Data storage using peptide sequences

Cheuk Chi A. Ng
State Key Laboratory of Chemical Biology and Drug Discovery, Food Safety and Technology Research Centre and Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

Wai Man Tam
Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

Haidi Yin
State Key Laboratory of Chemical Biology and Drug Discovery, Food Safety and Technology Research Centre and Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

Qian Wu
State Key Laboratory of Chemical Biology and Drug Discovery, Food Safety and Technology Research Centre and Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

Pui-Kin So
University Research Facility in Life Sciences, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

Melody Yee-Man Wong
University Research Facility in Chemical and Environmental Analysis, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

Francis C. M. Lau (✉ francis-cm.lau@polyu.edu.hk)
Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

Zhong-Ping Yao (✉ zhongping.yao@polyu.edu.hk)
State Key Laboratory of Chemical Biology and Drug Discovery, Food Safety and Technology Research Centre and Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

Method Article

Keywords: data storage, peptides, mass spectrometry, proteomics, error-correction scheme, peptide sequencing

Posted Date: June 25th, 2021

DOI: https://doi.org/10.21203/rs.3.pex-1543/v1

Abstract

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. Moreover, the amount of data generated has been increasing exponentially, from several ZB in 2008 to an expected 74 ZB in 2021, causing a much increased demand for data storage correspondingly. 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 ten to twenty 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.

Here we report the use of peptide sequences for digital data storage, a method that has not been reported before. The data-bearing peptides are commercially synthesized, and the data retrieval process is described here. As an example, we stored one dataset consists of (i) 848 bits of ASCII formatted text in 40 peptides, and (ii) another dataset consists of 13752 bits of the “silent night” music in MIDI format together with its title in ASCII format in 511 peptides. These files are available in Supplementary Files section.

Introduction

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 an 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 ten to twenty 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.

Here we report the use of peptide sequences for digital data storage, a method that has not been reported before[3]. The data-bearing peptides are commercially synthesized, and the data retrieval process is described here. As an example, we stored one dataset consists of 848 bits of ASCII formatted text in 40 peptides, and another dataset consists of 13752 bits of the “silent night” music in MIDI format together with its title in ASCII format in 511 peptides. These files are available in Supplementary Files section.

Reagents

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.

Equipment

LC: 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)

MS: Orbitrap Fusion Lumos mass spectrometer (ThermoFisher Scientific, San Jose, CA)

Procedure

1. Dilute the stock peptide mixture by 50% acetonitrile (ACN)/water with 0.1% formic acid (FA). The final concentration of the 40-peptide mixture was 10 ng/uL while the 511-peptide mixture was 1.02 ug/uL.

2. Prepare the mobile phase A: 0.2% FA in water, and mobile phase B: 0.2% FA in ACN.

3. Mount the C18 column on the UPLC system.

4. Set the LC method:

- Injection volume = 1 uL

- Flow rate = 0.3 mL/min

- Column temperature = 55 ^{o}C

- Gradient:

- 0 min: 10% B

- 2 min: 18% B

- 8 min: 22% B

- 48 min: 34% B

- 64 min: 40% B

- 75 min: 55% B

- 78 min: 80% B

- 83 min: 80% B

- 83 min: 80% B

- 83.01 min: 10% B

- 90 min: 10% B

5. Set the MS method:

General:

- spray voltage = +3600 V

- ion transfer tube temperature = 280 ^{o}C

- vaporizer temperature = 280 ^{o}C

MS1 scan parameters:

- scan range = *m/z* from 900 to 1400 Da

- resolution = 30K

Ion selection parameters:

- use quadruple

- if available, use advanced peak determination (APD) with default charge of +2

- top-speed mode with 3 s cycles

- mass tolerance = 25 ppm

- dynamic exclusion window = 4 s

- isolation window width = 1.6 Da

MS2 scan parameters:

- high-energy collision dissociation (HCD) at 28% of normalized collision energy

- stepped collision energy = 5%

- scan range = *m/z* from 240 to 2450 Da

- resolution of 15K.

6. Run the LC-MS/MS process using the set parameters. For better results, repeat the process 2-3 times.

7. Use MSConvert (a part of ProteoWizard), convert the .RAW spectral file to .ms1 and .ms2 files.

8. Using a custom script, preprocess the spectra in the .ms1 and .ms2 files. The preprocess includes 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.

9. Using another custom script, analyze the preprocessed spectra to obtain a list of correct sequence candidates, then group and score the candidates to obtain the correct sequences. If the LC-MS/MS process is repeated, all such candidates are included in the grouping and scoring processes.

10. Using another custom script, convert the correct sequences back to sequences of 0s and 1s, according to the predefined encoding (Fig. 1) and error-correction schemes. The original file is obtained.

Troubleshooting

Time Taken

Steps 1-2: < 1h

Steps 3-6: About 2 h

Steps 7-10: About 4 h per .RAW file

Anticipated Results

For 40 peptides, >38 peptides can be sequenced correctly;

For 511 peptides, >90% of the peptides can be sequenced correctly.

After error-correction according to the encoding scheme, the stored data can be retrieved correctly.

References

1. Hilbert, M. & López, P. The world's technological capacity to store, communicate, and compute information. *Science* **332**, 60 (2011).

2. https://www.statista.com/statistics/871513/worldwide-data-created/ (accessed on 28 May 2021).

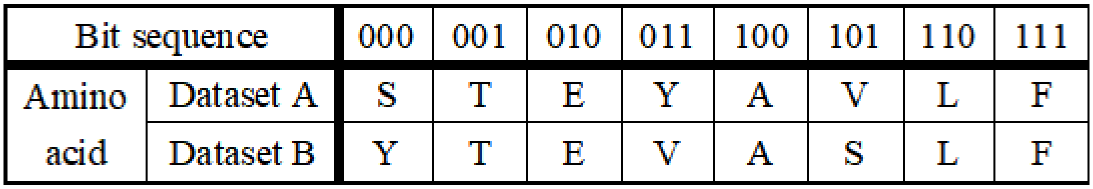
3. Yao, Z.P., Ng, C.C.A., Lau, C.M. & Tam, W.M. Data storage using peptides. US Provisional Patent Application No. 62/657,026 (Filed on 13 April 2018); PCT Application No. PCT/CN2018/119349 (Filed on 6 December 2018); US Non-Provional Patent Application No.16/224,957 (Filed on 19 December 2018).

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. **Funding:** This work was supported by Hong Kong Research Grant Council (Nos. R5013-19F, C5031-14E, 153040/15P, 153348/16P, 153041/17P, C4002-17G and R4005-18), China Resources Life Sciences Group Limited, and University Research

Facility in Chemical and Environmental Analysis and University Research Facility in Life Sciences of The Hong Kong Polytechnic University.

Figures



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

Figure 1

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

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- datasetAtext.txt
- datasetBsilentNight.mid
- datasetBtitle.txt