the third highest intensities. Moreover, in order to find the tag, N-terminal and C-terminal amino acids, a sequencing method called two-stage sequencing method is used together with the highest-intensity-tag based method.

Figure 4 shows a flowchart illustrating the method of highest-intensity-tag based sequencing. At steps 2, 3 and 4, the intensities of the preprocessed data were sorted from the largest to the smallest and values with J denoting the ranking of intensity. The mass/charge ratio with the highest intensity was then identified. At start, it was set as $J = 1$ and $i = 1$, and using only $W = w_i$ ($w_1 > w_2 > w_3 \ldots$) masses with the higher ranking in the tag-finding processing.

The method then proceeded to step 5 to find the highest-intensity-based tag. Starting from the mass of the putative y-ion with the highest intensity, the highest-intensity-based tag was found by simultaneously connecting the vertices in the forward direction pointing to the tail vertex of the path, and connecting the vertices in the backward direction pointing to the head vertex of the path, where the vertices had mass gap being the exact mass of any amino acid and preferably the length of the tag was as long as possible (Fig. 2d). The tags containing the amino acid with the highest intensity were obtained subsequently, which were called highest-intensity-based tags. With knowledge of the masses of the head and the tail amino acids of a highest-intensity-based tag, the method proceeded to step 6 to find the N-terminal amino acids that could connect the head of the path to the head of the tags in the forward direction by using the method described at Step 2 of the two-stage sequencing method. Similarly, for the tags with valid N-terminal amino acids, at step 7, the C-terminal amino acids of the sequences could be further found by connecting the tail of the tags to the tail of the path in the forward direction.

At step 8, the candidate paths could be constructed by combining the three parts: N-terminus, tag, and C-terminus. At step 9, one could follow steps 3 and 4 of the two-stage sequencing method to select and refine the sequences. Note that a larger value for W sometimes introduced one or more wrong amino acids in the head and/or tail parts of a tag, while a smaller value for W may give more reliable tag but the length of the tag may be limited. Therefore, after step 6, if no valid candidate could be found, one may attempt to reduce the value of W with $W = w_i$ by increasing i by 1, i.e., $i = i + 1$, and repeat the tag, N-terminus and C-terminus finding procedure until the candidate sequence could be found or $i = V$ (where V is the maximum number of iterations).

For the special case when the experimental mass with the highest intensity gave an unreliable message due to noise and uncertainty, a highest-intensity-based tag or a valid path with the highest-intensity-based tag could not be found. In this case, the mass with the second highest intensity was used by setting $J = J + 1$ and $i = 1$ to find the second highest-intensity-based tag and the candidates. This process would continue until the sequence could be found or $J = J_{\max}$ (where $J_{\max}$ is the maximum number of higher-ranking-intensity masses allowed to be the start point to find the tag).

Details of the two-stage sequencing method. Figure 5 shows a flowchart illustrating a method of two-stage sequencing. Four steps are involved in the two-stage sequencing method: (1) preprocessing, (2) candidate sequence generation, (3) sequence selection, and (4) candidate refining. As shown in Fig. 5, Steps 1–3 belong to the first stage (Stage 1), while Step 4 is processed in the second stage (Stage 2). In Stage 1 of the two-stage sequencing method, partial sequence is inferred using the preprocessed data after Step 1. In Stage 2, the remaining part of the sequence is determined using the raw data.

At Step 1, preprocessing is performed. At Step 2, the preprocessed data from step 1 is used to find the valid paths (sequences), and the number n of candidate sequences is counted. At Step 3, the effects of the following five factors are jointly considered when arriving at the score of a sequence candidate from Step 3.1 to Step 3.5: length of consecutive amino acids retrieved, number of amino acids retrieved, match error, average intensity of amino acids retrieved, and number of occurrences for different ion types with different offsets. The sequences with the longest length of consecutive amino acids retrieved are first selected (Step 3.1). Among the selected sequences, the sequences with the largest number of amino acids retrieved are then selected (Step 3.2). For the sequences with equal length of consecutive amino acids retrieved together with equal number of amino acids retrieved, the match error is evaluated, which is the mean error between the observed mass values for the amino acids retrieved from the experimental spectrum and the actual mass values of the amino acids normalized by the corresponding observed mass values (Step 3.3). If there is more than one sequence with identical match errors, the

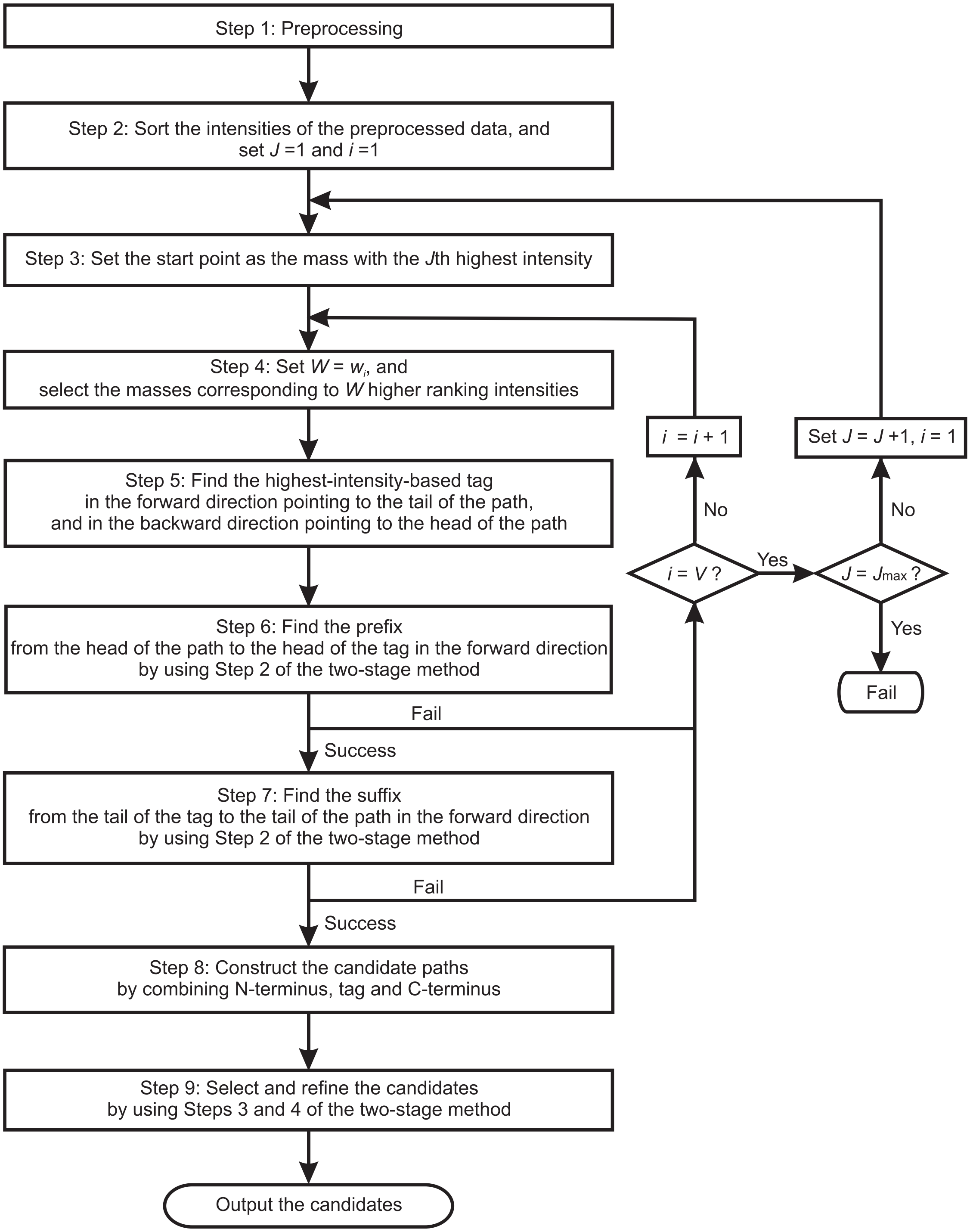

Step 1: Preprocessing
Step 2: Sort the intensities of the preprocessed data, and set J =1 and i =1
Step 3: Set the start point as the mass with the Jth highest intensity
Step 4: Set W = w_i, and select the masses corresponding to W higher ranking intensities
Step 5: Find the highest-intensity-based tag in the forward direction pointing to the tail of the path, and in the backward direction pointing to the head of the path
Step 6: Find the prefix from the head of the path to the head of the tag in the forward direction by using Step 2 of the two-stage method
Fail
Success
Step 7: Find the suffix from the tail of the tag to the tail of the path in the forward direction by using Step 2 of the two-stage method
Fail
Success
Step 8: Construct the candidate paths by combining N-terminus, tag and C-terminus
Step 9: Select and refine the candidates by using Steps 3 and 4 of the two-stage method
Output the candidates
i = i + 1
No
i = V ?
Yes
Set J = J +1, i = 1
No
J = Jmax ?
Yes
Fail

Fig. 4 A flowchart illustrating the method of highest-intensity-tag based sequencing. i represents the iteration number and V is the maximum number of iterations. W represents the number of masses with the higher ranking used in the tag-finding processing and w_i is the number of masses with the higher ranking for the ith iteration. J represents the ranking of intensity and J_{max} is the maximum number of higher-ranking-intensity masses allowed to be the start point to find the tag.

average intensity of amino acids retrieved is further calculated and a higher score is given to a sequence with a larger average intensity value (Step 3.4). In addition, multiple ion types are usually considered as the important factors in inferring an amino acid, which means that a mass value may correspond to different types of ions in the spectrum. Generally, the more the number of occurrences for different ion types of an amino acid is, the more likely the amino acid is correct. Therefore, for the sequences with equal score after the aforementioned evaluations of Steps 3.1–3.4, the number of occurrences for different ion types is counted to determine the sequence (Step 3.5). The mass offset sets for the N-terminal a-ion, b-ion, and c-ion type sets, i.e., {a, a-H_2O, a-NH_3, a-NH_3-H_2O}, {b, b-H_2O, b-H_2O-H_2O, b-NH_3, b-NH_3-H_2O}, and {c, c-H_2O, c-H_2O-H_2O, c-NH_3, c-NH_3-H_2O} are {−27, −45, −44, −62}, {+1, −17, −35, −16, −34}, and {+18, 0, −18, +1, −17}, respectively. According to the fragmentation method and the property of the data, all or some of the above ion types can be used flexibly.

Since the candidate sequences obtained at Step 2 are found by using the preprocessed data, which aim to provide more reliable information to generate the partial sequence, amino acid combinations (AACs) may present in the sequence due to insufficient data provided by preprocessing. At Step 4, if selected sequences with missing mass values exist, which means that the corresponding mass gaps are equal to the summation of at least two amino acids, the raw data may be used to find as many vertices as possible for the path in Stage 2. After finding the missing amino acids of AACs at Step 4.1, the sequences with the longest length of

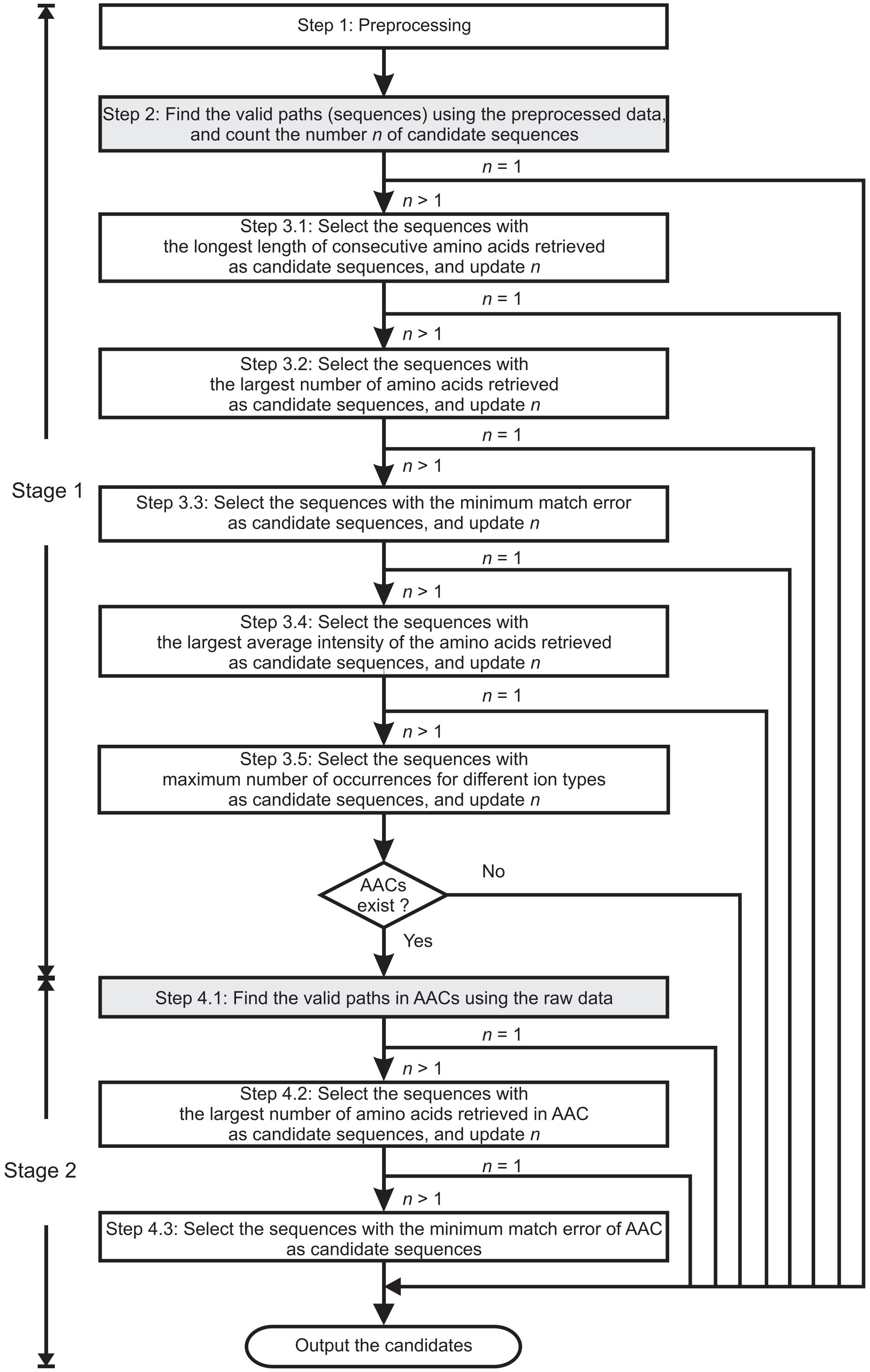

Stage 1
Step 1: Preprocessing
Step 2: Find the valid paths (sequences) using the preprocessed data, and count the number n of candidate sequences
n = 1
n > 1
Step 3.1: Select the sequences with the longest length of consecutive amino acids retrieved as candidate sequences, and update n
Step 3.2: Select the sequences with the largest number of amino acids retrieved as candidate sequences, and update n
Step 3.3: Select the sequences with the minimum match error as candidate sequences, and update n
Step 3.4: Select the sequences with the largest average intensity of the amino acids retrieved as candidate sequences, and update n
Step 3.5: Select the sequences with maximum number of occurrences for different ion types as candidate sequences, and update n
AACs exist ?
No
Yes
Stage 2
Step 4.1: Find the valid paths in AACs using the raw data
Step 4.2: Select the sequences with the largest number of amino acids retrieved in AAC as candidate sequences, and update n
Step 4.3: Select the sequences with the minimum match error of AAC as candidate sequences
Output the candidates

Fig. 5 A flowchart illustrating the method of two-stage sequencing. AAC stands for amino acid combinations.

consecutive amino acids retrieved in AACs are selected as candidate sequences (Step 4.2). If there still remain at least two candidate sequences after selection, a final decision is made based on the match error of the amino acids retrieved in AACs for each sequence (Step 4.3).

Grouping. After sequence recovery, there could be more than one valid 18-mer sequences from each spectrum or multiple valid 18-mer sequences containing the same address. Therefore, sequencing selection and grouping were performed to identify the correct peptide for each address.

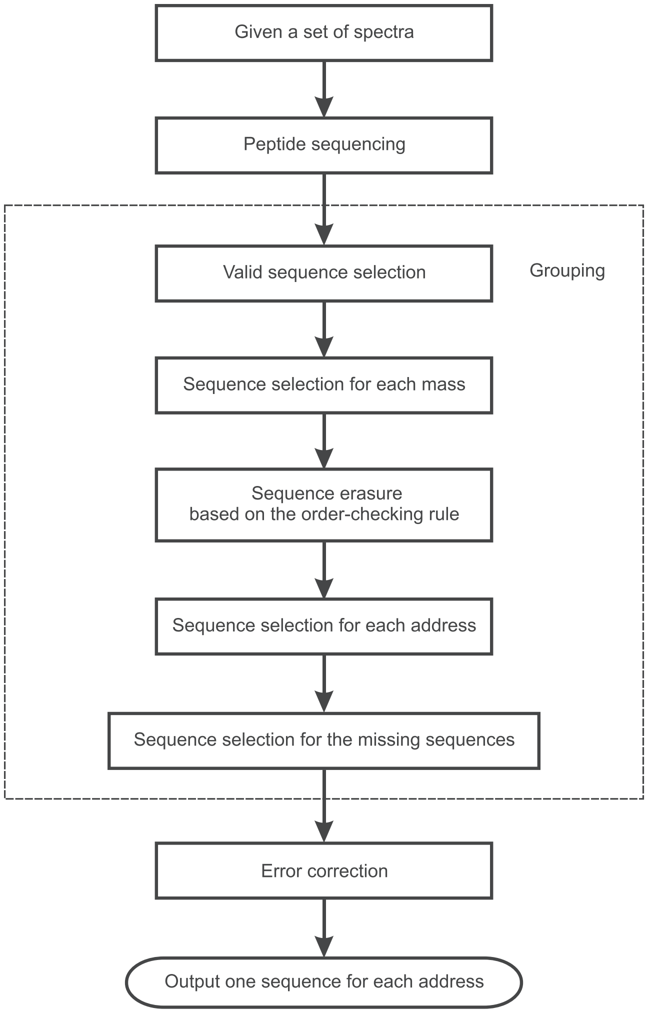

Given a set of spectra
Peptide sequencing
Valid sequence selection
Grouping
Sequence selection for each mass
Sequence erasure based on the order-checking rule
Sequence selection for each address
Sequence selection for the missing sequences
Error correction
Output one sequence for each address

Fig. 6 A flowchart illustrating the method of sequence grouping. The procedure of grouping is shown in the dashed square.

Given a set of spectra containing Ns (Ns = 40 or 511) peptide sequences. After sequencing, a set of sequence is obtained and a block of Ns × 16 is constructed for decoding. This process is called sequence grouping which is described below and in Fig. 6.

Valid sequence selection. To reduce the effect of the unreliable sequences caused by the noise and the uncertainty, the requirements for the valid sequence are listed as follows.

1. The sequence is of length-16.
2. For each sequence, one AAC with more than two missing amino acids is not allowed.
3. For each sequence, more than one AACs with two missing amino acids are not allowed.

After MS/MS analysis, a set of spectra containing 40 or 511 sequences is obtained, among which some spectra can generate length-16 sequences. If there are more than two missing amino acids in one AAC or two missing amino acids in more than one AACs, then the corresponding sequence will be ignored in the further selection.

Selection for each mass. For each mass value, if there are more than one output sequences with the highest score, then all these sequences are selected; otherwise, at most $L_{\max}$ (=2) sequences with higher scores will be considered for each of the selected spectra.

Erasure based on order checking. Based on the orders of the estimated symbol pairs $\{S_1, S_2\}$ and $\{S_{15}, S_{16}\}$ for 40 sequence set, 2 bits are generated according to the order-checking rule, which will be compared with the first bits of the estimated symbols S_3 and S_7, respectively. Similarly, 3 bits are generated based on the orders of the estimated symbol pairs $\{S_1, S_2\}$, $\{S_2, S_3\}$, and $\{S_{15}, S_{16}\}$ for 511 sequence set, which will be compared with the 3 bits of the estimated symbol S_4. If any one of the generated order-checking bits does not match the corresponding bit in an estimated sequence, the estimated sequence will be erased.

Selection for each address. According to the address represented by the first two and three elements of a sequence, the sequences are divided into 40 and 511 address groups, respectively. Then for each group, there are following cases possible:

Case 1: There is only one sequence.
Case 2: There are two or more sequences, some of which are the same, where:
 2a. there is only one result with two or more sequences; or
 2b. there are at least two different results, each with two or more sequences.
Case 3: All sequences in the group are different, where:
 3a. different sequences belong to the same spectrum; or
 3b. different sequences belong to different spectra.

For Case 1, the only sequence is recovered for the group. For Case 2a, the result with two or more sequences is selected. For Case 2b, the results with the largest number of sequences are first selected. Among the sequences corresponding to these results, the sequence with the highest score according to Steps 3.1–3.5, 4.2, and 4.3 of the two-stage sequencing method is further selected. For Case 3a, the sequence with the highest score according to Steps 4.2 and 4.3 of the two-stage sequencing method is selected. For Case 3b, the sequence with the highest score according to Steps 3.1–3.5, 4.2, and 4.3 of the two-stage sequencing method is selected.

Selection for missing sequences in the block. With knowledge of the sequences corresponding to each address, a N_s x 16 block of symbols can be constructed with each row of the block representing a sequence and each symbol representing an amino acid. In this block, some rows of the block may be missing due to the erasure by the order-checking process or the impurity of the data for peptide sequencing.

If there existed missing rows for some addresses in the block, the length-16 sequences generated by all spectra were considered to find these missing rows. For each address with missing row, the scores of the sequences were compared to make the decision.

Calculation of data density. In theory, each nucleotide in DNA could hold 2 bits while each amino acid in our designed peptides could hold 3 bits. The average molecular mass of nucleotides in DNA is 327 Da, while the average molecular mass of the eight amino acids used in this project is 132 Da. Putting these factors together, in principle, the storage density ratio of our method to the DNA method is (3/132)/(2/327) = 3.72.

About the density allowing flawless retrieval in this work, for dataset A, the total concentration of all 40 peptides was 10 ng/μL (0.25 ng/μL for each peptide) in the final mixture. Based on the injection volume of 5.0 μL, the total mass of peptides used was 10 × 5.0 = 50 ng. Therefore, the data density of peptides in this study was $848/(50 \times 10^{-9}) = 1.7 \times 10^{10}$ bits/g. For dataset B, the total concentration of all 511 peptides was 1.02 μg/μL (2 ng/μL for each peptide) in the final mixture. Based on the injection volume of 5.0 μL, the total mass of peptides used was 5.1 μg. Therefore, the data density of peptides in this study was $13752/(5.1 \times 10^{-6}) = 2.6 \times 10^9$ bits/g.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw spectral data generated in this study have been deposited at MassIVE and can be available at https://doi.org/10.25345/C5P54Z.

Code availability

The custom scripts for encoding data, peptide sequencing, and decoding data used in this paper are not publicly available due to the patent issues, but may be available for academic exchange and collaboration purposes by sending email requests to the corresponding authors, with the expected response time of around 1 week.

Received: 17 May 2021; Accepted: 22 June 2021;
Published online: 13 July 2021

References

1. Hilbert, M. & López, P. The World's technological capacity to store, communicate, and compute information. *Science* **332**, 60 (2011).
2. Hoist, A. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. https://www.statista.com/statistics/871513/worldwide-data-created/ (accessed 28 May 2021).
3. Clelland, C. T., Risca, V. & Bancroft, C. Hiding messages in DNA microdots. *Nature* **399**, 533 (1999).

4. Bornholt, J. et al. A DNA-based archival storage system. *SIGPLAN Not.* **51**, 637–649 (2016).
5. Regalado, A. Microsoft has a plan to add DNA data storage to its cloud. *MIT Technol. Rev.* (2017).
6. Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242 (2018).
7. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337**, 1628–1628 (2012).
8. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77 (2013).
9. Grass, R. N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem. Int. Ed. Engl.* **54**, 2552–2555 (2015).
10. Yaniv, E. & Dina, Z. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
11. Organick, L. et al. Probing the physical limits of reliable DNA data retrieval. *Nat. Commun.* **11**, 616 (2020).
12. Roy, R. K. et al. Design and synthesis of digitally encoded polymers that can be decoded and erased. *Nat. Commun.* **6**, 7237 (2015).
13. Huang, Z. et al. Binary tree-inspired digital dendrimer. *Nat. Commun.* **10**, 1918 (2019).
14. Cafferty, B. J. et al. Storage of information using small organic molecules. *ACS Cent. Sci.* **5**, 911–916 (2019).
15. Yao, Z. P., Ng, C. C. A., Lau, C. M. & Tam, W. M. Data storage using peptides. US Provisional Patent Application No. 62/657,026 (Filed on 13 April 2018); PCT Application No. PCT/CN2018/119349 (Filed on 6 December 2018); US Non-Provional Patent Application No.16/224,957 (Filed on 19 December 2018).
16. Service, R. F. Protein power. *Science* **349**, 372–373 (2015).
17. Warren, M. Move over, DNA: ancient proteins are starting to reveal humanity's history. *Nature* **570**, 433–436 (2019).
18. Nguyen, T. T. T. N., Petersen, N. J. & Rand, K. D. A simple sheathless CE-MS interface with a sub-micrometer electrical contact fracture for sensitive analysis of peptide and protein samples. *Anal. Chim. Acta* **936**, 157–167 (2016).
19. Sun, B., Kovatch, J. R., Badiong, A. & Merbouh, N. Optimization and modeling of quadrupole orbitrap parameters for sensitive analysis toward single-cell proteomics. *J. Proteome Res.* **16**, 3711–3721 (2017).
20. Valaskovic, G. A., Kelleher, N. L., Little, D. P., Aaserud, D. J. & McLafferty, F. W. Attomole-sensitivity electrospray source for large-molecule mass spectrometry. *Anal. Chem.* **67**, 3802–3805 (1995).
21. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
22. Yates, J. R. The revolution and evolution of shotgun proteomics for large-scale proteome analysis. *J. Am. Chem. Soc.* **135**, 1629–1640 (2013).
23. Frank, A. M., Savitski, M. M., Nielsen, M. L., Zubarev, R. A. & Pevzner, P. A. De novo peptide sequencing and identification with precision mass spectrometry. *J. Proteome Res.* **6**, 114–123 (2007).
24. Ma, B. et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342 (2003).
25. Bandeira, N., Pham, V., Pevzner, P., Arnott, D. & Lill, J. R. Automated de novo protein sequencing of monoclonal antibodies. *Nat. Biotechnol.* **26**, 1336–1338 (2008).
26. Breci, L. A., Tabb, D. L., Yates, J. R. & Wysocki, V. H. Cleavage N-terminal to proline:analysis of a database of peptide tandem mass spectra. *Anal. Chem.* **75**, 1963–1971 (2003).
27. Seidler, J., Zinn, N., Boehm, M. E. & Lehmann, W. D. De novo sequencing of peptides by MS/MS. *Proteomics* **10**, 634–649 (2010).
28. Tabb, D. L., Huang, Y., Wysocki, V. H. & Yates, J. R. 3rd Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **76**, 1243–1248 (2004).
29. Medzihradszky, K. F. & Chalkley, R. J. Lessons in de novo peptide sequencing by tandem mass spectrometry. *Mass Spectrom. Rev.* **34**, 43–63 (2015).
30. Ryan, W. E. & Lin, S. *Channel Codes: Classical and Modern* (Cambridge Univ. Press, 2009).
31. MacKay, D. J. C. & Neal, R. M. Near Shannon limit performance of low density parity check codes. *Electron. Lett.* **33**, 457–458 (1997).
32. Reed, I. S. & Solomon, G. Polynomial codes over certain finite fields. *J. Soc. Indust. Appl. Math.* **8**, 300–304 (1960).
33. Trauger, S. A. et al. High sensitivity and analyte capture with desorption/ionization mass spectrometry on silylated porous silicon. *Anal. Chem.* **76**, 4484–4489 (2004).
34. Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein sequencing. *Nat. Nanotechnol.* **13**, 786–796 (2018).
35. Callahan, N., Tullman, J., Kelman, Z. & Marino, J. Strategies for development of a next-generation protein sequencing platform. *Trends Biochem. Sci.* **45**, 76–89 (2020).
36. Swaminathan, J. et al. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* **36**, 1076–1082 (2018).
37. Ng, C. C. A. et al. Data storage using peptide sequences. *Protoc. Exch.* https://doi.org/10.21203/rs.3.pex-1543/v1 (2021).
38. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
39. Valkenborg, D., Jansen, I. & Burzykowski, T. A model-based method for the prediction of the isotopic distribution of peptides. *J. Am. Soc. Mass. Spectrom.* **19**, 703–712 (2008).
40. Tabb, D. L., Saraf, A. & Yates, J. R. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* **75**, 6415–6421 (2003).
41. Tabb, D. L., Ma, Z.-Q., Martin, D. B., Ham, A.-J. L. & Chambers, M. C. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J. Proteome Res.* **7**, 3838–3846 (2008).
42. Yan, Y., Kusalik, A. J. & Wu, F.-X. NovoHCD: de novo peptide sequencing from HCD spectra. *IEEE Trans. Nanobioscience* **13**, 65–72 (2014).

Acknowledgements

We thank Prof. Henry Lam (Hong Kong University of Science and Technology) and Prof. Jiang Xia (Chinese University of Hong Kong) for their helpful discussions, Dr. Bin Hu (Jinan University) for his assistance, and University Research Facility in Chemical and Environmental Analysis and University Research Facility in Life Sciences of The Hong Kong Polytechnic University for the technical supports. Z.-P.Y. would like to give thanks to Fudan8420 WeChat Group for the helpful discussion. This work was supported by Hong Kong Research Grant Council (Grant Nos. R5013-19F, C5031-14E, 153041/17 P, 15304020, C4002-17G, and R4005-18), Natural Science Foundation of China (Grant Nos. 81874306 and 81601828), and China Resources Life Sciences Group Limited.

Author contributions

C.C.A.N. and Z.-P.Y. initiated and designed the experiments; W.M.T. and F.C.M.L. developed the error-correction schemes; C.C.A.N. performed the LC-MS/MS analysis with the assistance from H.Y., Q.W., P.-K.S. and M.Y.-M.W.; W.M.T., F.C.M.L., C.C.A.N., Z.-P.Y. and H.Y. designed and optimized the in-house software; W.M.T. performed the peptide sequence assignments with the assistance from F.C.M.L., C.C.A.N. and H.Y.; C.C.A.N. and W.M.T. drafted the manuscript, and Z.-P.Y. revised the manuscript with the contributions also from F. C.M.L., H.Y., P.-K.S. and M.Y.-M.W.; Z.-P.Y. coordinated the whole project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-24496-9.

Correspondence and requests for materials should be addressed to F.C.M.L. or Z.-P.Y.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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