INVESTIGATION OF NEW PEPTIDE DESIGN, PEPTIDE STABILITY AND 2D LC-MS/MS FOR DATA STORAGE AND RETRIEVAL WITH PEPTIDES

DAI JUN

MPhil

The Hong Kong Polytechnic University

2022

The Hong Kong Polytechnic University

Department of Applied Biology and Chemical Technology

Investigation of new peptide design, peptide stability and 2D LC-MS/MS for data storage and retrieval with peptides

DAI Jun

A Thesis Submitted in Partial Fulfillment of the Requirements of the Degree of Master of Philosophy

July 2022

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

DAI Jun (Name of student)

Abstract

With the amount of data generated growing exponentially, there is an urgent need to develop new storage technology. Peptides have shown great potential as a storage material in the previous study. This study was conducted to improve data storage with peptide sequence and to prove the kinetic stability of peptides for data storage. Two mass spectrometry technologies were used for the quantitative analysis of peptides and *de novo* peptide sequencing.

We have improved the encoding scheme and peptide design based on the previous study to achieve higher capacity, and the successful retrieval of a 1088-bit text file suggested the feasibility of the new encoding scheme. With a new dataset, a 96,224 bits long PNG file, 4095 peptides sequences have been generated using the new encoding scheme. The 4095 peptides will be synthesized and tested in the next step.

Additionally, the stability of peptides has been demonstrated in this study. A triple quadrupole mass spectrometer (QqQ-MS) was used for quantitative analysis, and 11 peptides were chosen for the study. Accelerated aging experiments were performed to measure peptide decay kinetics. Our study showed that the half-life of these peptides was within a range of 10 - 100 years at -20℃ without any extra protective measure, and by freeze-drying with trehalose, the half-life at -20℃ could be improved to more than 200 years, which suggested the potential of peptide as the data storage medium.

Two-dimensional liquid chromatography (2D-LC) was applied for *de novo* peptide sequencing to retrieve data. 2D-LC has been reported as a powerful tool applied for protein identification with peptide fingerprinting. In this study, we applied 2D-LC in

the peptide-based data storage system to retrieve the information stored in 40 peptides, and the result was compared with that of ultra-performance liquid chromatography (UPLC) and nano-liquid chromatography (nano-LC). Theoretically, separation capacity could be higher with 2D-LC. However, the lower signal-to-noise ratio due to successive dilutions and salt residue caused lower recovery.

This study demonstrated the feasibility of peptides used for data storage and clarified the essential need for further research. The findings regarding peptide stability and 2D-LC are expected to have an active role in the development of relevant fields.

Research publications

Conference Publication

1. Dai, J.; Zhou, Y.; Ng, C. C. A.; Yao, Z. P. Investigation of Peptide Stability for Data Storage, The 28th Symposium on Chemistry Postgraduate Research in Hong Kong, Hong Kong Baptist University, 6 November **2021**. (Poster)

Acknowledgments

First and foremost, I would like to extend my sincerest gratitude to my supervisor Prof. Zhongping Yao, for giving me the guidance, patience, support, and valuable advice during my study. I joined Prof. Yao's group two years ago. In the past two years, I have learned comprehensive knowledge of mass spectrometry. And more significantly, the valuable experiences studying here make me learn the importance of being patient while conducting research.

I would like also thanks to Dr. Cheuk Chi Ng, Dr. Wai Man Tam, and Dr. Long Wu for their selfless assistance. I must specially thank Dr. Sirius Tse, Dr. Pui Kin So, Dr. Carol Tsang, and Dr. Chi-Hang Chow for their useful suggestions on my research.

I thank all my team members for their kind advice and support during the study. Special thanks are given to Dr. Cheuk Chi A. Ng and Mr. Eugene Zhen Yan Li for their valuable suggestions about my thesis.

Last but not least, gratitude should be given to my friends, for their kind support over the past years, which motivates me to keep moving forward and pursuing my goals.

Table of contents

List of abbreviations

Full Form	Abbreviation
3,5-disubstituted tetrahydro-2H-1,3,5-thiadiazine-2-thione	THTT
acetonitrile	ACN
alanine	Ala (A)
American Standard Code for Information Interchange	ASCII
arginine	Arg (R)
asparagine	Asp (N)
atmospheric pressure chemical ionization	APCI
Blu-Ray Disc	BD
chain ejection model	CEM
charged residue model	CRM
chemical ionization	CI
circular dichroism	CD
collision gas	CAD
collision-induced dissociation	CID
compact disc	CD
comprehensive two-dimensional liquid chromatography	LC x LC
curtain gas	CUR
cysteine	Cys (C)
Dalton	Da
declustering potential	DP
deoxyribonucleic acid	DNA
digital versatile disc	DVD
dimethyl sulfoxide	DMSO

direct current	DC
electron ionization	EI
electrospray ionization	ESI
elementary charge	e
entrance potential	EP
Extended Binary Coded Decimal Interchange Code	EBCDIC
field ionization	FI
formic acid	FA
gigabyte	GB
glutamic acid	Glu (E)
glutamine	Gln (Q)
glycine	Gly (G)
hard disk drive	HDD
heart-cutting two-dimensional liquid chromatography	LC-LC
high performance liquid chromatography	HPLC
higher energy collision dissociation	HCD
histidine	His (H)
Human Genome Project	HGP
human insulin-like growth factor I	hIGF-I
hydrophilic interaction chromatography	HILIC
internal standard	IS
ion evaporation model	IEM
ion exchange	IEX
ion routing multipole	IRM
ionspray voltage	IS

isoleucine	Ile (I)
leucine	Leu (L)
limit-of-detection	LOD
limit-of-quantitation	LOQ
liquid chromatography	LC
liquid chromatography coupled with tandem mass spectrometry	LC-MS/MS
liquid chromatography-mass spectrometric multiple reaction monitoring	LC-MRM
liquid chromatography-multiple reaction monitoring-MS	LC-MRM-MS
liquid chromatography-tandem mass spectrometry	LC-MS/MS
low-density parity-check	LDPC
Lysine	Lys (K)
mass-to-charge ratio	*m/z*
mass spectrometry	MS
matrix-assisted laser desorption/ionization	MALDI
megabyte	MB
methionine	Met (M)
mitochondrial DNA	mtDNA
molecular dynamics	MD
multilevel cell	MLS
multiple reaction monitoring	MRM
nano-liquid chromatography	nano-LC
parts-per-million	ppm
phenylalanine	Phe (F)
proline	Pro (P)

polymerase chain reaction	PCR
predicted local-distance difference test	pLDDT
radio frequency	RF
Reed-Solomon	RS
relative standard deviation	R.S.D.
reversed-phase	RP
selected reaction monitoring	SRM
serine	Ser (S)
single level cell	SLC
size exclusion chromatography	SEC
solid-state drive	SSD
strong cation exchange	SCX
substance P	SP
tandem mass spectrometry	MS/MS
terabyte	TB
threonine	Thr (T)
time-of-flight	TOF
triple quadrupole mass spectrometer	QqQ-MS
tryptophan	Trp (W)
two-dimensional liquid chromatography	2D-LC
two-dimensional liquid chromatography/mass spectrometry	2D-LC-MS
two-dimensional liquid chromatography coupled with tandem mass spectrometry	2D-LC-MS/MS
tyrosine	Tyr (Y)
ultra-performance liquid chromatography	UPLC

Unicode Transformation Format – 8-bit	UTF-8
valine	Val (V)
zettabyte	ZB

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 1: Introduction

1.1 Data explosion

Data is the new oil of the digital economy.[1] Even before humans learned how to speak, we have figured out ways to store information.[2] Alarmingly, the amount of data created is growing at an unprecedented rate. The total amount of data generated globally increased from 2 zettabytes (ZB) in 2010 to 64.2 ZB in 2020 (Figure 1-1).[3] Over the next three years up to 2025, global data creation is projected to grow to more than 180 ZB.[3] Another research indicated that global memory demand will exceed the projected silicon supply by 2040.[4] Therefore, there is a need for new data storage methods.

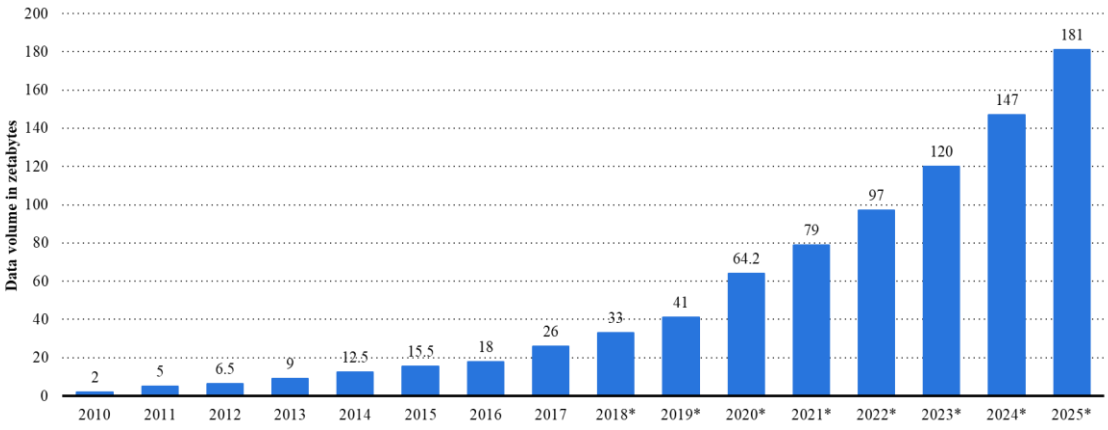


Figure 1-1. The total amount of data created, consumed, and stored globally each year. (Reprinted from ref[3])

1.2 Traditional data storage technology

Before the advent of computers, methods for information storage were limited to fewer forms, such as paper and film.[5] Nowadays, with the advancement of computer and communication technologies, all the data could be stored as strings of binary numbers, as shown in Figure 1-2.

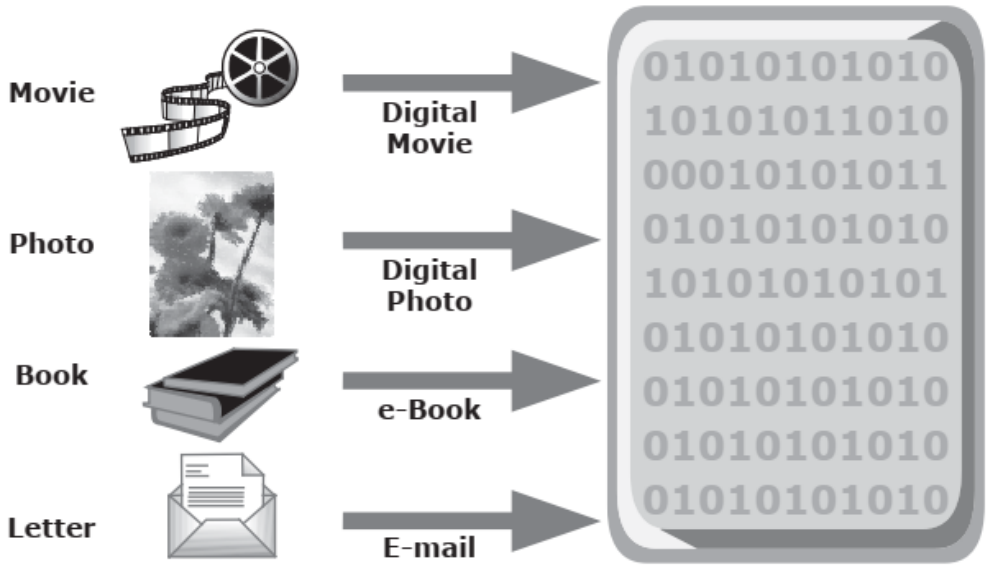


Figure 1-2. Information is translated into binary data for storage. (Reprinted from ref[5])

1.2.1 General introduction of digital data storage devices

The storage devices currently employed could be divided into four categories: magnetic tape, optical disk, hard disk drive (HDD), and solid-state drive (SSD). The product positioning of the four storage devices is different. Currently, magnetic tape has already withdrawn from the ranks of personal data storage and is more used for the backup of massive data. Due to its higher data density and lower cost, it is favored by some internet companies, such as Google and Sony. The optical disk is mostly used for personal data storage. Although it takes longer to retrieve data, the lower per-unit prices make it widely used over a long period. Recently, optical disks are being replaced by

HDDs and SSDs due to falling prices of HDDs and SSDs. HDD and SSD are relatively new storage methods, which are widely used in computer systems due to their low access times. Compared with HDDs, SSDs could achieve 1000 times faster transfer speeds but come with a higher cost.[6]

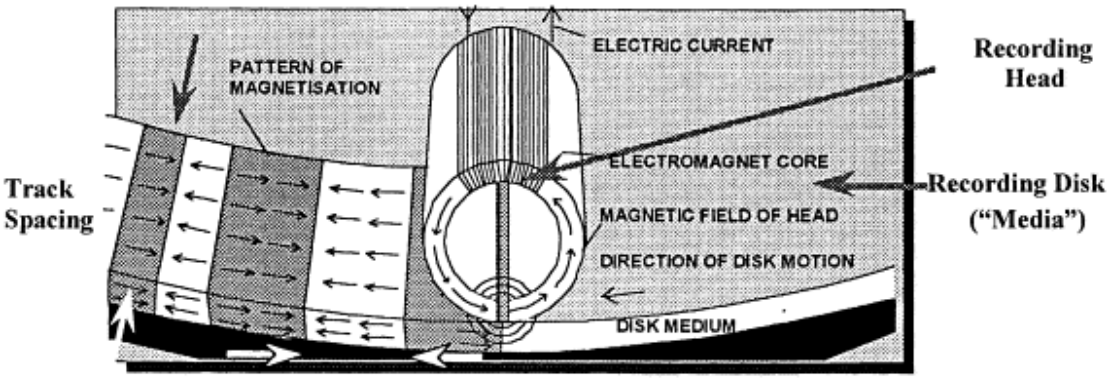


Figure 1-3. Schematic diagram of the working principle of magnetic tape. (Reprinted from ref[7])

The magnetic tape consists of two layers: the base film and the magnetic layer, as shown in Figure 1-3. The magnetic layer is the top coat and the base film is the substrate. Additionally, a back coat can be used to reduce tape friction and dissipate static charge.[8] To store the information, the binary data is translated into voltage pulses. The wire wound around the recoding head carries the electrical signal, which produces a magnetic field. Then, the magnetic layer is magnetized when exposed to the magnetic fields. Data retrieving is the inverse of a recording procedure. Magnetic tape has a great advantage when dealing with large amounts of data. In 2014, Sony announced that magnetic tape could hold 185 terabytes (TB) of data by creating a nano-grained magnetic layer.[9]

Binary data stored on the optical disk is encoded in the form of pits and lands, as shown in Figure 1-4. To retrieve data, a laser beam will be emitted from an optical disk drive. By distinguishing the differences in reflectivity, an optical disk drive could determine the 0 and 1 bits that represent the data.[10] Nowadays, most of the available optical disks are in one of the following three formats depending on the wavelength of the laser used for reading: compact discs (CDs), digital versatile discs (DVDs), and Blu-Ray Discs (BDs). BDs could achieve higher capacity using smaller pits and lands and reading with a blue-violet laser.[11]

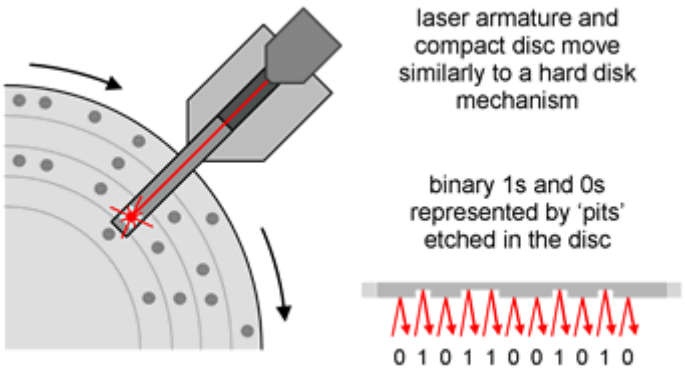
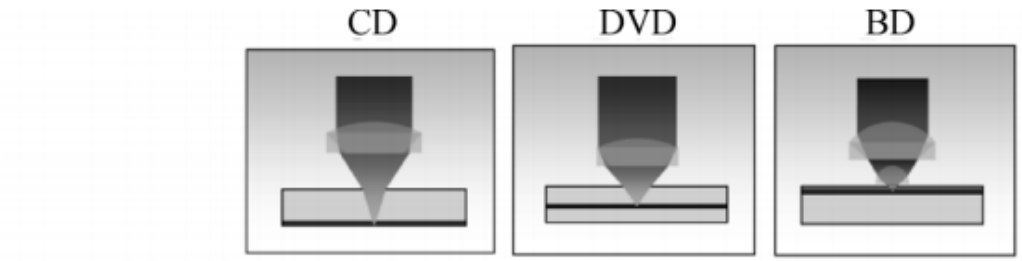


Figure 1-4. Schematic diagram of reading data from an optical disk. (Reprinted from ref[10])



	CD	DVD	BD
capacity/layer	0.7 GB	4.7 GB	25 GB
data rate	1.2 Mb/s	1X: 11 Mb/s	1X: 36 Mb/s
laser wavelength	780 nm	650 nm	405 nm
objective lens NA	0.45	0.60	0.85
cover thickness	1.2 mm	0.6 mm	0.1 mm

Figure 1-5. Comparison of main parameters of CD/DVD/BD. (Reprinted from ref[11])

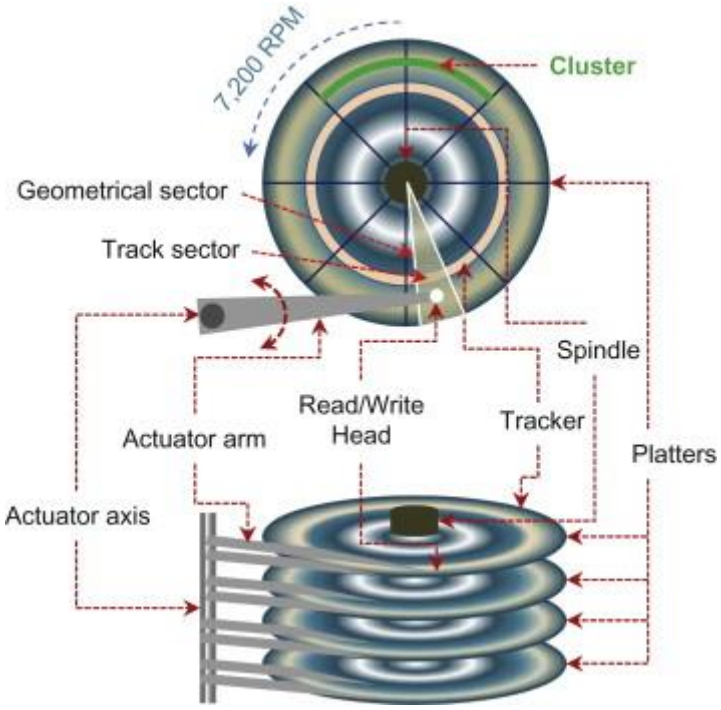


Figure 1-6. Schematic diagram of a hard disk drive. (Reprinted from ref[6])

HDD is another magnetic storage media. But different from magnetic tape, an HDD is made of several magnetic platters, as shown in Figure 1-6. The platters are paired with magnetic heads, which are used to retrieve data. The data is accessed in a random-

access manner, which lowers the access times. Similar to magnetic tapes, HDDs are also suitable for large-capacity data storage. These characteristics make the HDD the dominant storage device of the computer.

SSD is flash technology-based storage that uses integrated circuit components to store data. By 2004, since the cost of flash had plummeted, SSDs became an alternative to magnetic storage devices.[12] These SSDs have millions of memory cells based on floating-gate architectures.[13] They are designed to trap electrons. Each memory cell has an inlet with a valved transistor. [13] To store the data, the inlet transistor will open to let a charge enter the cell and remain there. The cells with no charge represent binary data "1", and those with a charge represent binary data "0".[13]

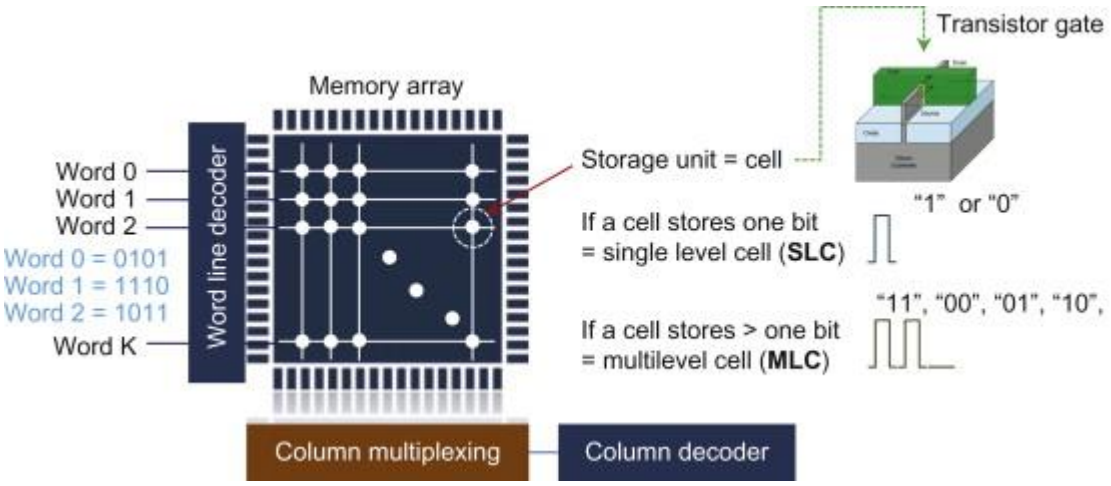


Figure 1-7. Schematic diagram of the working principle of SSD. (Reprinted from ref[6])

1.2.2 Stability and longevity of digital data storage devices

As mentioned in the previous section, we have multiple efficient methods to store the data. However, these methods are not durable enough. The optical disks and magnetic tapes are vulnerable to physical damage and are sensitive to temperature and

humidity.[14] HDDs are susceptible to magnetic fields.[15] Both data corruption and mechanical damage could be caused by the magnet. As for SSDs, if the memory cells are flashed too frequently, the transistor gate could become worn out and eventually be broken down.[6] A study suggested that magnetic tapes can only store data for 5 years, and HDDs might start degrading after 3 years.[16, 17] Besides, although these storage methods could achieve very high data density, e.g., magnetic tape could achieve a data density up to 148 GB per square inch, they are insufficient for rapidly growing data.[9] Therefore, a new data storage method that can achieve a higher data density and longevity is essential. Given the background, molecular digital data storage was proposed since storing data in molecule instead of mechanical structure could lead to much higher data density.

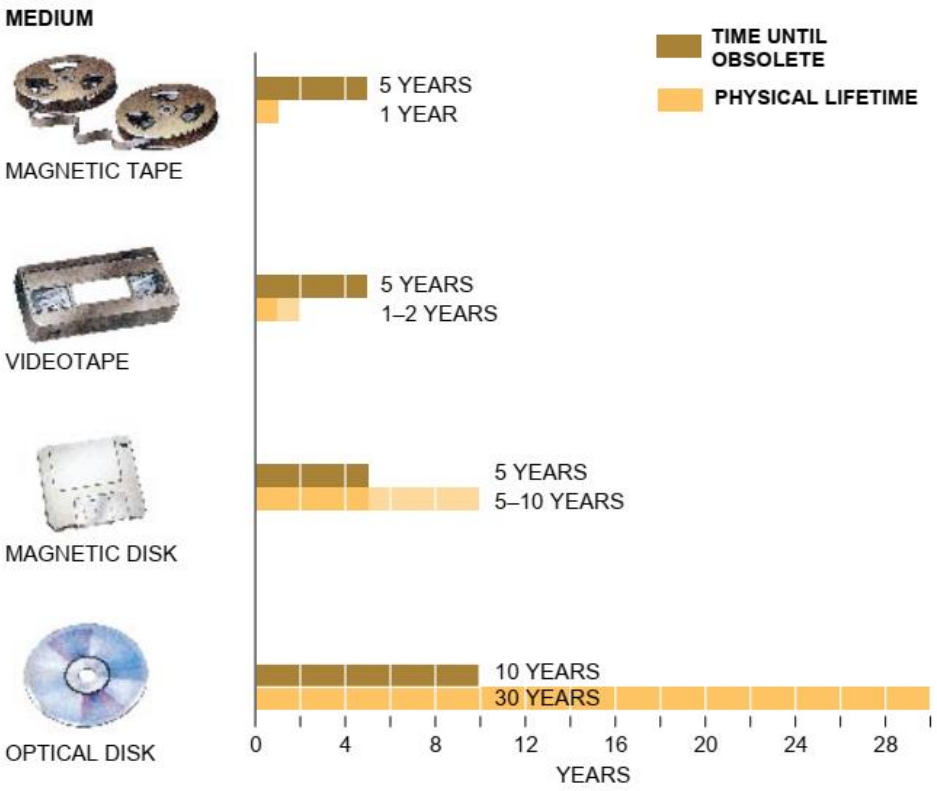


Figure 1-8. Expected lifetimes of common digital storage media. (Reprinted from ref[16])

1.3 Molecular data storage technology

1.3.1 DNA-based data storage

Deoxyribonucleic acid (DNA) is the most researched one among all the molecules for data storage. As the carrier of genetic information, DNA has been used for information storage for billions of years. Joe Davis is the first one to use DNA sequences to store information. In 1988, he translated a 35 bits code into a DNA sequence and stored the information within *E. coli*.[18] With the DNA sequencing technology well developed during the Human Genome Project, the cost of determining one megabase of DNA has been reduced from USD 5,292.39 in 2001 to USD 0.52 in 2010.[19] Besides, microchip-based DNA synthesis technologies also highly promoted the research using DNA as digital data storage material.[20] With the development of these technologies, Church and Goldman proposed a feasible idea in 2012: using an oligo library for data storage.[21, 22] The data stored in DNA was segmented into several strings with a certain length. The DNA sequence consisted of three parts: the sequence containing the information, the address sequence, and the 5' and 3' primer binding sites for library amplification and sequencing.[23] The information was retrieved with a total of 10-bit errors, which suggested the feasibility of the idea. After that, more and more studies on DNA-based data storage have been published. These studies proved that molecule digital data storage has several strengths compared with traditional storage methods. A significant advantage of a DNA-based data storage system is its higher storage density. The calculation result showed that the practical density of DNA memory would be close to that of living cells, about 10^{19} bit/cm^3, at least 6 orders of magnitude higher theoretical storage density than the densest media available now.[4] Another advantage is its stability and longevity: DNA has a half-life of more than 500 years in ancient fossil bone.[24] Another study showed that DNA can be stored at 70 ℃ for a week when encapsulated

in silica, which is thermally equivalent to storing DNA in Zurich (9.4 °C) for 2000 years.[25]

Much research has been done to optimize DNA-based data storage systems, and DNA could be sequenced and synthesized at a much lower price compared with that a few years ago. However, current storage systems are still not good enough for mass storage, because the cost to synthesize oligo library is excessively expensive compared with available storage methods, and it takes too long to get DNA sequenced and read.[19, 26, 27] Therefore, multiple other molecular approaches have been investigated.

Table 1-1. Summary of features of current DNA-based storage platforms.

Group	Year	Input data (MB)	Coding potential (bits/bp)	Redundancy	Length of playloads (bp)	Net density (bits/bp)
Church[21]	2012	0.65	1	1	115	0.83
Goldman[22]	2013	0.75	1.58	4	117	0.33
Grass[28]	2015	0.08	1.78	1	117	1.14
Yazdi[29]	2015	0.17	—	—	1000	1.575
Bornholt[30]	2016	0.15	1.58	1.5	120	0.88
Blawat[31]	2016	22	1.6	1.13	190	0.92
Erlich[27]	2017	2.15	1.98	1.13	152	1.57
Yazdi[32]	2017	0.0036	—	—	1000	1.72
Erlich and Grass[33]	2020	0.045	1.98	5.4	104	—

1.3.2 Peptide-based data storage

Among all the other molecules, peptides could be synthesized and sequenced with mature techniques, and they can achieve higher sequence complexity than DNA. Compare with DNA, there are more choices of natural monomers. Besides, extensive

studies have been conducted to incorporate unnatural amino acids into proteins. Therefore, attracted by the properties of peptides, Prof. Yao's group has proposed the idea: data storage using peptide sequences.[34]

Peptide synthesis

Peptides are chains of amino acids linked to each other by the peptide bond. In order to store information using peptides, synthesizing peptides with specific sequences is essential. The chemical synthesis of peptides was first achieved by Theodore Curtius and Emil Fischer.[35, 36] In 1901, Emil Fischer published his work on dipeptide synthesis by the hydrolysis of diketopiperazines of glycine.[35] Then, the azide-coupling method was developed by Theodore Curtius in 1904, which was the first practical method for peptide synthesis.[37] In 1903s, the introduce of a removable temporary amino protecting group provided a new thought.[38] In 1963, the chemical synthesis of peptides ushered in a new revolution. The strategy of peptide synthesis was dramatically changed by the work of Bruce Merrifield, who invented solid-phase peptide synthesis (SPPS).[39]

Currently, the SPPS strategy is still widely used for peptide synthesis. The principle of SPPS is shown in Figure 1-9. In the general procedure of SPPS, the first amino acid is attached to a resin, which acts as an insoluble support. Followed by N-terminal deprotection, the amino acid is coupled with another N-terminal protected amino acid and leads to peptide chain elongation. The cycles of deprotection and coupling reactions repeat, and the byproducts are removed after each cycle until the target peptide is synthesized.

Compared with solution phase peptide synthesis, the SPPS approach simplifies the steps of purification and enables automatization of the processes, which allows high-throughput peptide synthesis. These characteristics of SPPS enable the industrialization of peptide synthesis. Since peptides could be synthesized with a reasonable price, peptide-based data storage may become economically viable.

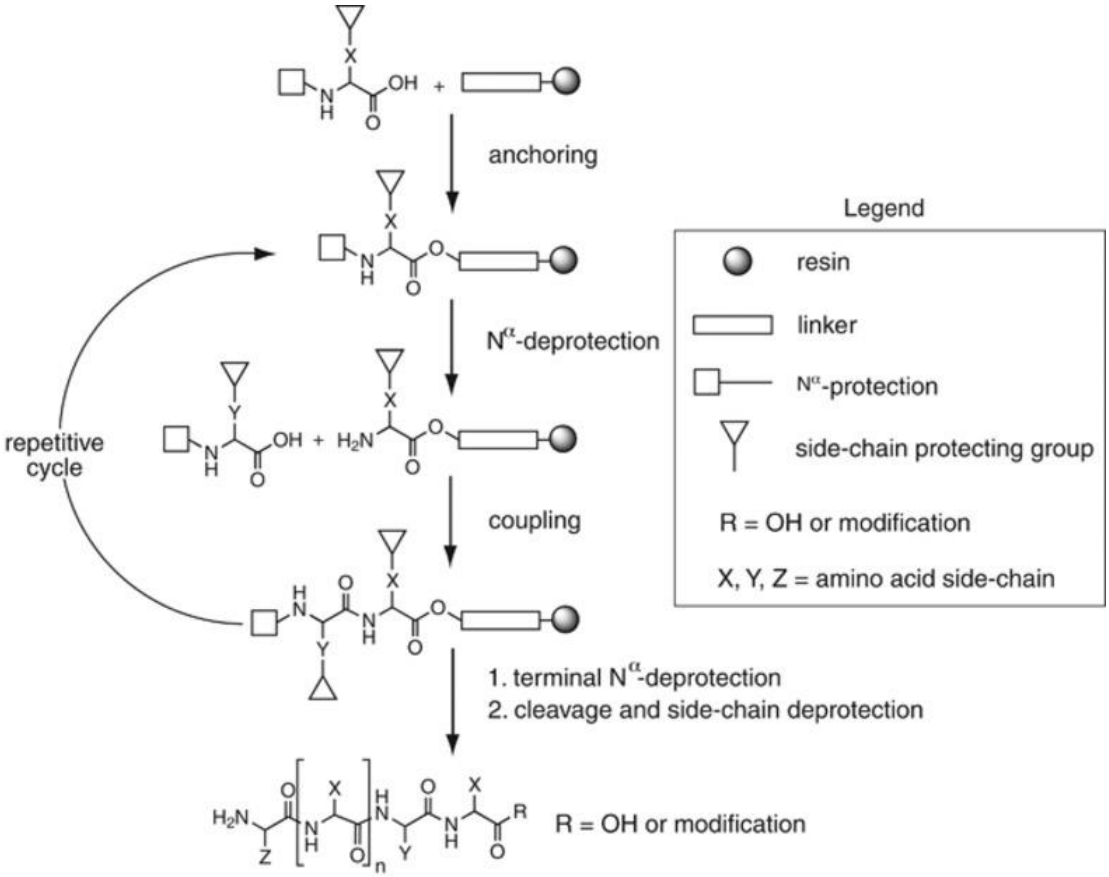


Figure 1-9. Principle of SPPS.[40]

Peptide sequencing

In 1958, Sanger won his first Nobel Prize for his work that completed the first amino acid sequence determination of insulin.[41] Ever since then, multiple peptide sequencing methods have been developed. In 1975, Hughes, J. and his colleagues determined the sequence of enkephalin with mass spectrometry.[42] It was the first time that a mass

spectrometer was used to decipher peptide sequences. After that, with the development of soft ionization techniques, biomolecules analysis with mass spectrometry became more feasible, and mass spectrometry became a powerful tool for peptide sequencing.

De novo peptide sequencing is a method that determines the peptide sequence with tandem mass spectrometry (MS/MS).[43] [44] First, peptides are ionized by the ion source. Then analyzed with the mass analyzer, the chosen precursor ions are fragmented.[45] The daughter ions are sent to another mass analyzer for the generation of MS/MS spectra (Figure 1-10). By comparing the masses of amino acids with the mass differences between pairs of fragment ion peaks, the amino acid sequence can be determined.[43] The robustness of MS/MS is further enhanced when coupled with liquid chromatography (LC). Typically, liquid chromatography-tandem mass spectrometry (LC-MS/MS) can be used to qualitatively identify thousands of proteins at once; even if the proteins without fully sequenced, a reference database enables accurate identification if the proteins are present in the database.[44] However, this strategy cannot be used in peptide data storage. Since there is no reference database for data retrieving, nearly all the amino acids should be recovered. Therefore, the peptide sequences need to be elaborately designed to ensure full coverage.

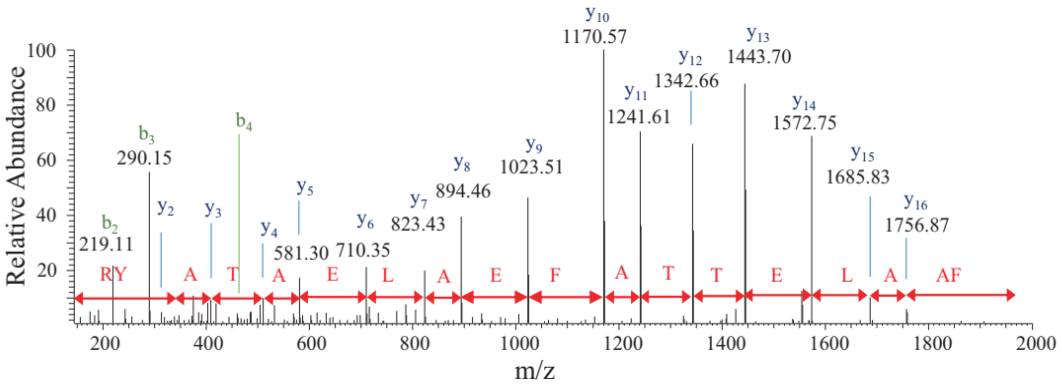


Figure 1-10. A representative MS/MS spectrum for peptide sequence determination. (Reprinted from ref[34])

During data retrieving, peptides were read in typical "beads on a string" manner. The "0" and "1" bits were assigned following specific rules, and every three bits were translated into an amino acid to get the peptide sequences (Figure 1-11). Prof. Yao's group has used peptide mixtures instead of oligo library for data storage. Solid-phase peptide synthesis was applied for peptide synthesis and LC-MS/MS was applied for peptide sequencing. The details of mass spectrometry were introduced in the next section.

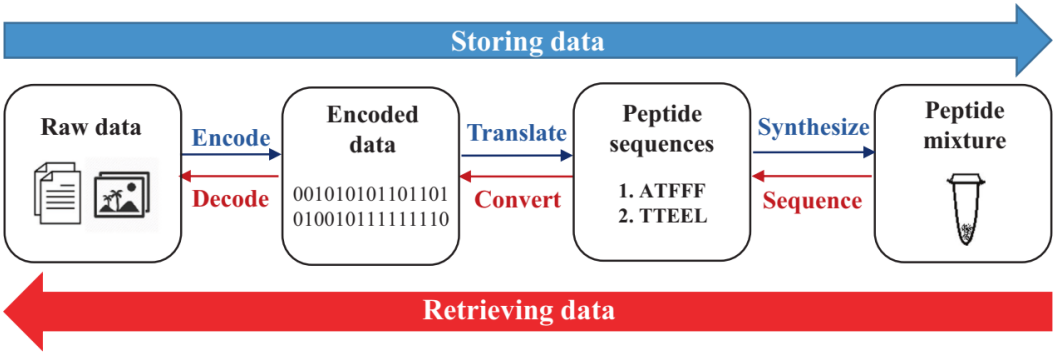


Figure 1-11. Schematic diagram of data storage with peptide sequences. (Reprinted from ref[34])

1.4 Mass spectrometry

1.4.1 General introduction

Mass spectrometry is an indispensable analytical tool in many fields of science. By determining the mass-to-charge ratios (*m/z*) and intensity of the peaks in the spectrum of analyzed samples, MS can provide multiple information, including the molecule masses, the compositions, and the concentrations of the analytes. Therefore, mass spectrometry can be used for both qualitative and quantitative analysis. There are three basic components in a typical mass spectrometer: an ion source, an analyzer, and a detector.

1.4.2 Ion source

Since the mass analyzer can only handle the charged species, the ion source is needed to convert analytes into gas-phase ions. Multiple ionization methods have been developed, e.g., electron ionization (EI), chemical ionization (CI), field ionization (FI), electrospray ionization (ESI), matrix-assisted laser desorption/ionization (MALDI), and atmospheric pressure chemical ionization (APCI), that could ionize the analytes to form ions in several ways. In this study, ESI was used for peptide sequencing and quantitative analysis.

The principle of ESI was proposed by Malcolm Dole in 1968.[46] By dissolving the non-volatile polymer in the volatile solvent (benzene: acetone=1:2) and producing highly charged droplets of the solvent, the droplets were rapidly evaporated to form the intact gaseous ions. Since there was no high energy involved, it is a suitable solution for macromolecules and labile molecules ionization. The research was further developed by John Fenn, who received the 2002 Nobel Prize in Chemistry for his contribution to

the development of ESI.[47] The ion desolvation process is shown in Figure 1-12. The liquid sample enters the electrospray chamber through a needle. The voltage of the needle is maintained at a few kilovolts relative to the surrounding cylindrical electrode, resulting in a strong field at the needle tip that charges the surface of the emerging liquid and disperses it into a fine spray of charged droplets. Then, the droplets migrate toward the inlet end of the capillary driven by the electric field. During the process, a countercurrent drying gas is used for hastening evaporation of the fine droplets, which causes the droplet surface charge repulsion increases. Consequently, the droplet is torn apart and produces charged daughter droplets when the charge repulsion overbears the surface tension. This sequence of events repeats until the droplets are transferred into gas-phase ions.

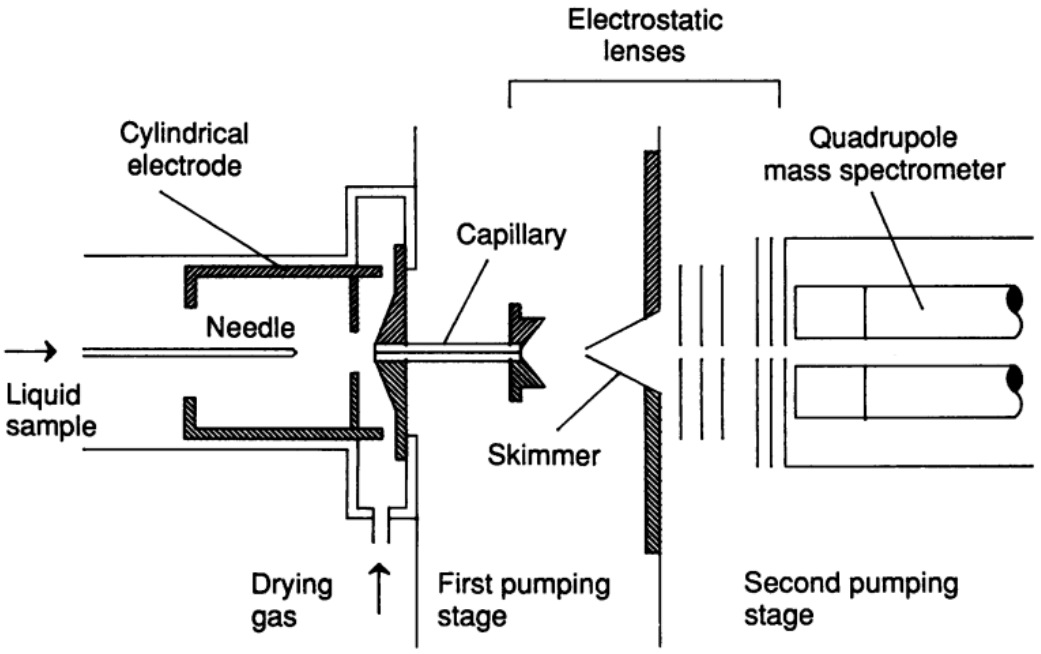


Figure 1-12. Sketch of the ion desolvation process. (Reprinted from ref[47])

The droplet is torn apart when the Rayleigh limit is reached. The Rayleigh limit could be described using the following equation, in which Z_R is the number of elementary charges (e), R is the droplet radius, ε_0 is the vacuum permittivity, and γ is the surface tension.[48]

$$Z_R = \frac{8\pi}{e}\sqrt{\varepsilon_0 \gamma R^3} \tag{1-1}$$

Furthermore, three widely accepted ion release mechanisms have been established to describe the generation of smaller and highly charged offspring droplets. Small analyte ions follow the ion evaporation model (IEM), large analyte ions follow the charged residue model (CRM), and a chain ejection model (CEM) has been proposed to explain the formation of disordered polymers (Figure 1-13).[49, 50] The IEM was proposed by Iribarne and Thomson in 1976.[49] The model suggests that, with the solvent evaporation, the electric field emanating from a Rayleigh-charged nanodroplet (with R less than 10 nm) will be high enough to cause the direct emission of small solvated ions from the charged droplet surface.[49, 50] A molecular dynamics (MD) simulation conducted in 2011 well supported this model.[51] For the ion formation of large globular analytes, such as folded proteins, CRM, which was proposed by Dole, was widely accepted to describe the process.[46, 50] For the CRM, the droplets will continually tear apart until the charged droplets contain only one analyte ion; and as the solvent shell evaporates, the droplets will lose their charge. During the process, CRM nanodroplets remain close to the Rayleigh limit. CEM is used to describe the ion formation of unfolded proteins.[50, 52] A study conducted by Elias Ahadi and Lars Konermann in 2011 suggested that the behavior of polymers with compact conformations was consistent with the CRM, while the disordered polymers behaved in a completely different way.[52] With a disordered protein, unfolded chains immediately migrate to the droplet surface, followed by

ejection of the remaining protein and separation from the droplet. In the meantime, unfolded proteins can provide signals with higher intensity.[52]

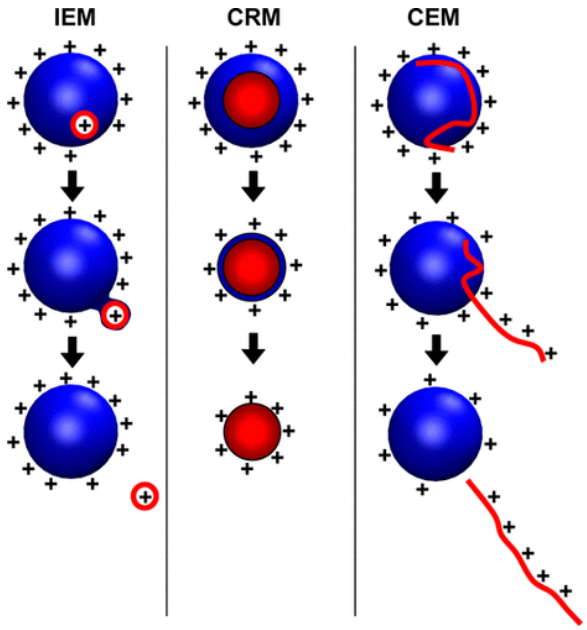


Figure 1-13. Summary of ESI mechanisms. (Reprinted from ref[50])

1.4.3 Mass analyzer

The mass analyzer is used to separate the ions according to their *m/z* values. There are several types of analyzers; the most commonly used types are quadrupole, ion trap, orbitrap, and time-of-flight (TOF). In this study, peptide quantitative analysis was performed with a triple quadrupole mass spectrometer (QqQ-MS), and peptide sequencing was performed with an orbitrap mass spectrometer.

Quadrupole

Quadrupoles are mass analyzers capable of mass-selective operation. A linear quadrupole mass analyzer consists of four rod electrodes mounted in the xy-plane and extending in the z-direction. The rods are connected to radio frequency (RF) and direct

current (DC) generators (Figure 1-14). The voltage applied to the rods changes periodically, therefore, repulsion and attraction in both the x- and the y-directions alternate periodically with time, and the motion of ion can be calculated as following[53]:

$$\frac{d^2x}{dt^2} + \frac{e}{m_i r_0^2}(U + V cos\omega t)x = 0 \quad (1\text{-}3)$$

$$\frac{d^2y}{dt^2} + \frac{e}{m_i r_0^2}(U + V cos\omega t)y = 0 \quad (1\text{-}4)$$

In the equation, *U* means DC voltage, *V* means RF voltage, and ω is the frequency. For a given set of *U*, *V*, and ω, the ion with a certain *m/z* value can be selected to pass the quadrupole. Ions of different m/z can reach the detector sequentially as magnitudes of the RF and DC voltages increase.

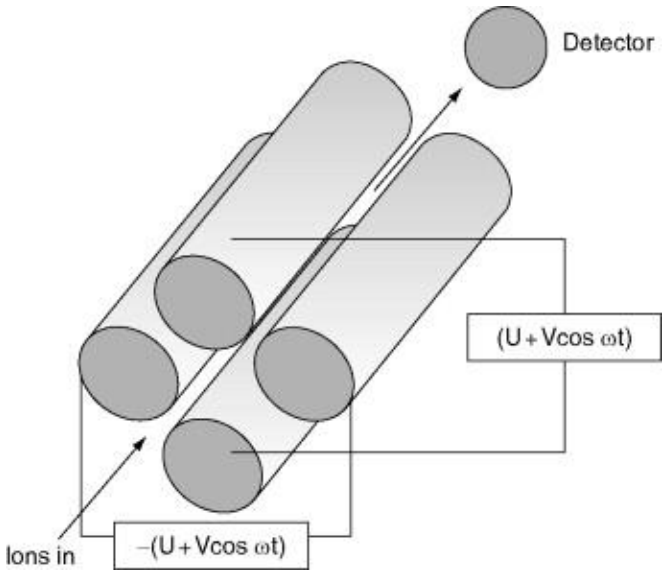


Figure 1-14. Schematic diagram of a quadrupole mass spectrometer. (Reprinted from ref[54])

In QqQ-MS, three quadrupoles are coupled together. It is a commonly used analytical tool, especially for accurate quantitation.[55] It is a mass spectrometer designed for tandem mass spectrometry, which means the precursor ions are fragmented to the

daughter ions for the second mass spectrometric analysis. In a QqQ-MS, the first quadrupole (Q1) and the third quadrupole (Q3) are operated independently for ions scan or selection, while the second quadrupole (Q2) works as a collision cell with collision-induced dissociation (CID). A triple quadrupole mass spectrometer provides four scan modes as shown in Figure 1-15. Two of them, product ion scan selected reaction monitoring (SRM) were used in this study. In product ion scan, Q1 is used to select ions with a particular *m/z* value, and then fragmented in Q2 and the resulted fragment ions are scanned in Q3.[56] Product ion scan can provide structure information for sample identification and peptides sequencing. In SRM, both Q1 and Q3 serve as mass filters. Q1 selects precursor ions (peptides) with a selected *m/z* value which are subsequently fragmented in Q2, then the fragment ions with a selected *m/z* value will be detected in Q3.[57] Multiple reaction monitoring (MRM) is also possible with more than one SRM operated at the same time, and it is a common method in the quantitative analysis of targeted compounds due to its sensitivity and specificity. Therefore, the QqQ-MS was chosen for peptide quantitative analysis in this study.

Orbitrap

The orbitrap analyzer is the latest newcomer to the family of mass analyzers. It was first presented to the public in 1999.[58] In 2005, it was commercialized by Thermo Fisher Scientific. The orbitrap is an ion trap, but instead of using RF or a magnet to hold the ions, it traps the ions with an electrostatic field. The orbitrap mass analyzer consists of a cup-like outer electrode and a spindle-like central electrode (Figure 1-16).[59] With voltage applied between the central and outer electrodes, the ions move around the central electrode in a nearly circular spiral inside the trap, and the ions with different *m/z* values move in different trajectories. Compared with TOF and quadrupole, it

provides higher resolving power, mass accuracy, and sensitivity.[58] Besides, it is also capable of performing tandem mass spectrometry just like an ion trap. These features allow both identification and quantitation analysis, which makes it a powerful tool in proteomics and metabolomics studies. Therefore, the orbitrap mass spectrometer was chosen for peptide sequencing in this study.

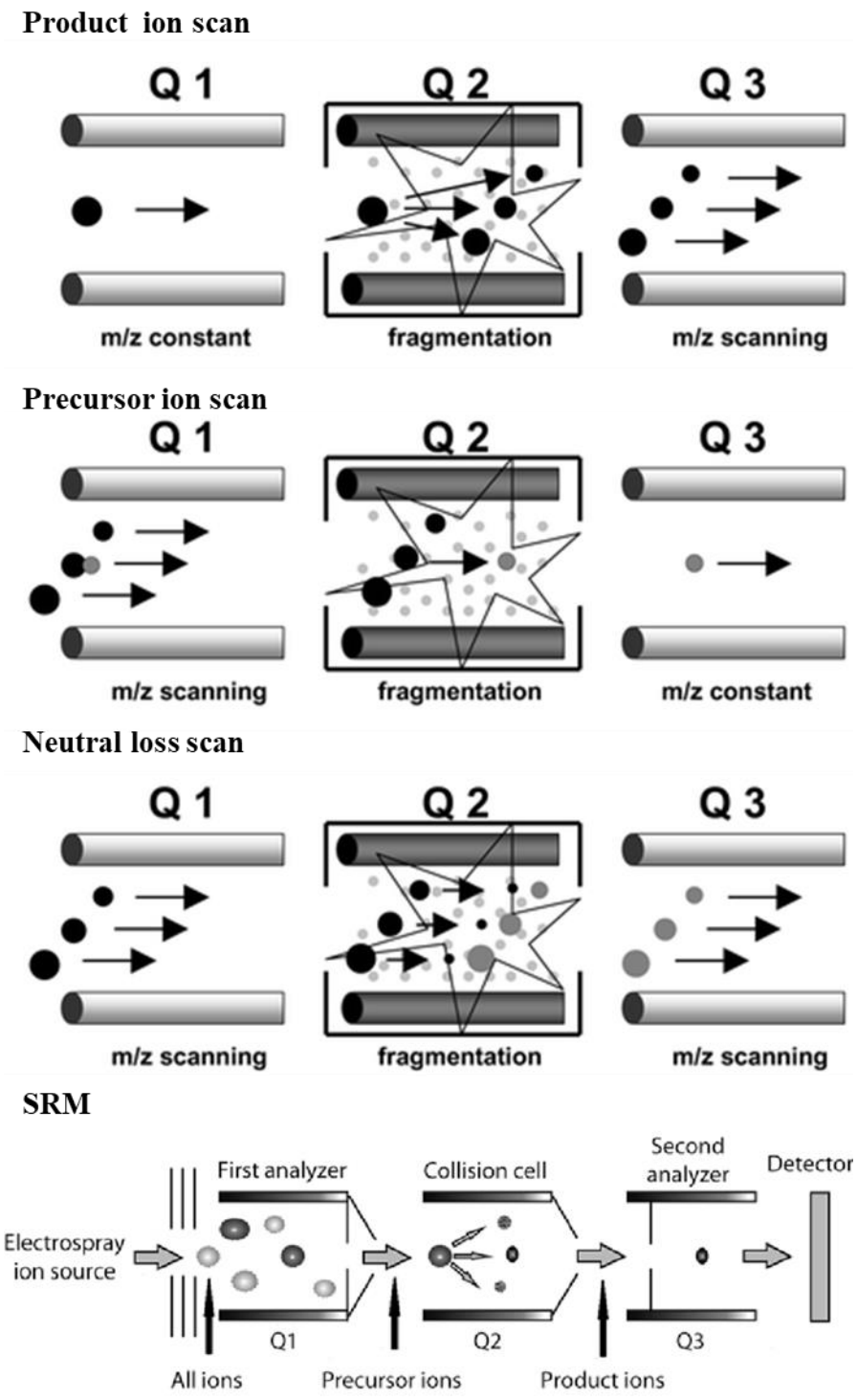


Figure 1-15. Schematic of the main detection modes of a triple quadrupole mass spectrometer. (Reprinted from ref[56, 57])

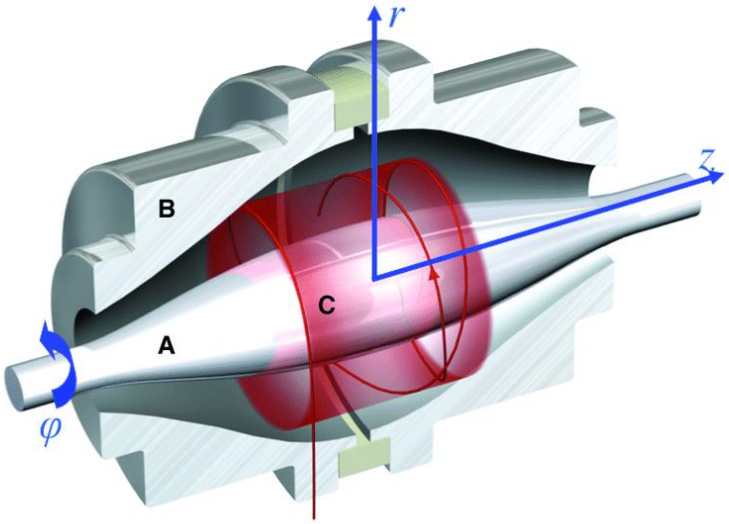


Figure 1-16. An orbitrap mass analyzer consists of three parts: (a) a central electrode, (b) an outer electrode, and (c) an insulating ceramic ring. (Reprinted from ref[59])

1.5 The objectives and outline of this thesis

This study is conducted to improve the peptide-based data storage systems. Three objectives have been established for this project: (1) To improve the encoding scheme and peptide design; (2) To explore the stability and durability of peptides; (3) To improve the protocols for liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis of data-bearing peptides.

In this Chapter, the general background of data storage and mass spectrometry have been introduced. The fundamentals of Orbitrap and QqQ-MS, which were the two major instruments used in this study, have also been introduced.

In Chapter 2, the encoding scheme and peptide design were improved based on the previous study. The peptides were designed with varying lengths and better hydrophilicity to be more suitable for higher capacity data storage. In addition, more amino acids were incorporated into the encoding scheme.

In Chapter 3, the kinetic stability of peptides for data storage was explored and compared with that of DNA. QqQ-MS was used for quantitative analysis. Additionally, the effects of different drying methods and storage methods were discussed. The half-life of peptide was deduced according to the Arrhenius equation.

In Chapter 4, to attempt to improve peptide separation, two-dimensional liquid chromatography coupled with tandem mass spectrometry (2D-LC-MS/MS) was applied for peptide sequencing. A peptide mixture containing 40 peptides translated from a text file was separated by three LC methods: UPLC, nano-LC, and 2D-LC.

Separation effect and data recovery were discussed to evaluate the suitable method for higher capacity data retrieval.

In Chapter 5, the research findings were summarized, and the prospects were discussed.

Chapter 2: Peptide design for data storage

2.1 Introduction

2.1.1 Previous study

The process of conversion from binary data to peptide sequences is called transcoding. The data is transcoded into peptide sequences following a designed coding scheme.[34] Each peptide sequence consisted of four parts: the sequence containing the information, the address sequence, the error correction scheme, and the ends of the peptides.[34]

Since the peptides should not be designed too long to guarantee efficient peptide synthesis and sequencing, the peptide sequences were designed as strings with a certain length. Although shorter-length peptides are easier to synthesize and sequence, longer-length peptides could bear more data. Therefore, the peptides were designed to contain 18 amino acids for the balance of these factors.

Amino acids used for data storage were selected from a pool of 20 natural amino acids due to cost considerations. Cysteine (C) and methionine (M) were excluded because they are prone to disulfide bridge formation and oxidation. Proline (P) was excluded because proline-containing peptides are relatively difficult to be synthesized.[60] glutamine (Q) and Asparagine (N) are excluded since they may lose amine in the process of fragmentation in MS/MS.[43] Histidine (H), arginine (R), and Lysine (K) were excluded because they could cause a serious decrease in peak intensity if placed at the N-terminal or in the middle.[43, 61] In addition, isoleucine (I) was excluded since it could not be distinguished from leucine (L). Therefore, a total of 11 amino acids were remaining. To achieve one-to-one mapping of amino acids to bit sequences, only 8 amino acids, which were alanine (A), glutamic acid (E), leucine (L), phenylalanine (F),

serine (S), threonine (T), tyrosine (Y), and valine (V), were chosen for data encoding. Each amino acid was mapped to three bits of binary data, as shown in Table 2-1. These 8 amino acids were chosen to balance the hydrophilicity of peptides.

Table 2-1. The mapping of amino acids to bit sequences.

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

The address code and error correction code are essential for data encoding and data retrieving. To retrieve the binary data, the sequence containing the information needs to be found and arranged together in a specific order. The address sequence is the indicator used to place each peptide in its original order during data retrieving. During data encoding, each peptide will be assigned a number as the address code. The number will be translated into a binary number, and then mapped to amino acids.

Additionally, in the process of information transmission, there is always a chance that the retrieved data will contain errors.[62] Therefore, error correction schemes were incorporated during encoding to ensure the data could still be fully retrieved as long as the received data meet the error rate requirement. Two error correction codes were applied in two datasets: low-density parity-check (LDPC) code and Reed-Solomon (RS) code.[34] The LDPC code was developed by Gallager in 1962.[63] It is a forward error correction code. There were two parts: order-checking codes and redundant codes.[34] The order-checking codes were used to distinguish the N-terminal and C-terminal of each peptide, and the redundant codes were used to safeguard data and promote

consistency. This error correction scheme was used on the first dataset, as shown in Figure 2-1. Based on the result, we made some assumptions about possible errors. Under these assumptions, we improved the error correction scheme based on the RS code.[34] The RS code was developed by Reed and Solomon in 1960.[64] It is a block-based error correction code with a wide range of applications in digital communications and data storage. There were also two parts in the error correction scheme: order-checking codes and RS codes. The order-checking codes remained the same. The RS codes were divided into four parts according to the different error rates in different regions of peptides. This error correction code allowed the 10% error rate overall.

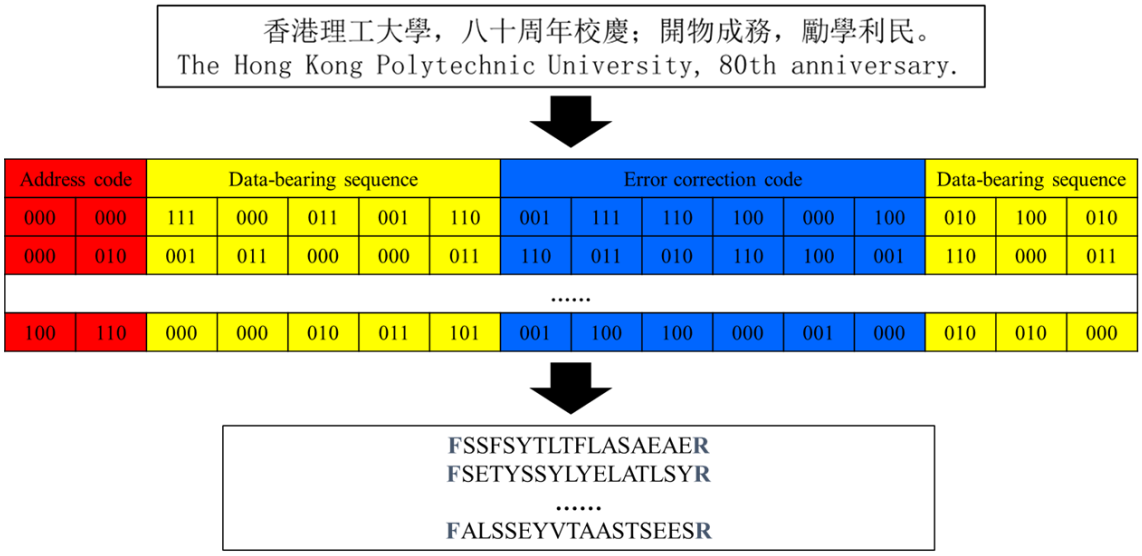


Figure 2-1. The process of encoding data into peptide sequences.

Finally, since the amino acid at N-terminal and C-terminal were prone to missing during fragmentation, F and R were fixed at the N-terminal and C-terminal of the peptides respectively and stored no data.[34] R was fixed at C-terminal because C-terminal arginine could increase the signal intensity of the y-ion series.[61] And to balance the hydrophobicity, F was fixed at N-terminal.

Following the specific pattern, the MS/MS spectra could be analyzed and decoded automatically with in-house software developed in the previous study instead of manual interpretation.

Two datasets, dataset A (40 peptides) and dataset B (511 peptides) were designed for a primary test. Dataset A was an 848-bit long text file(Figure 2-2)[34]. Dataset B was a 13,752-bit long MIDI file.[34]

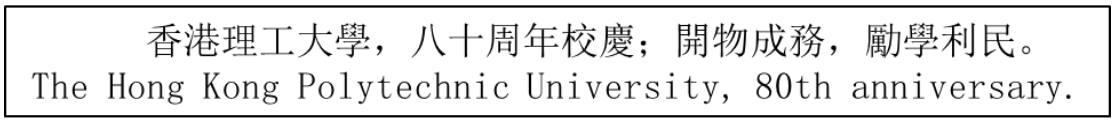


Figure 2-2. The message of dataset A. (Reprinted from ref[34])

The result showed that dataset A could be fully retrieved even without error correction code, while only 93.7% of the amino acids in dataset B were correctly sequenced without error correction code.[34] Even though, with the error correction code the incorrect or lost amino acids were acceptable at less than 10%, the results demonstrated the limitation of the current peptide design.[34] If the address code was corrupted during a sequencing run or two peptides could not be distinguished from each other with the current LC method, the entire peptide would be missed. It was the major reason causing the missing information in dataset B. This problem will become more serious as the amount of stored data becomes much larger. Therefore, the encoding scheme certainly should be improved.[34]

2.1.2 Current study

In this study, the previous peptide design and encoding scheme have been improved to achieve higher throughput and better retrieval coverage. Based on the previous result, the encoding scheme and peptide design were improved in three aspects: (1) protecting the address code; (2) varying the peptide mass; (3) balancing the hydrophilicity. First, the address code was moved from close to the N-terminal to the middle of the sequence. Second, since some peptides shared the same mass or same retention time, distinguishment using LC-MS/MS can be challenging. Therefore, the length of peptides was designed to vary between 17 to 21 amino acids and was cross-checked by their address code. Third, histidine (H) instead of F was fixed at N-terminal to balance the hydrophilicity, because some peptides were very hydrophobic which made it hard to be sequenced or synthesized. Moreover, glycine (G) has been incorporated into the encoding scheme. The mapping table was shown in Table 2-2. During the first pass of encoding, L would not be incorporated into the peptide sequence. Then, in the second pass of encoding, L would be used to replace G in the peptides which have redundant masses with others.

Table 2-2. The mapping of amino acids to bit sequences in this study.

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

The ultimate goal of this study was to encode and retrieve a dataset with the use of approximately 5,000 peptides. Since the cost of synthesizing about 5,000 peptides is very high, the validation was performed step by step. First, the new encoding scheme was evaluated by statistical calculations, including the number of peptides with

redundant masses and the coding efficiency. Then, a new dataset, in a total of 40 peptides was synthesized and sequenced to evaluate the new encoding scheme. In the future, about 5,000 peptides will be synthesized to store data.

2.2 Methods

2.2.1 Dataset

Two datasets are chosen for this study. Dataset C is a 96,224-bit long PNG formatted picture for the PolyU logo (Figure 2-3). Dataset D is the updated version of dataset A, a 1088-bit long text for the name and the motto of The Hong Kong Polytechnic University in both Chinese and English.

a)

b)

香港理工大学：开物成务，励学利民。
The Hong Kong Polytechnic University: To learn and to apply, for the benefit of mankind.

Figure 2-3. (a) The picture of dataset C and (b) the message of dataset D.

2.2.2 New encoding scheme

The transcoding from binary data to peptide sequence was shown in Figure 2-4. The address code was moved to the middle of the sequence. The error correction code including order-checking codes and RS codes was hidden within the data bear sequence. Following the new encoding scheme, dataset D was transcoded into 40 peptides and dataset C was transcoded into 4,095 peptides.

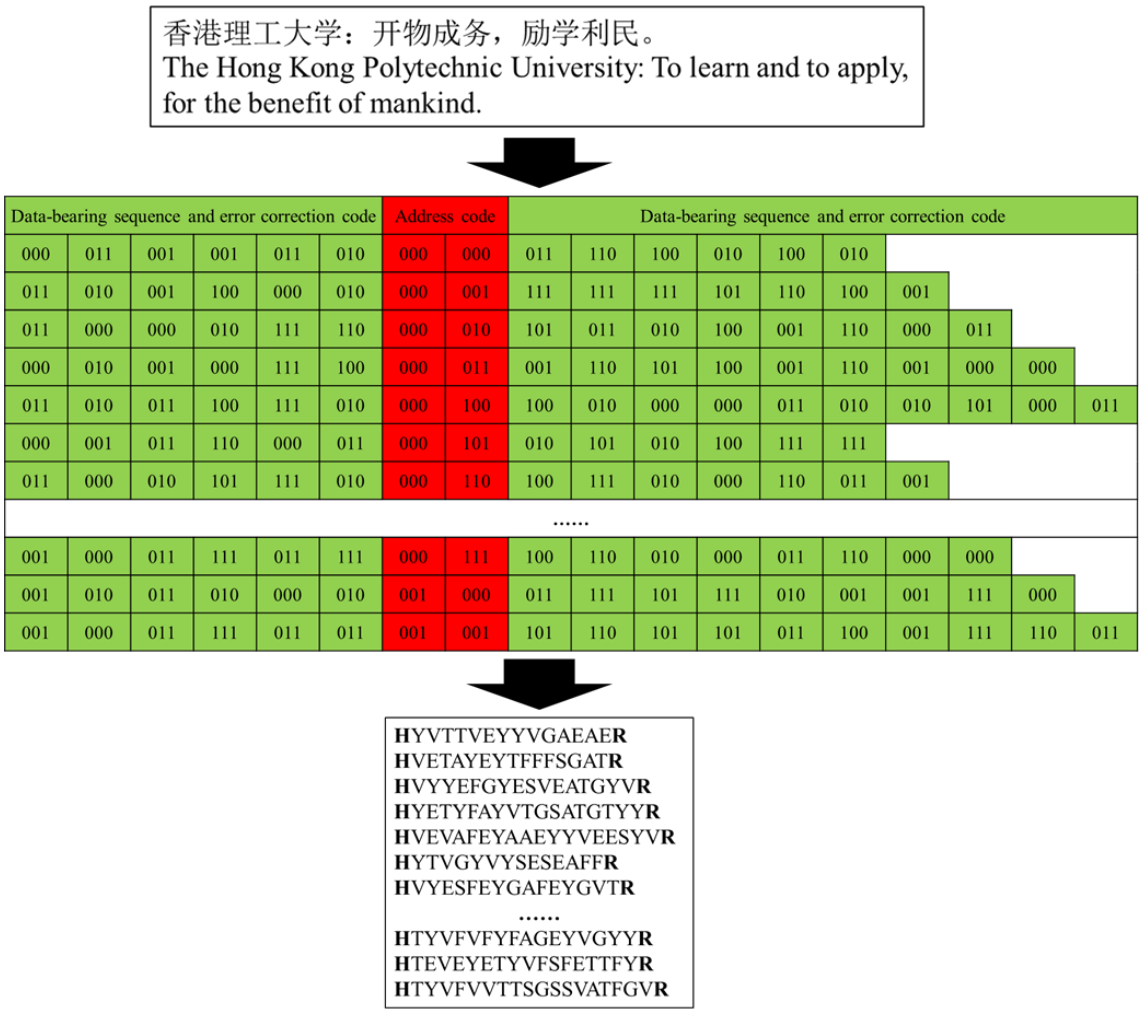


Figure 2-4. Structure of sequences in dataset D.

2.2.3 Materials and chemicals

For the primary test, we have generated 40 peptides transcoded from dataset D following the new encoding scheme. The peptides used in this study were purchased from Synpeptide Co. Ltd. (Shanghai, China). The peptide lengths ranged from 16 to 20 amino acids without any modification. The sequences of the peptides are listed in Table A1 of the Appendix section. Formic acid (FA) was purchased from VWR LLC. (France). HPLC grade acetonitrile (ACN) was purchased from Duksan Inc. (South Korea). Water was purified with the MilliQ Direct Laboratory Water Purification system.

2.2.4 Sample preparation

The peptides synthesized were dissolved with dimethyl sulfoxide (DMSO) (10 µg/mL) and diluted with 50% ACN with 0.2% FA to 5 nmol/ml for LC-MS/MS analysis. A blank sample containing 50% ACN with 0.2% FA was prepared at the same time. The samples were freshly prepared before analysis.

2.2.5 Instrumental setup

This experiment was performed with an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher, MA, USA) coupled with a two-dimensional ultra-high pressure liquid chromatography (2D UPLC) (UltiMate 3000 DGLC) system (Thermo Fisher, MA, USA). The scan range was set between 400-1500 *m/z*. The ionspray voltage (IS) was set at 2.3 kV for positive ions. The ion transfer tube temperature was set at 280℃. The fragmentation method was HCD with stepped collision energy: 27, 30, and 34. With 10 µL of each sample solution injected into the UPLC system, the peptide mixtures were separated with a C18 column (Thermo Fisher Hypersil GOLD AQ, 100×2.1 mm, 1.9 µm particle size). The LC parameters were summarized in Table 2-3.

Table 2-3. LC parameters for peptide sequencing.

	H at N-terminal	**F at N-terminal**
Solvent A	0.2% FA in water	0.2% FA in water
Solvent B	0.2% FA in ACN	0.2% FA in ACN
Flow rate	0.3 mL/min	0.3 mL/min
Temperature	55℃	55℃
Gradient	0-27 min: 5-50% B	0-3 min: 5% B
	27-27.1 min: 50-95% B	3-27 min: 5-50% B
	27.1-32 min: 95% B	27-30 min: 50-95% B
	32-32.1 min: 95-5% B	30-35 min: 95% B
	32.1-40 min: 5% B	35-35.1 min: 95-5% B
		35.1-40 min: 5% B

2.3 Results and discussion

Two groups of peptides sequences have been generated by using dataset C with the old and new encoding schemes. The distribution of mass of the peptides was analyzed as shown in Figure 2-5. The result showed that the mass of peptides encoded with the new encoding scheme was more dispersed than that encoded with the old encoding scheme. The number of peptides with redundant masses had also been counted for analysis. The redundant masses were defined as that the mass difference between two peptides was less than 25 parts-per-million (ppm) as follows:

$$\text{mass difference} = \frac{M_{(\text{peptide A})} - M_{(\text{peptide B})}}{M_{(\text{peptide B})}} \times 10^6 \qquad (2\text{-}1)$$

It showed that the new encoding scheme could highly reduce the number of peptides with redundant masses, which would make the peptide-based data storage system achieve higher capacity. In total, with the new encoding scheme, there were 67.06% of peptides with redundant masses less than 25 ppm. While with the old encoding scheme, there were 96.21% of peptides with redundant masses less than 25 ppm. However, there were still too many peptides with redundant masses, which might cause peptides missing if these peptides were co-eluted. The result implied that the LC method should be improved in further study.

The net information density of datasets A, B, and C have also been calculated for reference. The data size, number of peptides, peptide length, and net information density of the three datasets were summarized in Table 2-4. Normally, coding efficiency was higher in a larger amount of data, as dataset A and dataset B showed. However, with more peptides in the mixture, the address code became longer which cause a waste of coding efficiency. The result possibly indicates the existence of a

critical threshold. With more peptides stored in one mixture, the coding efficiency would be lower.

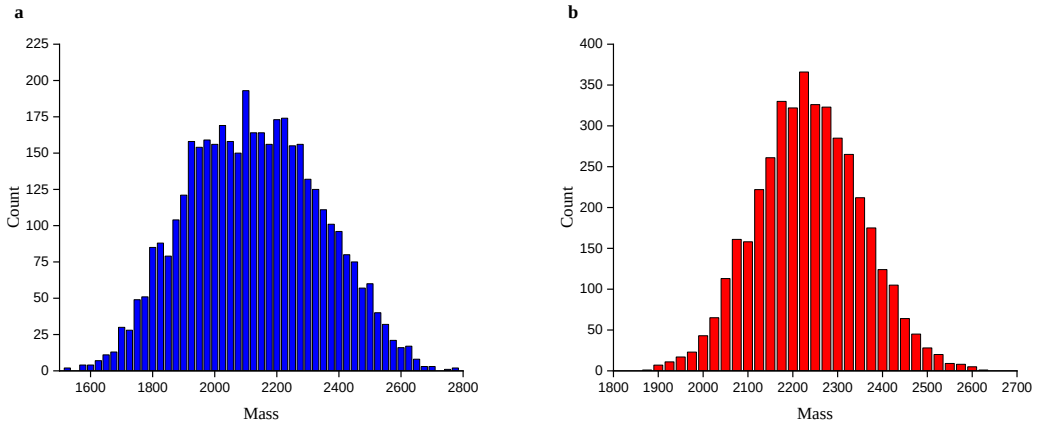


Figure 2-5. The distribution of masses of the peptides encoded with (a) new encoding scheme and (b) old encoding scheme.

Table 2-4. The statistical data of three datasets.

	Length (bits)	Number of peptides	Peptide length	Coding efficiency (%)
Dataset A	848	40	18	39.3
Dataset B	13,752	511	18	49.8
Dataset C	96,224	4,095	17-21	41.2

Then the 40 peptides transcoded from dataset D with the new encoding scheme were synthesized and sequenced. The chromatograms for analysis of the peptide mixture were shown in Figure 2-6. The sequencing result was shown in Table 2-5, and only 70.5% (451/640) of amino acids have been correctly sequenced. Since only 10% of sequence error or missing was allowed based on the error correction code, the data couldn't be fully retrieved in this case. Among all of the peptides, 10 peptides were

fully correctly sequenced, and 21 peptides were sequenced with less than 3 amino acids missing. However, 5 peptides were completely missing and could not be sequenced. After checking, we found three of them, no. 13, 19, and 39 could not be sequenced because the signal intensity of the peaks of MS/MS spectra was too low, at less than 1×10^6. Since the in-house software was designed to distinguish the b-ion series and the y-ion series from noise according to signal intensity, the peptide could not be sequenced if the signal intensity was too low or incomplete. The remaining two peptides, No. 17 and 37, could not be found in MS spectra. After we changed the LC method, these two peptides could be found and sequenced. The reason was that these two peptides were not selected as valid peptides by the software due to lower intensity compared with the co-eluted peptides. However, the predominant reason for data missing was that the signal of the y-ion with low *m/z* was bad. There was too much noise, and the signal intensity was relatively low, as shown in Figure 2-7. The result was consistent with a study that suggested that N-terminal fixed H could cause a sharp decrease in peak intensity.[61] This phenomenon didn't appear when F was fixed at N-terminal. Considering that a new amino acid G has been incorporated into the peptide, the peptide is sufficiently hydrophilic even though F is fixed at the N-terminus. Therefore, F instead of H was fixed at N-terminal to get better MS/MS spectra.

Then, the 40 peptides with F fixed at N-terminal transcoded from dataset D with the new encoding scheme were synthesized and sequenced. The peptides with F fixed at N-terminal were sequenced in the same method as the peptides with H fixed at N-terminal, apart from the LC method, because the hydrophilicity of peptides has changed. The result was shown in Table 2-5. The result showed that 95.63% (612/640) of amino acids have been correctly sequenced. The data could be fully retrieved with error

correction code. Among all of the peptides, 33 peptides were fully correctly sequenced, and 4 peptides were sequenced with less than 3 amino acids missing. The other 6 peptides were sequenced with 5 to 7 amino acids missing. All the peptides could be sequenced. Most of the amino acids missing were caused by missed fragmentation, e.g., the order of the second and third amino acids and the order of the third and second last amino acids in a peptide may be exchanged. It is a common phenomenon caused by unsuccessful fragmentation of the first three and last three amino acids in the MS/MS spectra.

Table 2-5. The number of peptides with 0-18 correct amino acids after sequencing (H at N-terminal).

No. of correct amino acids	No. of peptides (H at N-terminal)	No. of peptides (F at N-terminal)
0	5	0
1	0	0
2	0	0
3	0	0
4	1	0
5	0	0
6	0	0
7	0	0
8	1	0
9	1	0
10	2	0
11	4	2
12	7	1
13	4	1
14	5	10
15	4	8
16	3	6
17	2	7
18	1	5

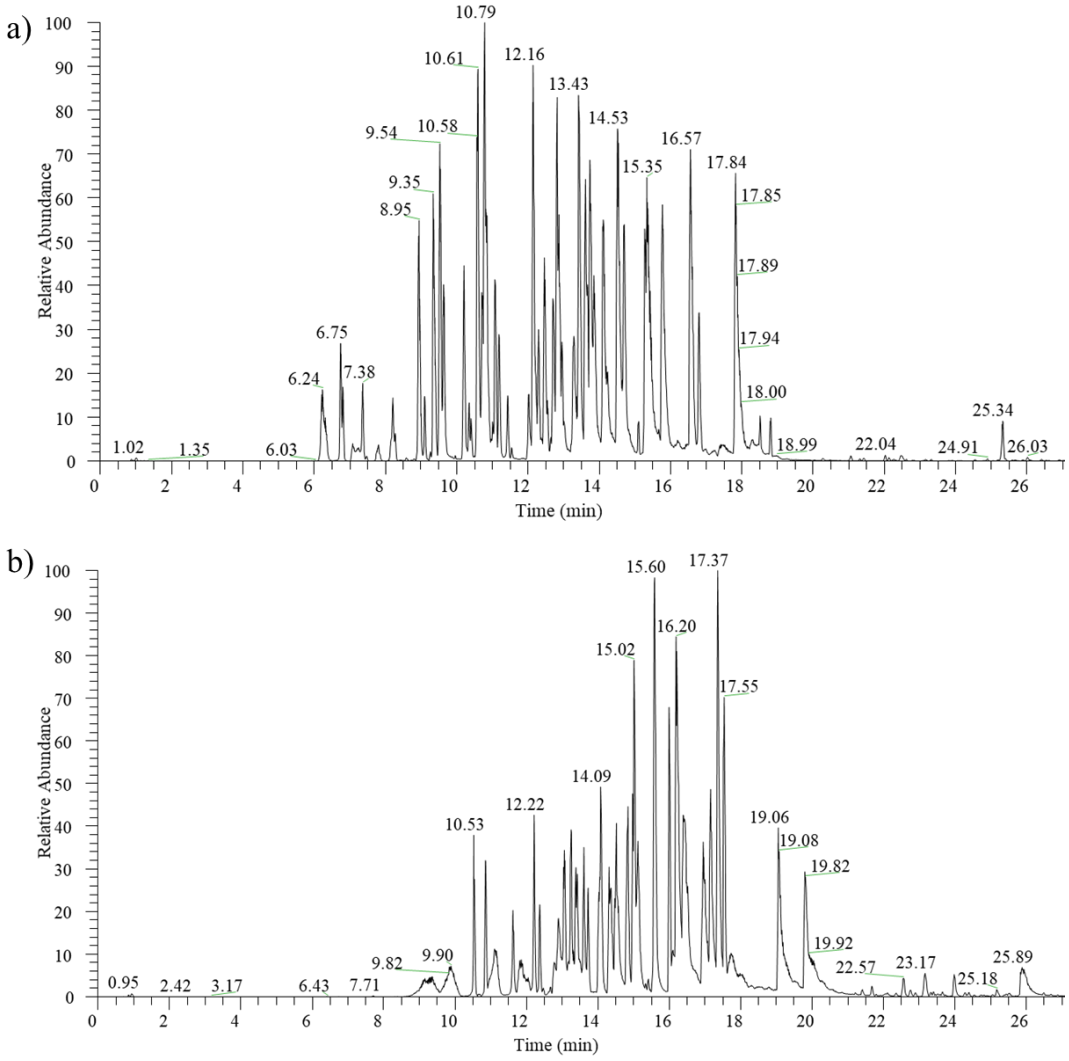


Figure 2-6. The chromatograms for analysis of two peptide mixtures encoding dataset D: (a) H at N-terminal, (b) F at N-terminal.

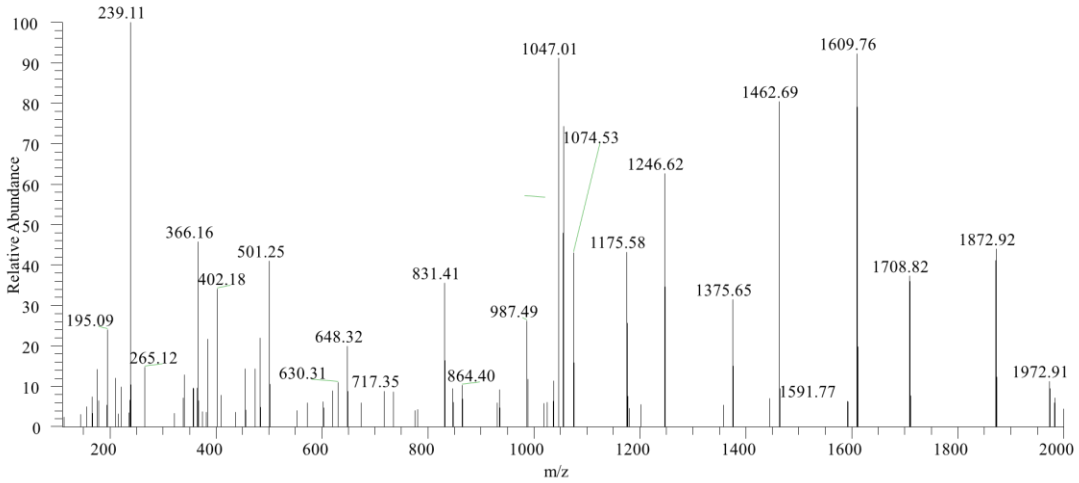


Figure 2-7. The MS/MS spectrum of peptide No. 10.

2.4 Conclusions

In this study, the encoding scheme and peptide design of the peptide-based data storage system have been improved for LC-MS/MS analysis. The result showed that the peptides generated with the new encoding scheme could significantly reduce the occurrence of redundant masses compared with the old encoding scheme. However, the net information density of dataset C was less than that of dataset B because more amino acids were used to encode address code which reduces coding efficiency. It suggested that a threshold might exist, and the practical way to store data with peptides is to divide the peptides into multiple mixtures, with each mixture containing several hundred to several thousand peptides, instead of encoding the data with one mixture.

Besides, MS/MS spectra and sequencing results have been tested in this study. Although the peptides were more hydrophilic with H fixed at N-terminal, only 70.5% of data was retrieved without error correction code. While with F fixed at N-terminal, 95.63% of the amino acids were correctly sequenced. With the error correction code, which allowed a maximum of 10% of amino acids misread or missing, the original information could be fully retrieved. Most of the amino acids missing were caused by missed fragmentation.

In future, we will encode and decode dataset C with 4,095 peptides. According to the result, there will still be 2,746 peptides with redundant masses. Therefore, the LC method will be further optimized to achieve higher capacity.

Chapter 3: Peptide stability

3.1 Introduction

In the past decades, peptides have gained a wide range of applications in medicine and biotechnology.[65] Therefore, lots of research have been conducted on the stability of peptides. However, most of the research was focused on the kinetic stability of peptides *in vivo* and the aqueous phase.[66] Therefore, our knowledge about its solid-state stability is insufficient.

There is no doubt that the peptides used to bear data should be formulated in the solid state, because it not only can achieve higher data density but also can provide longer-term storage. The previous study suggested that freeze-drying could successfully suppress the hydrolysis of the amide bond and generation of isomers.[67] However, even in the solid state, reactions can occur and lead to degradation. The major reactions include deamidation, peptide bond cleavage, and oxidation.[68, 69]

3.1.1 Chemical pathways of peptide degradation

Deamidation

Although the concept of pH has no meaning in the solid state, the pH of the buffer solution determines the extent of ionization of the peptides both in solution and in solid state. Chemical instability in the solid state due to deamidation has been found related to the pH of the buffer solutions.[70]

There have been studies on the stability and mechanism of degradation of peptides, one of them being on the Asp-hexapeptide (Val-Tyr-Pro-Asp-Gly-Ala) in solid formulations freeze-dried from acidic solutions ranging from pH 3.5-8.0 (Figure 3-1).[67]

While under acidic conditions (pH 3.5 and 5.0), the major product observed was Asu-hexapeptide. Hydrolysis of the Asp-Gly amide bond has also occurred but only trace amounts of tetrapeptide and dipeptide have been detected.[67] While under basic conditions (pH 8.0), the isoAsp-hexapeptide was the dominant degradation product observed.[67] Besides, the result showed that the degradation pathways of Asp-hexapeptide in solid state appear to be similar to that in solution.[67]

Similar to the degradation of Asp-hexapeptide, it has also been proved that the solid-state degradation of Asn-hexapeptide (Val-Tyr-Pro-Asn-Gly-Ala) was dependent on the pH of the bulk solution (Figure 3-1).[70] At a lower pH value (pH 3), Asn-hexapeptide was deamidation to generate the Asu-hexapeptide via hydrolysis of the Asn side chain.[70]

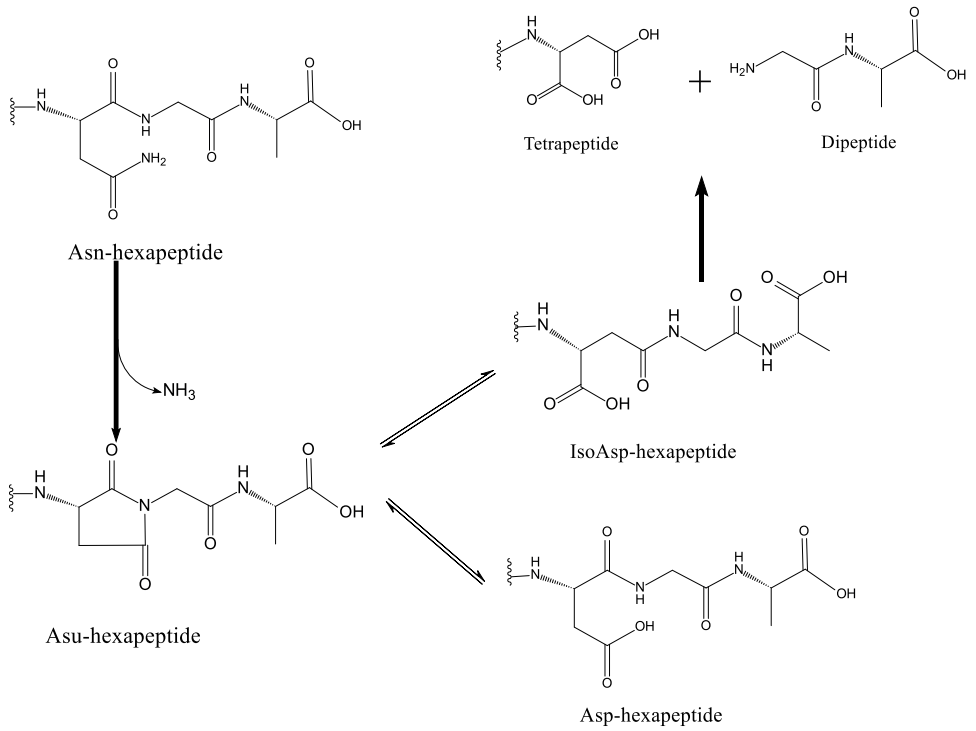


Figure 3-1. Degradation pathway of Asu-hexapeptide and Asp-hexapeptide. (Reprinted from ref[67, 70])

Peptide bond cleavage

Cleavage of the peptide bond has also been found in the solid state, as shown in Figure 3-1. A previous study showed that the C-terminal serine residue on the B-chain (Trp_{28}-Ser_{29}-COOH) of freeze-dried human relaxin could undergo hydrolytic cleavage while storage at 40°C.[71] The proposed mechanism for this reaction was that the Trp-Ser bond was hydrolysis via a cyclic intermediate, which initiated by the reaction of the Ser hydroxyl group with glucose.[68, 71] Therefore, the cleavage of the peptide bond could be observed in the freeze-dried sample.

Oxidation

Another major degradation pathway for peptides is the oxidation of peptide residues (Figure 3-2). The side chains of Cys, His, Met, Tyr, and Trp residues in peptides are potential oxidation sites.[72]

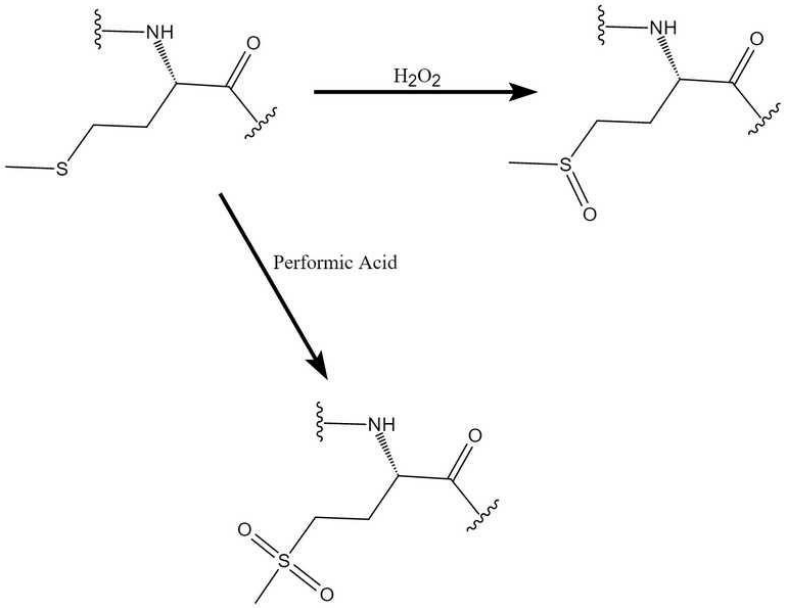


Figure 3-2. Pathway of methionine oxidation. (Reprinted from ref[72])

A study conducted with human insulin-like growth factor I (hIGF-I) showed that both oxygen content and light exposure affected the oxidation rate.[73] The oxidation rate increased by a factor of 30 while exposure to light, which suggested that molecular oxygen and photooxidation might be involved in the generation of radicals.[73] Although oxygen content was related to oxidation rate, oxidation still could be observed even though the samples were sealed in vacuum or nitrogen atmosphere.[74]

An early study about methionine oxidation of bovine serum albumin in the solid state showed that crystallization was another important factor in protein oxidation.[75] Three peptides were selected in the study, and both amorphous and crystalline materials were produced for the investigation. The result showed that nearly no oxidation was observed in the crystalline material, and amorphous material degrade much faster with each peptide.[75]

Solid-state oxidation can be minimized with certain excipients. It was suggested that some excipients could form hydrogen bonds with the protein surface to preserve the native conformation of the protein, and then bury the amino residues from exposure to oxidation. Therefore, some excipients could be used to inhibit protein oxidation. A previous study has shown that elastase could be stabilized with 10% dextran. The dextran could keep tryptophan(s) in hydrophobic regions that were unavailable for oxidation.[76] However, it was also found that some excipients, e.g., ascorbic acid can, surprisingly, promote the oxidation of methionine.[76, 77]

3.1.2 Peptides storage methods

Solution peptide and protein formulations are susceptible to physical and chemical degradation.[78] To achieve peptide formulations with long-term stability, peptides should be formulated in the solid state, because chemical reactions occurring in the liquid state get drastically reduced in solid form.[79] Various methods for manufacturing solid peptide formulations have been developed, including film drying, freeze-drying, spray coating, spray drying, spray freeze-drying, supercritical fluid drying, and vacuum drying. Among these methods, freeze-drying, also called lyophilization, was widely used due to its advantages over other drying methods.[79]

A freeze-drying cycle typically includes three stages: freezing, primary drying, and secondary drying.[80] During the freezing stage, ice crystals start forming. During the process, most of the water was separated from a matrix of glassy into ice crystals.[81, 82] During the primary drying stage, the formed crystalline ice earlier is removed by sublimation. The chamber pressure drops to well below the vapor pressure of ice resulting in sublimation. Additionally, the temperature of the shelf increases to supply the heat needed.[82] After the primary drying stage, there may still be about 15 - 20% of unfrozen water left in the product. The remained water is then removed in the secondary drying stage. The required low moisture is achieved with the help of higher temperature and lower pressure.[82] Although most freeze-drying cycles consist of the abovementioned steps, two freeze-drying cycles can vary significantly from each other. Because these three steps may be optimized based on the properties of the products.

It has been proved that freezing and drying may induce protein denaturation during freeze-drying.[83] To protect a protein from freezing and/or dehydration denaturation, an excipient(s) may be used.[84] Excipients giving protection to protein against freezing

stress are called cryoprotectants, whereas those giving protection against drying stress are called lyoprotectants. [85] There are various types of excipients, e.g., surfactants, non-aqueous solvents, sugars, polymers, inorganic salts, metal ions polyols, and some amino acids, which can offer different effects during freeze-drying and the following storage.[84, 86] Therefore, to achieve better stability, two or more excipients may have to be used to protect proteins from denaturation during freeze-drying.[84]

One major mechanism of protein stabilization by excipients is the formation of an amorphous glass during freeze-drying.[87] A study showed that it was the extreme viscosity at the amorphous glassy state, which increased protein stability by slowing down interconversion of conformational substates and conformational relaxation of a protein.[88] Besides, a study conducted with insulin showed that amorphous insulin was far more stable than crystalline insulin against deamidation and dimer formation at different water contents.[84, 89] However, some studies showed the opposite results, e.g., a research about methionine oxidation in the solid state showed that the amorphous peptides degraded much faster compare with crystalline peptides, even nearly no oxidation occurred in the crystalline form.[75] Another stabilization mechanism was the water replacement hypothesis.[84, 90] According to this hypothesis, these excipients are serving as water substitutes to prevent intra- or inter-protein hydrogen bonding, which preserves the native structures of proteins.[91]

Additionally, peptide modification may be another method to further improve the stability of the peptides. A study about the kinetics of solid-state stability of seven derivatives of 3,5-disubstituted tetrahydro-2H-1,3,5-thiadiazine-2-thione (THTT) of glycine has shown that in the case of N-3 alkyl substituents, the methyl group decreased

the half-life at 25°C to 0.2 month, whereas chain elongation (ethyl, propyl) was accompanied by increasing the half-life at 25°C to 25 and 40 months.[92]

The objective of this study was to verify the stability and durability of peptides for data storage. In this study, we explored the kinetic stability of peptides in solid state. Liquid chromatography-multiple reaction monitoring-MS (LC-MRM-MS) was used for quantification analysis of peptide degradation. Accelerated aging experiments were performed to measure peptide decay kinetics. Since several previous studies suggested that the degradation of peptides in solid state could be demonstrated with first-order kinetics, the stability of peptides was calculated and evaluated following first-order kinetics and the Arrhenius equation. The stability of peptides with different drying methods and storage methods was also explored in this study. The half-life of peptide was deduced according to the Arrhenius equation and compared with literature data on DNA stability.

3.2 Methods

3.2.1 Materials and chemicals

The peptides used in this study were purchased from Synpeptide Co. Ltd. (Shanghai, China). All the peptides were 18 amino acids in length without any modification. One of the peptides, FE12, which was used as an internal standard (IS), was labeled with ^{13}C and ^{15}N at C-terminal. The sequence and purity of the peptides are listed in Table 3-1. The parafilm (4 in. × 125 ft.) was purchased from Bemis Inc. (Zurich, Switzerland). Formic acid was purchased from VWR LLC. (France). HPLC grade acetonitrile was purchased from Duksan Inc. (South Korea). Water was purified with the MilliQ Direct Laboratory Water Purification system. D- (+)-Trehalose dihydrate was purchased from TIN HANG Ltd. (Hong Kong, China). The 1.5 mL cryogenic microtubes with deep caps were purchased from Sangon Biotech Co. Ltd. (Shanghai, China).

Table 3-1. The sequence and purity of the peptides used in this study.

Peptide	Sequence	Purity (%)
FE1	HEEEEEEEEEEEEEEER	70
FE2	HSTEYETYSAFVLLAVFR	70
FE3	RALEAVSTAESLAVLFYH	70
FE4	FFYVSYFATYYVTATYYR	70
FE5	FFYVEYFAEYYVEAEYYR	70
FE6	HYFLVALSEATSVAELAR	70
FE7	HASVTLEFYVSATELYFR	70
FE8	FTALTELESEAVEAAVER	70
FE9	FTSAVYAASESEEESFER	70
FE10	FYYLSLLSTLLYAVYAVR	70
FE11	FEELALYVEEYVVYEYTR	70
FE12	FYYLSLLSTLLYAVYAVR	70

3.2.2 Sample preparation

The peptides synthesized were dissolved with dimethyl sulfoxide (10 µg/mL) and diluted with 50% ACN with 0.2% FA. Then FE1 to FE5 were mixed together as mixture A, and FE6 to FE11 were mixed together as mixture B. Both groups were divided into several samples (10 nmol of each peptide). These samples were dried with a vacuum concentrator (Labconco, MO, USA) at 20℃ for 4 hours. The dried samples were stored at different temperatures ranging from 70℃ to 90℃ with a constant temperature metal bath (Jingxin, Shanghai, China). The peptide used as IS was also dried by vacuum drying with the same method and stored at -80℃. All the samples were stored for 4 weeks. The stored samples were dissolved with DMSO and diluted with 50% ACN with 0.2% FA. Then, the peptide solutions were pre-heat at 50℃ for one minute before the analysis to make sure the peptides were fully dissolved.

To compare the effect of different storage methods, part of the samples was prepared using freeze-dry with freeze-dryer FD-1A-50 (BIOCOOL, Beijing, China). The peptide was dissolved with FA and diluted with water to 0.1% FA. Then, add trehalose to the peptide solution. Initial peptide and trehalose concentrations were 20 µM and 0.008% w/v (1/2 mass ratio). The samples were pre-frozen at -80 ℃ for 12h. Start the freeze-dryer to allow the chamber temperature to drop to -50℃ under vacuum conditions. The samples were sublimated for 24 hours at -50℃, then, the shelf temperature was increased to about 30 ℃ to exclude the residual water. The prepared samples were stored at -80 ℃ before being analyzed.

3.2.3 Instrumental setup

In this study, quantitative analysis experiments were performed with a Sciex 6500+ triple quadrupole mass spectrometer (Sciex, MA, USA). For parameter optimization, the peptide standards were directly infused into the mass spectrometer with the syringe pump. The flow rate was set as 10 μL/min. The curtain gas (CUR) flow was ion source gas, and the collision gas (CAD) flow was set as medium. The ionspray voltage (IS) was set at 5kV. The temperature of heater gas was 550℃. The ion source gas 1 and ion source gas 2 were 60 μL/min.

Table 3-2. MRM channels and parameter settings of the peptides.

Peptide	MRM Channel	DP (V)	EP (V)	CE (V)	CXP (V)
FE1	1189.5→1594.5	180	8	64	46
FE2	1066.6→1386.6	180	9	57	30
FE3	988.8→319.1	200	8	52	15
FE4	1163.4→1371.4	200	9	48	20
FE5	1226.6→1455.3	210	9	35	17
FE6	989.2→301.2	195	11	54	11
FE7	1066.7→1395.4	220	14	56	49
FE8	982.5→1431.4	220	5	48	30
FE9	1019.5→1370.4	200	6	47	31
FE10	1078.2→508.3	210	10	31	20
FE11	1158.5→1450.4	205	8	50	31
FE12	1083.2→851.6	180	9	36	19

For quantitative analysis experiments, the peptide mixtures were mixed together with IS before being injected into the mass spectrometer. The liquid chromatography-mass spectrometric multiple reaction monitoring (LC-MRM) strategy was used for

quantitative analysis. The parameter setting and MRM channels used in this study were shown in Table 3-2.

All the stored samples were dissolved with DMSO and diluted with 50% ACN with 0.2% FA. Then, 5 μM FE12, which was used as the internal standard, was mixed with the test sample with a volume ratio of 1:1. The mixed sample solutions are used for LC-MRM analysis. With 2 μL of each sample solution injected into the UPLC system, the peptide mixtures were separated with a C18 column (Agilent AdvanceBio Peptide Map, 150×2.1 mm, 2.7 μm particle size). Solvent A was water containing 0.2% FA, and solvent B was ACN containing 0.2% FA. The flow rate was set as 0.3 mL/min. The initial gradient set at 5% B from 0 to 1 min, followed by a linear increase from 5% B to 95% B at 1 to 8.5 mins. Then decreased from 95% B to 5% B at 8.5 to 8.6 mins, and remained at 5% B from 8.6 to 12 mins. For mixture A, dwell time was set as 60 ms for each channel, in a total of 0.7801 s for one cycle. For mixture B, dwell time was set as 50 ms for each channel, in a total of 0.7701 s for one cycle. During the analysis, the sample was stored at 4℃ and the column temperature remained at 55℃.

3.2.4 Calibration curves

The calibration curve for quantitative analysis of each peptide was constructed by analyzing at least six different concentrations of each peptide. While at each concentration, three sets of experimental data were obtained. The MRM chromatograms were processed with SCIEX OS software. Calibration curves were constructed by the relative peak area of targeted peptides and IS versus concentration of targeted peptides. All the sample solutions were analyzed in one batch.

3.2.5 Accuracy and precision

The accuracy and precision were determined by at least three sets of samples at high, medium, and low concentrations respectively. The concentrations of the standard samples were measured at least three times. The average value of the data was calculated as the measured value. The accuracy was defined as the closeness of the measured value and the actual value and calculated as follows:

$$\text{Accuracy}= \frac{\text{the measured value}}{\text{the actual value}} \times 100\% \quad (3\text{-}1)$$

And the relative standard deviation (R.S.D.), i.e., precision, was calculated as follows:

$$\text{R.S.D.}= \frac{\text{the standard deviation of the measured value}}{\text{the measured value}} \times 100\% \quad (3\text{-}2)$$

3.2.6 Limit-of detection (LOD) and limit-of-quantitation (LOQ)

The LOD and LOQ were used to describe the minimum concentration that is reliable. They were evaluated by comparing the intensity between signal and noise (I_{signal}/I_{noise}). The LOD and LOQ are defined as the concentration of analyte that can achieve a I_{signal}/I_{noise} value of three and ten, respectively. LODs and LOQs were evaluated by conducting at least three repeated experiments.

3.2.7 Data fitting and statistical analysis

The previous study showed that the degradation of peptides in solid state conformed to first-order kinetics, and a critical temperature threshold might exist.[67] Because the decay rate is related to multiple variables, including temperature, humidity, and oxygen content. While at higher temperatures, the decay rate is relatively independent from water and oxygen. Therefore, the analyzed peptides were stored above 70℃. The

calculated decay rate is likely to be close to the actual degradation rate in the absence of oxygen and water. According to the rate law for first-order reaction:

$$N = N_0 \times e^{-kt} \tag{3-3}$$

the kinetic decay rate (k) was calculated by:

$$\ln\left(\frac{N_0}{N}\right) = kt \tag{3-4}$$

Additionally, according to Arrhenius equation:

$$k = A \times e^{-E_a/\mathrm{R}T} \tag{3-5}$$

activation energies of each peptide were calculated by:

$$\ln k = -\frac{E_a}{\mathrm{R}} \times \frac{1}{T} + \ln A \quad (\mathrm{R}=8.314 J/Kmol) \tag{3-6}$$

3.3 Results and discussion

3.3.1 Optimization of MRM conditions

The *m/z* of molecular ions of peptides was confirmed by infusing the peptide standards directly into the mass spectrometer. The molecule ions with 2 positive charges were selected and fragmented because they were found with the highest signal intensity. The fragmented ions were then monitored. Two fragment ions of each molecule ion with stable and intensive signals were chosen for quantitative analysis of the peptide. Between the two fragment ions, one was used for quantitative measurement, the other one was used for result verification. The relative molecular mass of the peptides and the *m/z* of the molecular ions and fragment ions were listed in Table 3-3. The MS/MS result of FE1 was shown in Figure 3-3 as an example.

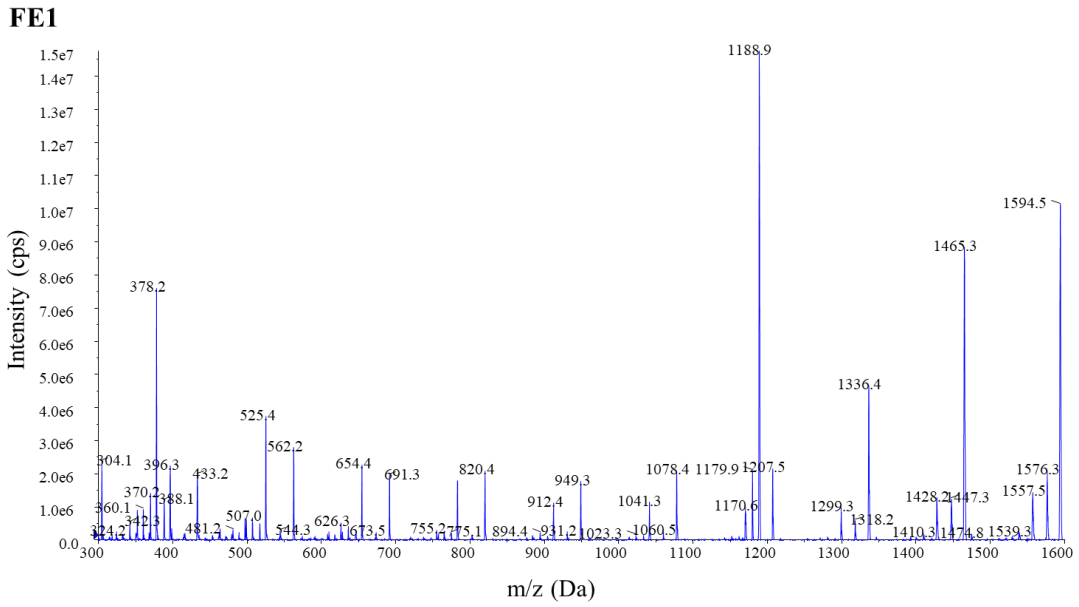


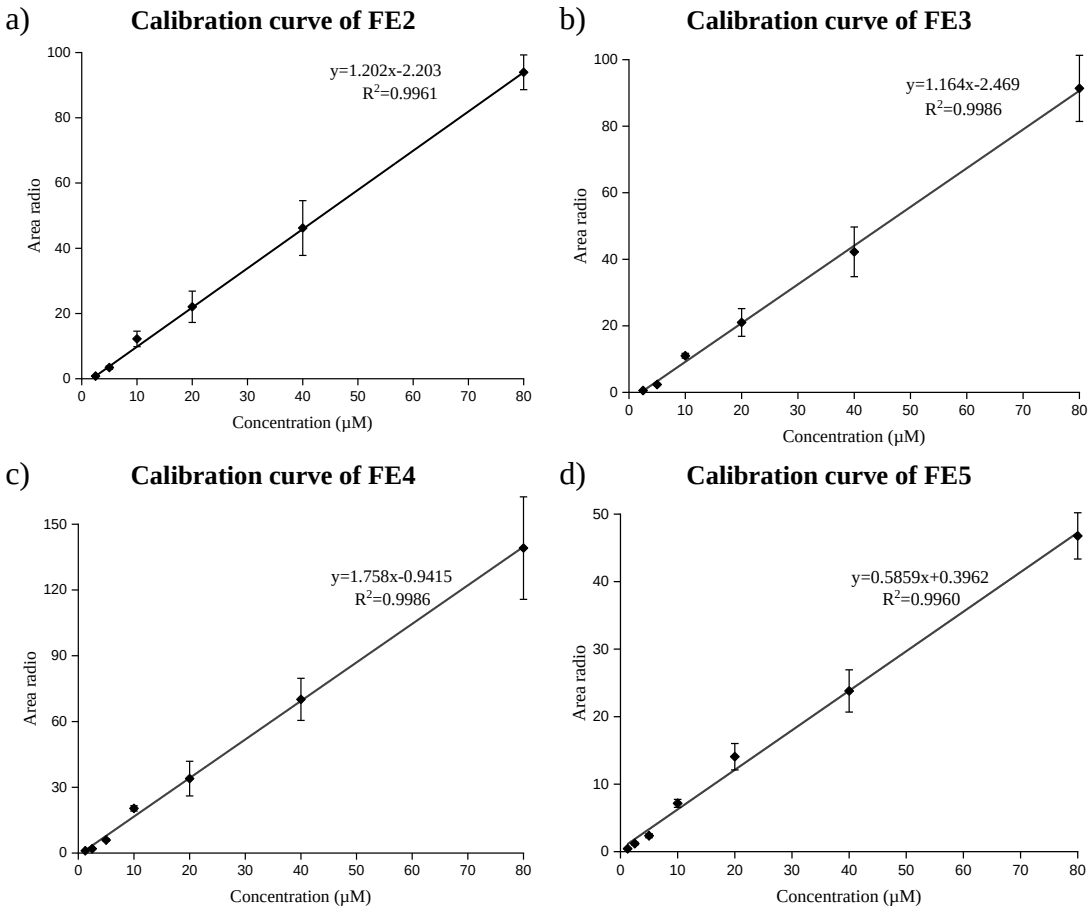
Figure 3-3. MS/MS result of FE1 as an example.

Table 3-3. Relative molecular mass, molecular ions, and fragment ions of targeted peptides in this study.

Peptide	Molecular mass (Da)	Molecular ion (*m/z*)	Fragmented ion (*m/z*)	
			Qualitative analysis	Quantitative analysis
FE1	2377.2	1189.5	1465.3	1594.5
FE2	2133.3	1066.6	1285.6	1386.6
FE3	1977.2	988.8	466.3	319.1
FE4	2325.6	1163.4	501.2	1371.4
FE5	2451.6	1226.6	501.5	1455.3
FE6	1977.2	989.2	1416.5	301.2
FE7	2133.3	1066.7	395.2	1395.4
FE8	1965.1	982.5	403.3	1431.4
FE9	2039.0	1019.5	1299.4	1370.4
FE10	2155.5	1078.2	841.5	508.3
FE11	2315.5	1158.5	1350.5	1450.4
FE12	2166.5	1083.2	518.4	851.6

3.3.2 Quantitation of targeted peptides

The calibration curves for quantitative analysis were constructed by analyzing at least six different concentrations of each peptide. While at each concentration, three sets of experimental data were obtained. 11 peptides were divided into two mixtures, mixture A and mixture B, to simplify the method and save time. 5 μM IS was mixed with these mixtures with a volume ratio of 1:1. Calibration curves, which were constructed by the relative peak area of targeted peptides and IS versus concentration of targeted peptides, were shown in Figure 3-4. The linear range and linearity (in term of R^2) of each peptide were shown in Table 3-4.



(To be continued)

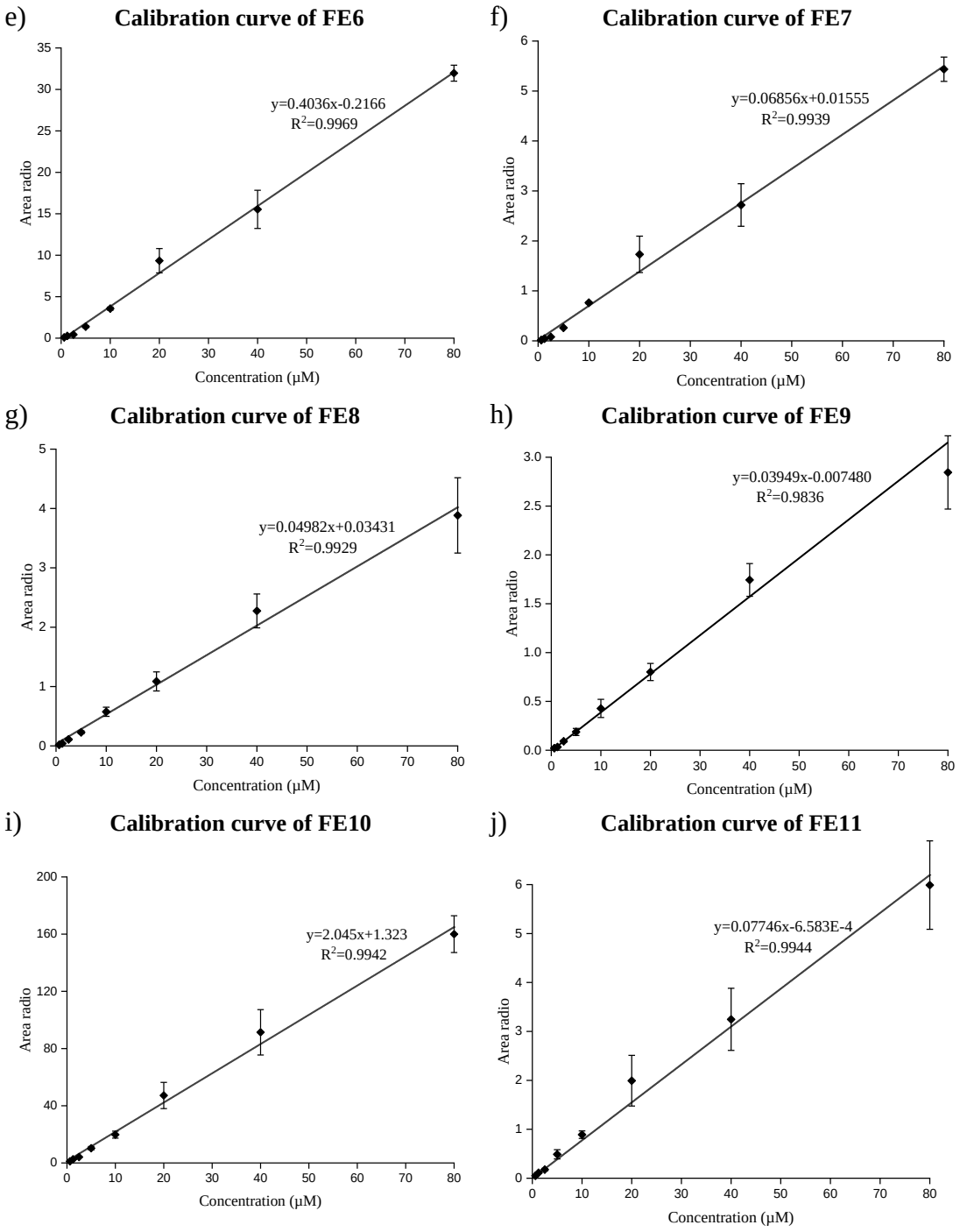


Figure 3-4. Calibration curves for the quantitative analysis of targeted peptides: (a) FE2, (b) FE3, (c) FE4, (d) FE5, (e) FE6, (f) FE7, (g) FE8, (h) FE9, (i) FE10 and (j) FE11.

Table 3-4. The linear range, and linearity (in term of R^2) of each peptide.

Peptide	Linear equation	Linear range (μM)	R^2
FE2	y=1.202x-2.203	2.5-80	0.9961
FE3	y=1.164x-2.469	2.5-80	0.9986
FE4	y=1.758x-0.9415	1.25-80	0.9986
FE5	y=0.5859x+0.3962	1.25-80	0.9960
FE6	y=0.4036x-0.2166	0.625-80	0.9969
FE7	y=0.06856x+0.01555	0.625-80	0.9939
FE8	y=0.04982x+0.03431	0.625-80	0.9929
FE9	y=0.03949x-0.007480	2.5-80	0.9836
FE10	y=2.045x+1.323	0.625-80	0.9942
FE11	y=0.07746x-6.583E-4	0.625-80	0.9944

During the experiment, we found that there were two peaks shown in the HPLC result of FE1. The first larger peak was labeled 0.92 minute, while the smaller peak was labeled 3.43 minutes. To find out the season, these two peaks were sequenced by using LC-MS/MS with FE1 standard solution. The LC-MS/MS results are shown in Figure 3-5.

The result showed that the sequences of the two peaks were exactly the same, which suggested that there were two isomers of FE1, which is consistent with a previous study that suggested peptide isomers might be generated during histidine-containing peptides synthesized.[93] Therefore, FE1 was not suitable for quantitative analysis.

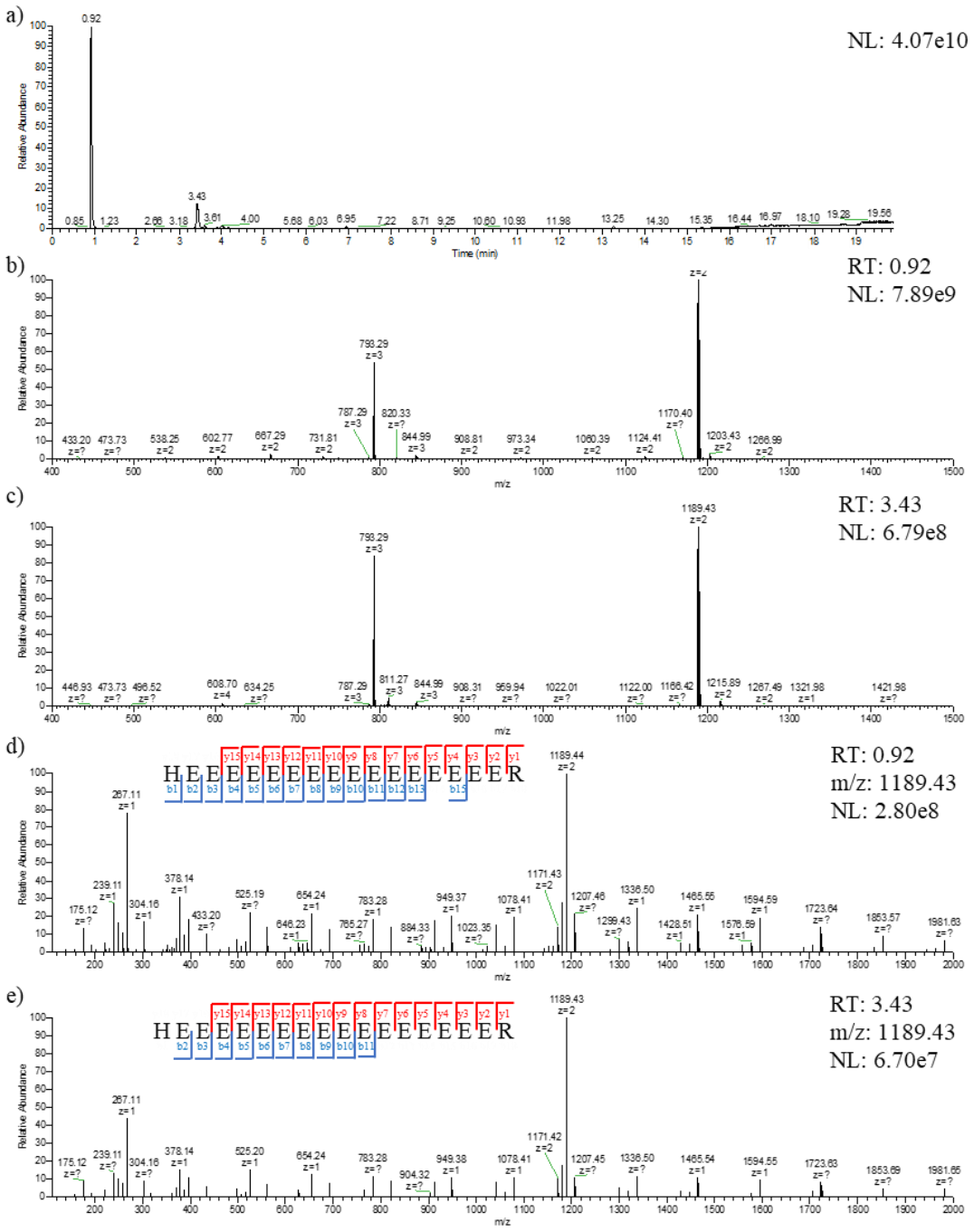


Figure 3-5. (a) The HPLC-MS profile of FE1 standard. (b) MS spectrum of the first peak. (c) MS spectrum of the second peak. (d) MS/MS spectrum of the first peak. (e) MS/MS spectrum of the second peak.

3.3.3 Accuracy and precision

Standard solutions of the targeted peptides at low, middle, and high concentrations were used as quality control samples. The precision and accuracy of targeted peptide were summarized in Table 3-5. For analysis of these peptides, the precision was in general within 20%, except for FE5 at high concentration with a value of 20.49%, which was slightly higher compared with other values. The accuracy of the targeted peptides was in the range of 85 - 106%, except for FE4 at low concentration, FE7 at low concentration, and FE9 at middle concentration. In general, for most peptides, the accuracy at middle concentration was closer to 100%, except for FE4 and FE9, whose accuracy at high concentration was closer to 100%. When constructing a calibration curve, the highest absolute error contributes the most to the overall curve fit.[94] Because the error of the high-concentration standard dominates the calibration curve, the accuracy at low concentrations is compromised.[94] The result suggested that the calibration curve of FE9 might be unreliable.

Table 3-5. The accuracy and precision of each peptide.

Peptide	Concentration (μM)	Determined concentration±S.D. (μM) (n=3)	Accuracy	RSD
FE2	5	4.672±0.0998	93.46%	2.135%
	10	9.141±0.3205	91.41%	3.506%
	60	53.44±5.611	89.07%	10.50%
FE3	5	4.288±0.1856	85.75%	4.329%
	10	9.036±0.3796	90.36%	4.202%
	60	52.17±8.246	86.95%	15.81%
FE4	5	3.577±0.1222	71.53%	3.417%
	10	9.121±0.3310	91.21%	3.629%
	60	62.96±10.34	104.9%	16.42%
FE5	5	4.802±0.1966	96.05%	4.095%
	10	9.741±1.118	97.41%	11.48%
	60	57.51±11.78	95.84%	20.49%
FE6	5	4.291±0.2621	85.82%	6.107%
	10	10.49±0.7438	104.9%	7.092%
	60	47.54±8.578	79.23%	18.05%
FE7	5	3.146±0.08077	62.92%	2.567%
	10	10.27±0.4972	102.7%	4.840%
	60	51.41±4.162	85.69%	8.095%
FE8	5	4.314±0.8096	86.28%	18.77%
	10	10.45±0.6967	104.5%	6.668%
	60	54.56±4.803	90.94%	8.802%
FE9	5	4.825±0.4875	96.50%	10.10%
	10	12.66±1.913	126.6%	15.11%
	60	58.72±1.545	97.86%	2.632%
FE10	5	4.320±0.5512	86.41%	12.76%
	10	10.24±1.459	102.4%	14.25%
	60	53.29±8.147	88.81%	15.29%
FE11	1	1.059±0.1344	105.9%	12.70%
	5	4.496±0.5020	89.92%	11.16%
	10	9.982±0.04985	99.82%	0.4995%
	60	63.50±6.755	105.8%	10.64%

3.3.4 LOD and LOQ

The LOD and LOQ of each peptide using the current method were evaluated experimentally with the standard solution samples at low concentrations. The LOD and LOQ of a targeted peptide were evaluated as the concentrations that could produce signals with I_{signal}/I_{noise} value ≥ 3 and I_{signal}/I_{noise} value ≥ 10. The LOD and LOQ of each peptide were summarized in Table 3-6. The determination of LOD and LOQ of FE3 was shown in Figure 3-6 as examples. The LOD and LOQ of FE3 were 1.0 nM and 2.0 nM respectively. In general, the LODs of these peptides were within the range of 1 - 2 nM, and the LOQs of these peptides were within the range of 2 - 5 nM, except for FE10, whose signal-to-noise ratio was very high compared with others.

Table 3-6. The LOD and LOQ of each peptide.

Peptide	LOD (nM)	LOQ (nM)
FE2	1.0	2.5
FE3	1.0	2.0
FE4	1.0	3.0
FE5	2.0	5.0
FE6	1.0	2.5
FE7	1.0	3.0
FE8	1.0	2.5
FE9	2.0	5.0
FE10	0.1	0.2
FE11	1.0	2.0

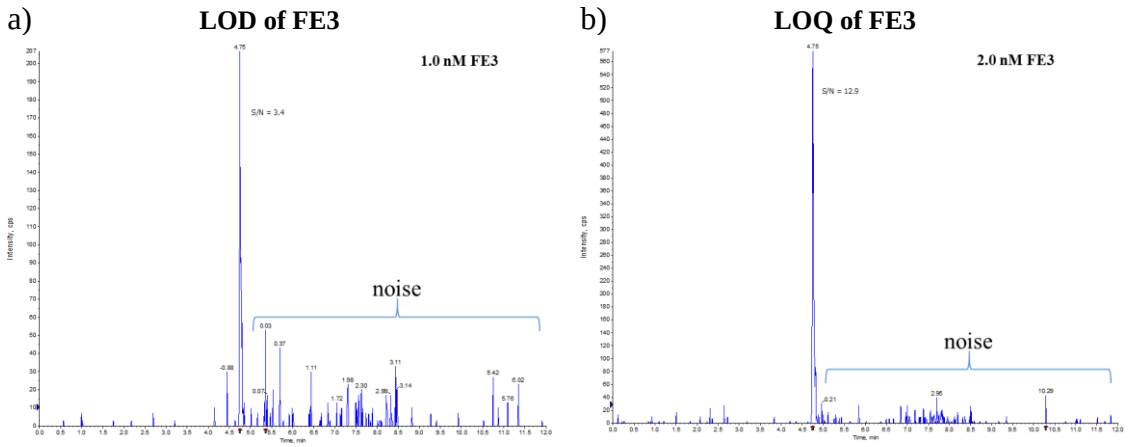


Figure 3-6. Spectra for the evaluation of (a) LOD and (b) LOQ of FE3.

3.3.5 Kinetic stability of peptides

The decay rates of each peptide were measured at different temperatures. The previous study showed that the degradation of peptides in solid state conformed to first-order kinetics, and a critical temperature threshold might be within a range of 50 - 60℃.[67] Therefore, the analyzed peptides were stored above 70℃. According to the rate law for first-order reaction, the kinetic decay rate (k) was calculated by:

$$\ln\left(\frac{N_0}{N}\right) = kt \quad (3\text{-}4)$$

The determination of decay rates of FE6 was shown in Figure 3-7 as examples. From equation (3-4), the relative FE6 concentrations were fitted to a first-order decay rate expression ($\ln(C/C_0)$ vs. duration). The calculated decay rates and linearity of targeted peptides were summarized in Table 3-7. Outlier is identified with a standardized residual that is larger than 3 (in absolute value). Among these peptides, some outliers have been found in the plot of FE2, FE4, FE5, and FE8. These outliers have been removed before linear fitting. In general, the decay rates increased at higher temperatures. While above 70℃, the temperature dependence of the kinetic constants followed the Arrhenius model.

From the temperature dependence of the decay rates, the activation energy (E_a) of each peptide could be calculated according to the Arrhenius equation.

$$lnk = -\frac{E_a}{R} \times \frac{1}{T} + lnA \quad (\text{R}=8.314 J/Kmol) \quad (3\text{-}6)$$

In this equation, A was the pre-exponential factor, E_a was the activation energy and R was the gas constant. From equation (3-6), The ln(k) values were plotted versus 1/T. The determination of E_a and A of FE6 was shown in Figure 3-8 as an example. The

calculated activation energy and pre-exponential factor of each peptide were summarized in Table 3-8.

The discrete distribution of calculated activation energy was analyzed as shown in Figure 3-9. The E_a of most targeted peptides was within a range of 52 ± 15 kJ/mol, except for FE9. The calculated activation energy of FE9 was much higher than the others. Additionally, the amino acid composition of FE9 was very similar to FE8 and FE11, whose activation energy was much lower than FE9. Therefore, we think the higher E_a might be caused by a special structure.

The average level of E_a of 18-amino-acid peptides was calculated with the decay rates of all the targeted peptides, as shown in Figure 3-10. Using the same method to process the data, after eliminating those data that were outside the expected confidence limits (red points), the average decay rate (per year) for peptides with 18 amino acids was therefore related to temperature as follows:

$$lnk = -5740 \times \frac{1}{T} - 2.308 \qquad \mathrm{R} = 8.314 J/Kmol \tag{3-7}$$

With E_a and A of each peptide calculated, the half-life at different temperatures of each peptide can be calculated by:

$$t_{1/2} = \frac{ln2}{k} \tag{3-8}$$

The half-life of the targeted peptide was shown in Figure 3-11. Among all these peptides, the half-life of FE9 was notably different from the others, while at -20℃, the half-life could be as long as 10,000 years. In general, the stability of most of the targeted peptides was similar, e.g., the half-life at -20℃ of most peptides was estimated to be within a range of 10 – 100 years. The average half-life at various temperatures of

peptides with 18 amino acids was calculated with the former equation (3-7) as shown in Figure 3-12.[24] The result suggested that the average half-life of 18-amino-acid peptides was only about 15 years at -20℃. It is obviously not good enough. A study suggested that the half-life of mitochondrial DNA (mtDNA) was 512 years, it was because the mtDNA was encapsulated within ancient fossil bone, which protected the solid DNA from the environment.[28] Previous study showed that DNA has the greatest chance of preservation if encapsulated in fossil bone, and the decay rate was almost 400 times slower than that of *in vitro* DNA.[24, 28] Therefore, the result indicated the necessity of developing a suitable peptide storage method.

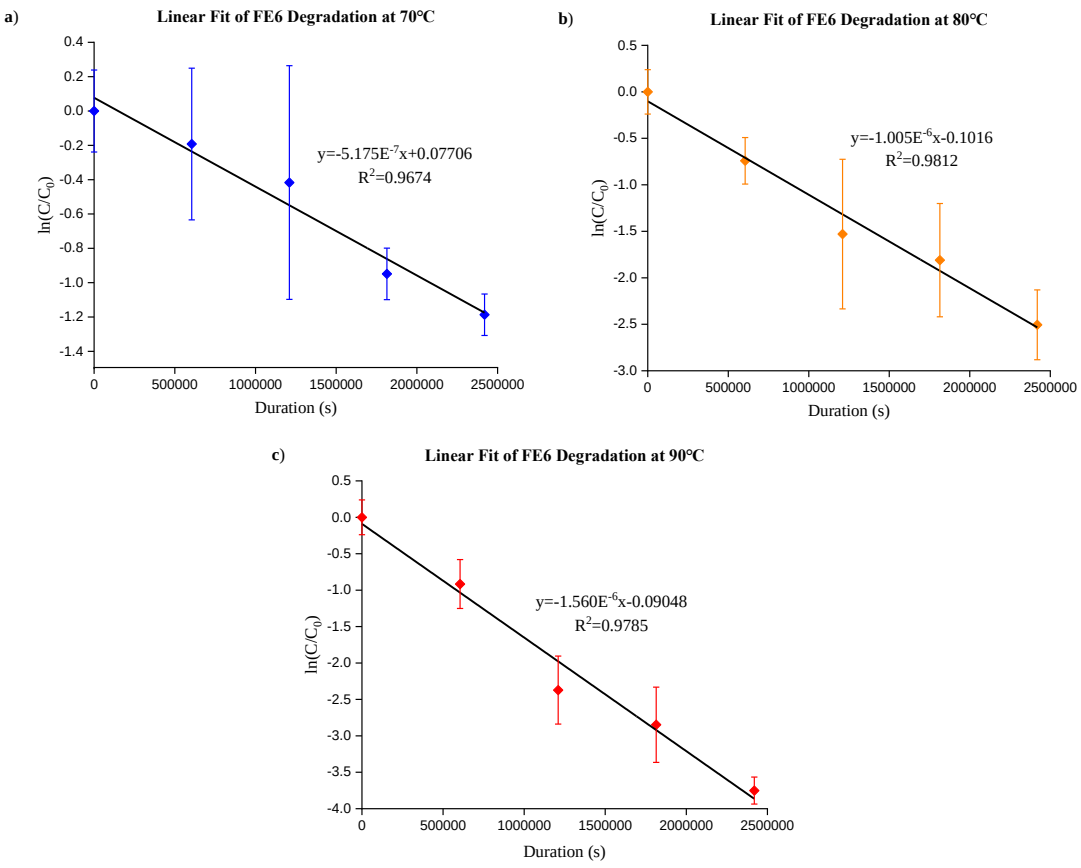


Figure 3-7. Curves for the determination of kinetic decay rate of FE6 at (a) 70℃, (b) 80℃, and (c) 90℃.

Table 3-7. The calculated decay rates, and linearity (in term of R^2) of each peptide.

Peptide	Temperature (°C)	Decay Rate (s^{-1})	R^2
FE2	70	5.946E-07	0.9658
	80	9.246E-07	0.9394
	90	1.425E-06	0.9838
FE3	70	2.504E-07	0.7747
	80	5.172E-07	0.7788
	90	6.205E-07	0.9453
FE4	70	5.388E-07	0.8138
	80	8.412E-07	0.9605
	90	1.124E-06	0.9593
FE5	70	5.925E-07	0.9848
	80	9.624E-07	0.8448
	90	1.382E-06	0.9097
FE6	70	5.175E-07	0.9674
	80	1.005E-06	0.9812
	90	1.560E-06	0.9785
FE7	70	5.120E-07	0.9485
	80	7.603E-07	0.9465
	90	1.763E-06	0.9765
FE8	70	4.425E-07	0.8885
	80	7.649E-07	0.9942
	90	1.196E-06	0.9741
FE9	70	2.294E-07	0.7857
	80	4.077E-07	0.9836
	90	1.317E-06	0.9907
FE10	70	5.170E-07	0.8680
	80	8.074E-07	0.8647
	90	1.288E-06	0.9835
FE11	70	7.390E-07	0.9651
	80	8.914E-07	0.9298
	90	1.508E-06	0.9704

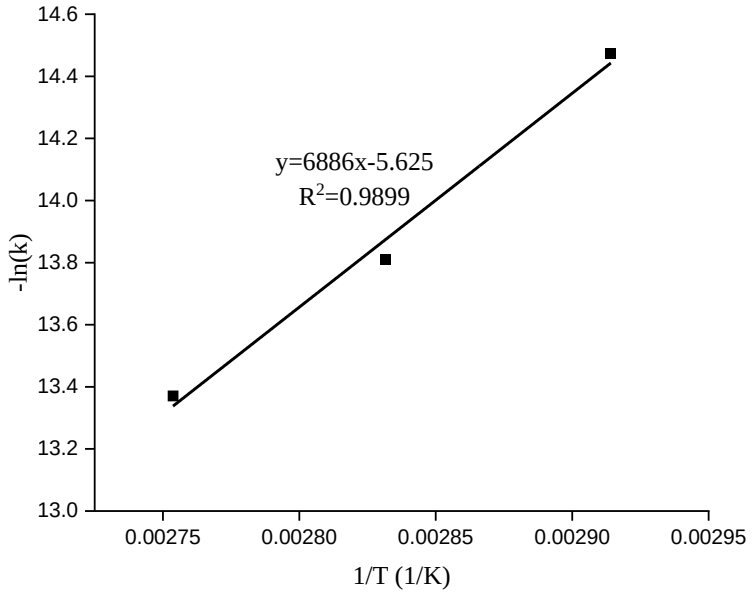


Figure 3-8. Correlation between decay rate and temperature of FE6.

Table 3-8. The calculated activation energy (E_a), pre-exponential factor (A) and linearity (R^2) of each peptide.

Peptide	E_a (kJ/mol)	A	R^2
FE2	45.26	4.594E+00	0.9999
FE3	47.25	4.247E+00	0.9032
FE4	38.13	3.514E-01	0.9889
FE5	43.91	2.913E+00	0.9954
FE6	57.25	2.772E+02	0.9899
FE7	63.83	2.464E+03	0.9518
FE8	51.55	3.145E+01	0.9983
FE9	90.21	1.117E+07	0.9560
FE10	47.25	7.993E+00	0.9991
FE11	36.78	2.773E-01	0.9216

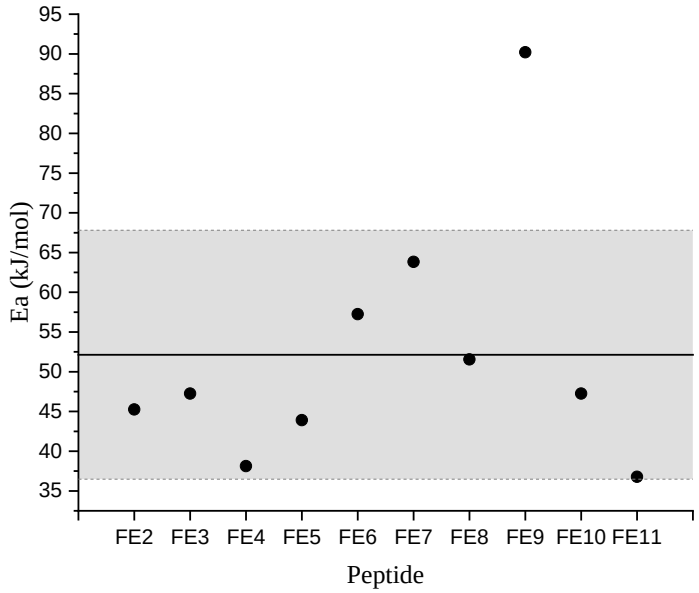


Figure 3-9. The discrete distribution of calculated activation energy.

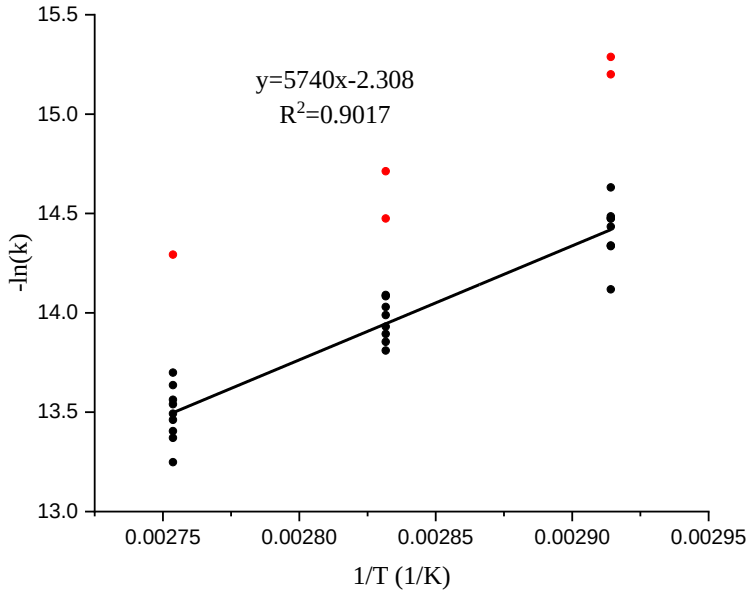


Figure 3-10. Correlation between decay rate and temperature of all the targeted peptides.

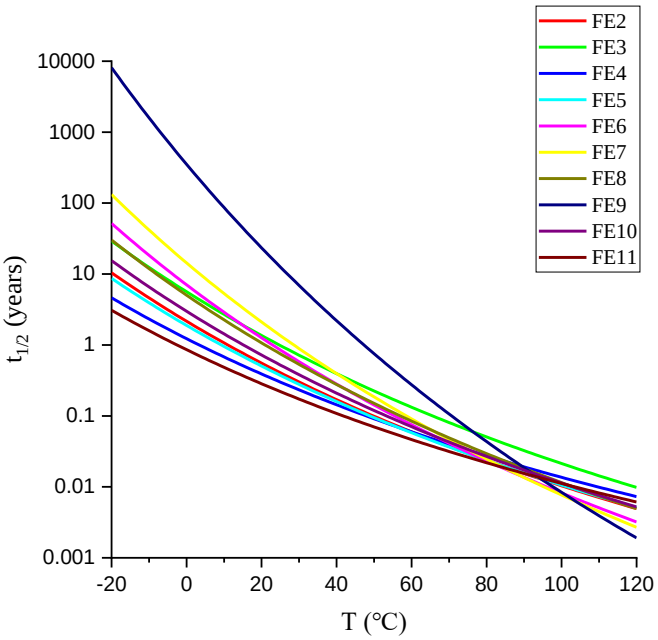


Figure 3-11. The half-life of targeted peptides according to the Arrhenius Equation.

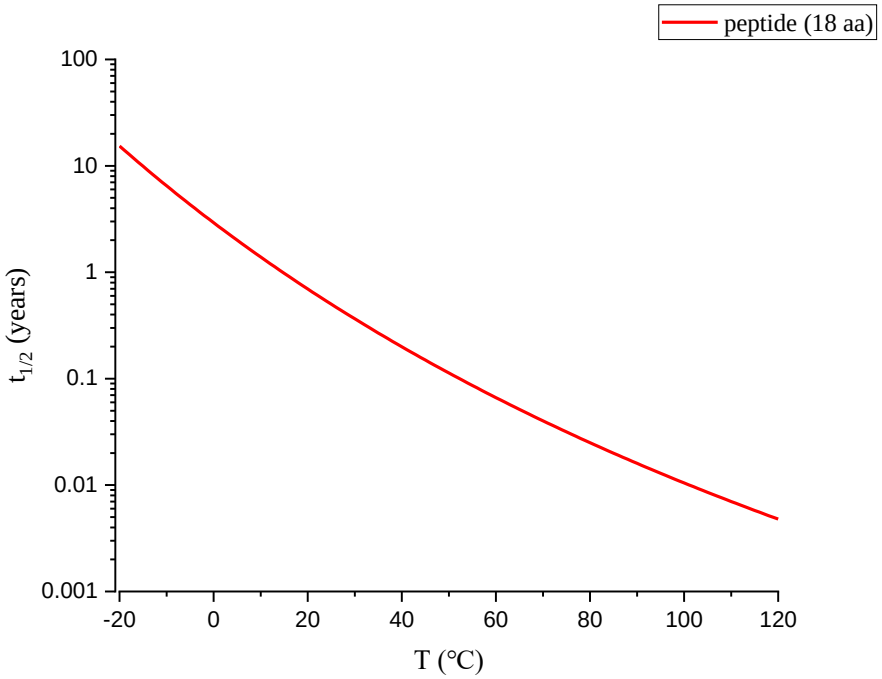


Figure 3-12. The half-life of 18-amino-acid peptides according to the Arrhenius Equation.

3.3.6 Effects of storage methods

The stability of peptides with different storage methods was also explored in this study. Since the decay rate of FE10 was very close to the average level, FE10 was chosen for this study. Four groups of samples were prepared: A, B, C, and D. All the peptides in this study were synthesized in one batch to minimize the system error. In group A, the peptide powder was prepared with vacuum drying; In group B, the samples were prepared with freeze-drying; In group C, the samples were freeze-dried with trehalose; In group D, the freeze-dried samples were stored in nitrogen. Since the stability of peptides under different conditions was varied, the experimental period was also varied. The peptides in group A degraded faster, the sampling interval was set as one week. As for the others, the sampling interval was set as three weeks. Therefore, samples were analyzed in two batches, group A was analyzed in one batch, and the rest three groups were analyzed in one batch. Two calibration curves for each batch were constructed. The calibration curves were shown in Figure 3-13, and the accuracy and precision of the results were shown in Table 3-9. The linear ranges of calibration curves were 5/32-10 μM and 5/32-20 μM. The concentration of all the samples analyzed was within the range. To minimize the systematic error, all the samples, including the standard sample for calibration curves construction and the unknown samples, were analyzed in one batch.

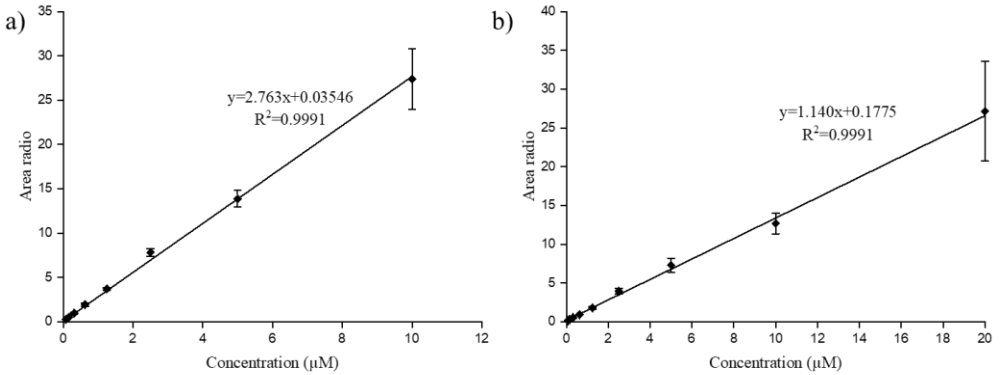


Figure 3-13. Calibration curves for the quantitative analysis of FE10: (a) for group A, (b) for group B, C, and D.

Table 3-9. The accuracy and precision of FE10.

Calibration curve	Concentration (μM)	Determined concentration±S.D. (μM) (n=3)	Accuracy	RSD
A	0.2	0.2149±0.01054	107.4%	4.906%
	2	2.0004±0.4052	100.0%	20.26%
	10	8.5958±0.3732	85.96%	4.342%
B	0.5	0.5172±0.02624	103.4%	5.073%
	2	2.219±0.6190	111.0%	27.89%
	10	13.89±0.3250	138.9%	2.339%

The degradation of FE10 under different conditions was shown in Figure 3-14. The result suggested that the peptides freeze-dried with the trehalose were the most stable, only 36.60% of peptides degraded after being stored at 90℃ for 66 days. The result also indicated that freeze-drying was a better method than vacuum drying for peptide storage, even without any excipients, the peptide dried with freeze-drying was much more stable than the peptide dried with vacuum dry. Additionally, no obvious difference was observed between group B and group D, which might suggest that sealed

in nitrogen was useless for peptide storage. However, it might be because the amino acids prone to oxidation have been excluded in this study.

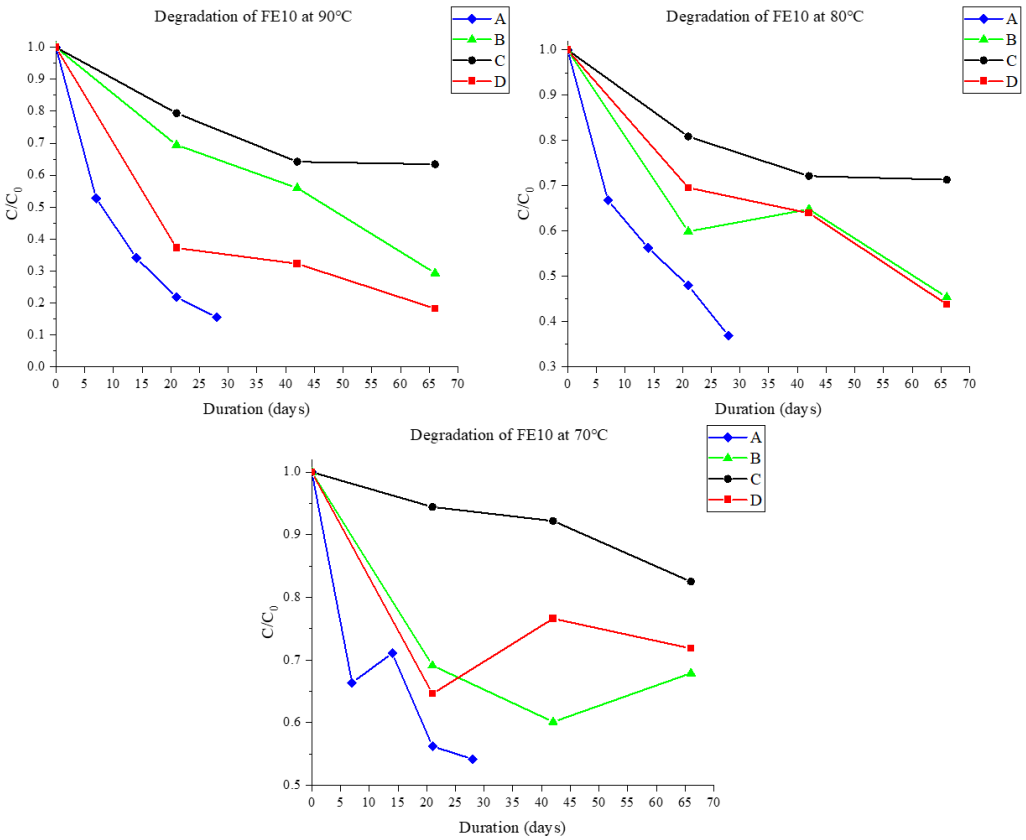


Figure 3-14. The degradation of FE10 under different conditions.

Then, the activation energy and pre-exponential factor of FE10 under different conditions were calculated according to first-order kinetics and the Arrhenius equation. Here only group A and group C were chosen for comparison. The determination of decay rates of FE10 of group C was shown in Figure 3-15 as examples. And the determination of E_a and A of FE10 of group C was shown in Figure 3-16 as an example. The calculated E_a and A were summarized in Table 3-11. The activation energy of FE10 in the previous study was 47.25 kJ/mol, very close to the calculated activation energy of this study, which suggested good repeatability. The result showed that, although

individual decay rates differ, the activation energy is nearly identical for different storage methods, which conformed with a previous study that the activation energy remained the same under different conditions.[95] Then, the half-life at different temperatures has been calculated following equation 3-8. At -20℃, the half-life of peptide generated with freeze-dried with trehalose was more than 200 years, about 3 times longer than that of peptide generated with vacuum dry. The result also proved the potential of peptides for data storage. The freeze-drying method could play a very good protective role for peptide data storage, and by further optimizing the freeze-drying formula, the stability could be further improved, and the data could be stored for a longer period. However, limitations also exist, e.g., introducing excipient would lower data density.

Table 3-10. The calculated decay rates, and linearity (in term of R^2) of FE10.

Group	Temperature (℃)	Decay Rate (s^{-1})	R^2
A	70	2.681E-07	0.9653
	80	3.852E-07	0.9689
	90	7.621E-07	0.9872
C	70	3.353E-08	0.9444
	80	5.568E-08	0.8734
	90	8.256E-08	0.8821

Table 3-11. The calculated activation energy (E_a), pre-exponential factor (A) and linearity (R^2) of FE10.

Group	E_a (kJ/mol)	A	R^2
A	53.95	4.134E+01	0.9638
C	46.73	4.411E-01	0.9969

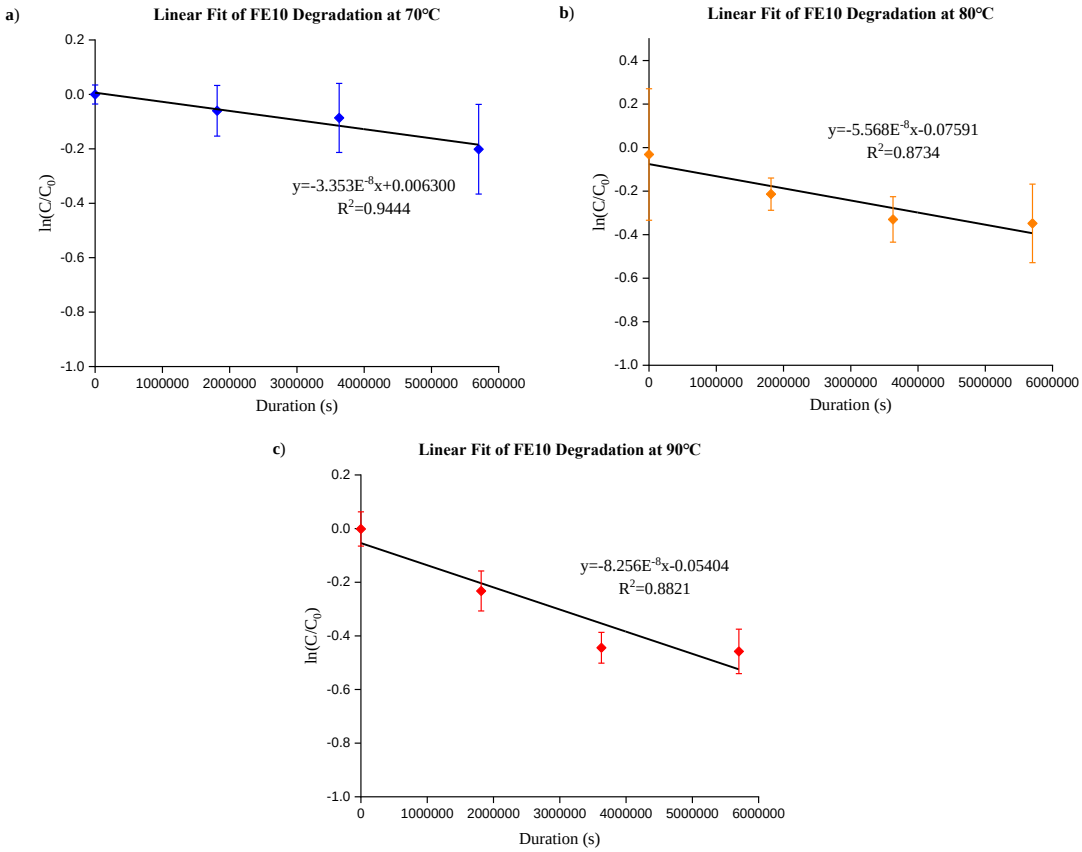


Figure 3-15. Determination of kinetic decay rate of FE10 freeze-dried with trehalose at different temperatures: (a) 70°C, (b) 80°C, and (c) 90°C.

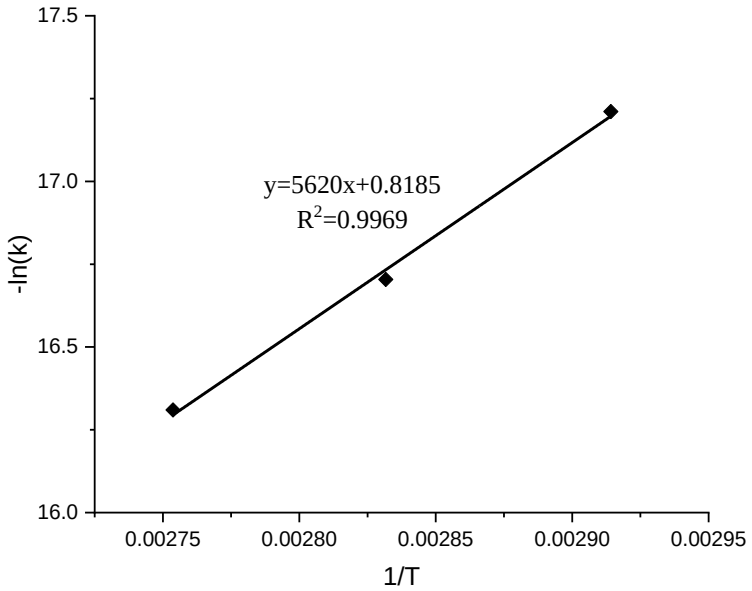


Figure 3-16. Correlation between decay rate and temperature of group C.

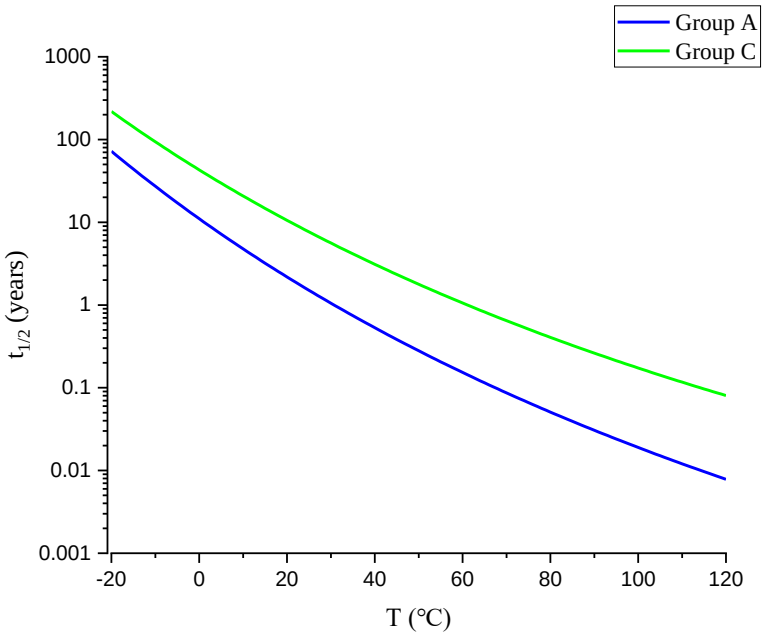


Figure 3-17. The half-life of FE10 of group A and Group C according to the Arrhenius Equation.

3.3.7 The durability of peptide-based data storage

The durability of dataset D was explored in this study. Based on the study in Chapter 3, section 3.3.6, the peptides freeze-dried with trehalose were the most stable. In this part, the peptides encoded dataset D were freeze-dried with trehalose for storage. To shorten the experiment, the peptide mixture was stored at 90℃. The peptides were sequenced every four weeks, and each group was sequenced three times. The recovery of dataset D was shown in Table 3-12, and the LC chromatograms generated for different durations were shown in Figure 3-18.

Table 3-12. The recovery of dataset D.

Duration	No. of correct amino acids	Recovery	Recovery (after error correction)
0 weeks	629	98.28%	100%
4 weeks	609	95.16%	100%
8 weeks	566	88.44%	98.44%
12 weeks	556	86.88%	96.88%

Peptide sequencing results were summarized in Table A2 of the Appendix section. In the beginning, 11 amino acids were not correctly retrieved, mostly because of missed fragmentation. After being stored at 90℃ for 4 weeks, the degradation of two peptides, peptides No. 19 and No. 30, was significant. Although the signal was not good enough for full retrieval, these two peptides could still be sequenced. While after being stored at 90℃ for 8 weeks, three peptides could hardly be sequenced. These three peptides included peptides No. 19, No. 20, and No. 30. Other peptides could still be sequenced after being stored at 90℃ for 12 weeks. Since error correction code allowed correct

data retrieval even if 10% of data was missing, dataset D could be fully retrieved after being stored at 90℃ for 4 weeks, which was thermally equivalent to storing the information at -20℃ for 64 years. Additionally, if we used more bits for error correction codes, e.g., allow 15% of missing amino acids, or develop a more advanced error correction code, the data could be fully retrieved after being stored at 90℃ for 12 weeks, which is thermally equivalent to storing the information at -20℃ for about 200 years.

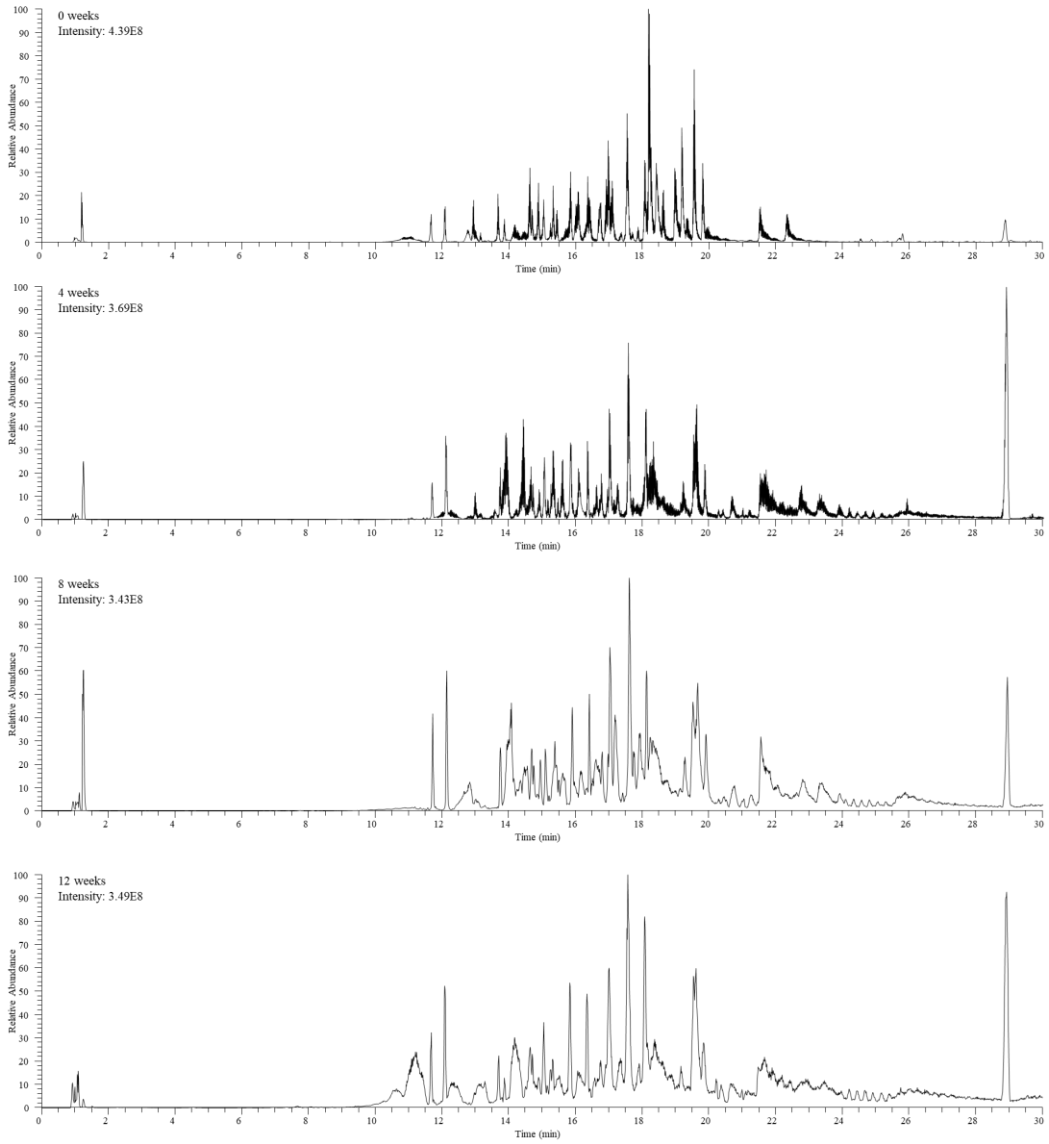


Figure 3-18. The LC chromatograms were obtained with different durations: 0 weeks, 4 weeks, 8 weeks, and 12 weeks.

3.3.8 Peptide structure

In the previous result, we found that the stability of FE9 was very different from the others. Our primary hypothesis was that unusual stability was caused by a special structure. Therefore, in this study, we used AlphaFold 2 to predict the structure of these peptides to verify this hypothesis. AlphaFold 2 was an AI system developed by DeepMind that could realize high-quality predictions of protein 3D structure from its amino acid sequence.[96]

X-ray crystallography was a powerful method for protein structure determination. About 85% of all known protein structures were determined with X-ray crystallography.[97] However, the sample must be crystallizable and an organized single crystal must be obtained. Circular dichroism (CD) spectroscopy was another valuable method for peptide structure analysis; however, a valid reference database was necessary. Since the peptide generated for data storage were designed following specific rules, these could be very different from natural peptides. Besides, CD spectroscopy could only be used to analyze peptide structure in solution, since structural changes could be caused by freeze-drying, it was not a suitable method for structure analysis of peptides in solid state.[98]

AlphaFold 2 was an AI system used to predict crystal structures of proteins. Since freeze-drying has been well developed as a crystallization technique, the peptide crystal structures predicted with AlphaFold 2 were reliable. Besides, AlphaFold 2 could be used for protein with a length from 16 to 2700 amino acids, and even disordered protein. For example, the crystal structure of a secretory abundant heat-soluble SAHS protein from *Ramazzottius varieornatus* (RvSAHS1), which belonged to tardigrade-specific

intrinsically disordered proteins, could be accurately predicted, as shown in Figure 3-19.[99] The confidence of each residual estimate was within the range of 0 to 100, which was called the predicted local-distance difference test (pLDDT).[96] Regions with pLDDT > 90 were expected to be predicted with high accuracy. Regions with pLDDT between 70 and 90 were expected to be predicted well. Regions with pLDDT between 50 and 70 were low confidence and should be treated carefully. Regions with pLDDT < 50 should not be interpreted and possibly suggested the existence of disorder regions. As shown in Figure 3-19, most regions were highly reliable, and the region highlighted in the orange was the his-tag, which was naturally disordered. In addition, a study proved that AlphaFold 2 could be used for predicting peptide structures between 16 – 60 amino acids.[100]

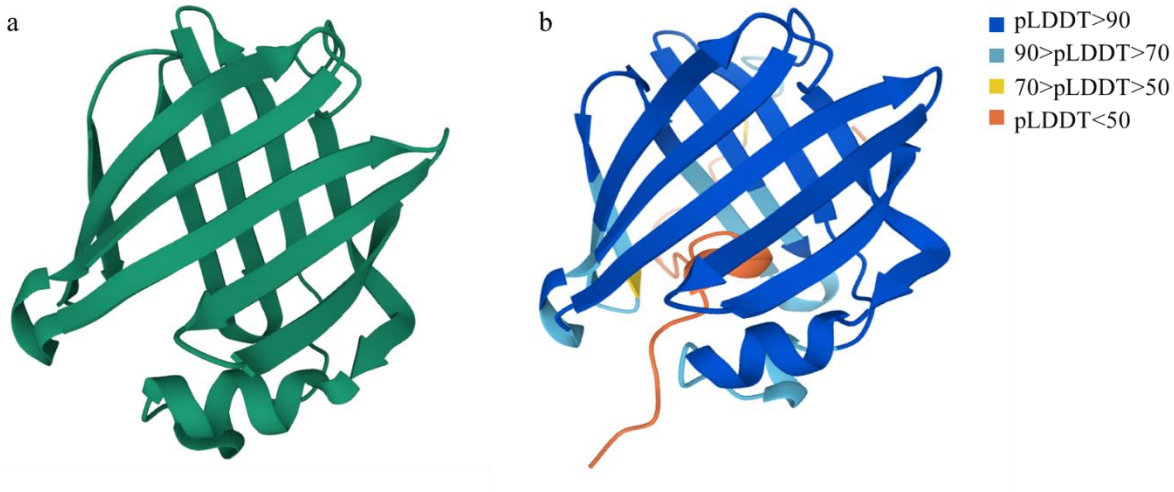


Figure 3-19. Crystal structure of RvSAHS1 (a) determined with X-ray diffraction and (b) predicted with AlphaFold 2. (Reprinted from ref[99])

Then, the structures of the 10 peptides used in the stability test were predicted with AlphaFold 2. Each peptide generated 5 predicted structures and chose the most reliable prediction for analysis. All the pLDDT per position was above 70. The result was

summarized in Figure 3-20. From the predicted structures, we could not find an obvious difference between FE9 and the others. However, we found that all the predicted structures were mostly alpha helix.

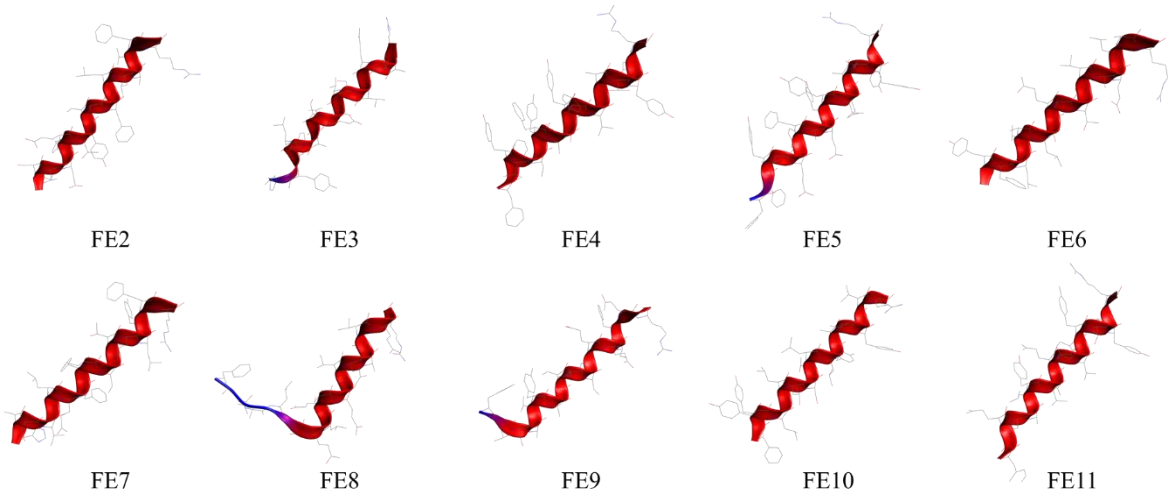


Figure 3-20. The predicted structures of the 10 peptides used in the stability test.

Then, we wondered if all the peptides following the encoding scheme exhibit abundant α-helical structure. Therefore, we generated 2 groups, in a total of 100 peptides to verify the hypothesis. The sequences of the peptides are listed in Table A3 of the Appendix section. The control group consisted of 50 peptides with random sequences, and the experimental group consisted of 50 peptides generated with dataset C following the new encoding scheme. The result was shown in Figure 3-21. 54% of the peptides with random sequences were natively disordered, however, only 14% of the peptides generated following the encoding scheme were natively disordered. The result suggested that the peptides generated following the encoding scheme were more likely to show α-helical structure. A previous study suggested that the stability of the proteins in solution could be improved by designing more helical content, e.g., more resistant to

urea or temperature denaturation.[101] Helical structure might have a similar effect on stability in the solid state. We will explore it in further study.

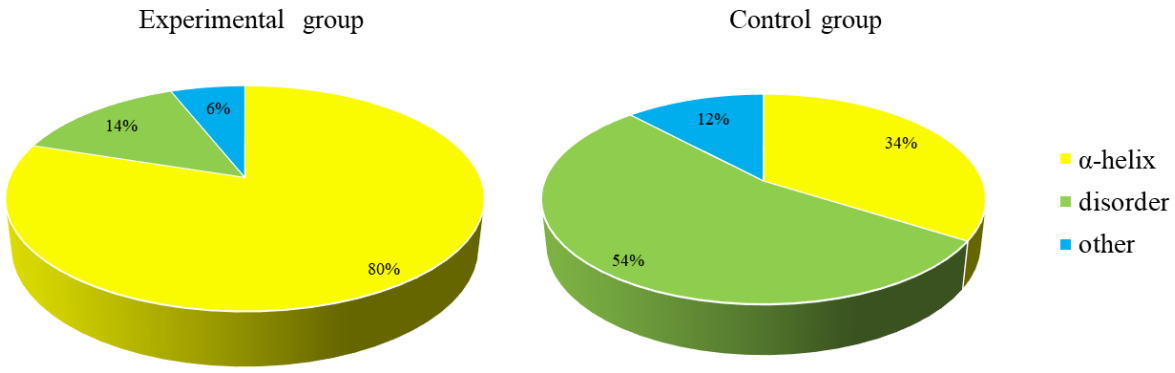


Figure 3-21. The predicted structure of peptides with AlphaFold 2.

3.4 Conclusions

The stability of peptides has been explored in this study. The kinetic decay rate and half-life of peptides have been investigated to explore the possibility of a peptide-based data storage system. The result showed the great potential of peptides as storage material.

The kinetic decay rates of 18-amino-acid peptides at various temperatures have been measured in this study. LC-MRM-MS method was used for quantitative analysis of the 11 targeted peptides. With the decay rates at various temperatures, the average half-life of the peptide was calculated. According to the result, the half-life of these peptides was more than 15 years at -20℃ in solid state, which suggested that it is not good enough for long-term data storage. However, these peptides were stored with a vacuum drying method. Using the freeze-drying method with trehalose, the stability of peptides was highly improved – the half-life at -20℃ was estimated to be more than 200 years. And with further optimization of the freeze-dried formula, peptide kinetic stability could be further improved. The result also showed that some peptides could be extremely stable. It might relate to peptide sequence or structure. Further study is needed to find out the reason. Besides, we found that the peptides generated following the encoding scheme were more likely to show the α-helical structure. In the future, we will further discuss the relationship between α-helix and peptide stability.

To have a wide range of applications, the storage method should be economical and simple. Freeze-dry was the most widely used storage method for peptides and proteins. In the future, we will further develop a suitable method for peptide-based data storage system preservation to provide both high data density and better kinetic stability.

Additionally, we will work with the China Academy of Space Technology to explore the possibility of applying peptide-based data storage technology for space exploration.

Chapter 4: Comparison of 2D-LC with UPLC and nano-LC for analysis of data-storing peptides

4.1 Introduction

Nowadays, two-dimensional liquid chromatography/mass spectrometry (2D-LC-MS) has become a state-of-the-art option for the identification of proteins in proteomics, peptidomics, and related areas. The concept of 2D-LC was first proposed by Giddings in 1987.[102] Compared with 1D-LC, 2D-LC is often a more powerful technique with higher peak capacity and higher resolving power. The 2D-LC instrumentation is a combination of two separated chromatographic systems, which are linked together to allow the fractions from 1D to travel to 2D for further separation. In Giddings' study, multidimensional separation was divided into two categories based on different working principles, and their roles were complementary.[102] Therefore, 2D-LC was also divided into two kinds: heart-cutting two-dimensional liquid chromatography (LC-LC) and comprehensive two-dimensional liquid chromatography (LC x LC).[103]

In LC-LC, only the selected fraction is injected into the 2D column. Therefore, LC-LC is usually applied for targeted analysis in complex matrix, e.g., drug metabolite in serum.[104] This method is relatively simple but limited to a few target compounds. In LC x LC, all fractions from 1D are sequentially sent to 2D, which means that the entire sample undergoes two different separations (Figure 4-1). The 1D column is used for the preliminary separation, the 2D column is used for further separation. Different from LC-LC, LC x LC is used for untargeted type analysis, e.g., protein and metabolism identification in proteomics and metabolomics, since it enables the full analysis of sample composition.[104] Theoretically, the total peak capacity for LC x LC is a linear combination of peak capacities in both separation dimensions (e.g., if peak capacity is

50 in 1D and 80 in 2D, the theoretical peak capacity in LC x LC is 4000).[102] As a matter of fact, the total peak capacity also depends on the orthogonality of two LC modes. Therefore, to achieve higher peak capacity in 2D-LC, the selectivity of the two columns should be as orthogonal as possible.[105] The most common LC separation technique used for peptide separation is reversed-phase chromatography (RP), therefore, 2D-LC is usually set as a combination of RP-LC with another liquid chromatography, e.g., strong cation exchange chromatography (SCX), hydrophilic interaction chromatography (HILIC) and size exclusion chromatography (SEC).[106, 107] Recently, a study has been conducted by Martin Gilar, which aimed to explore the orthogonality of various combinations. The result showed that three combinations could provide suitable orthogonality: SCX-RP, HILIC-RP, and RP-RP system.[108]

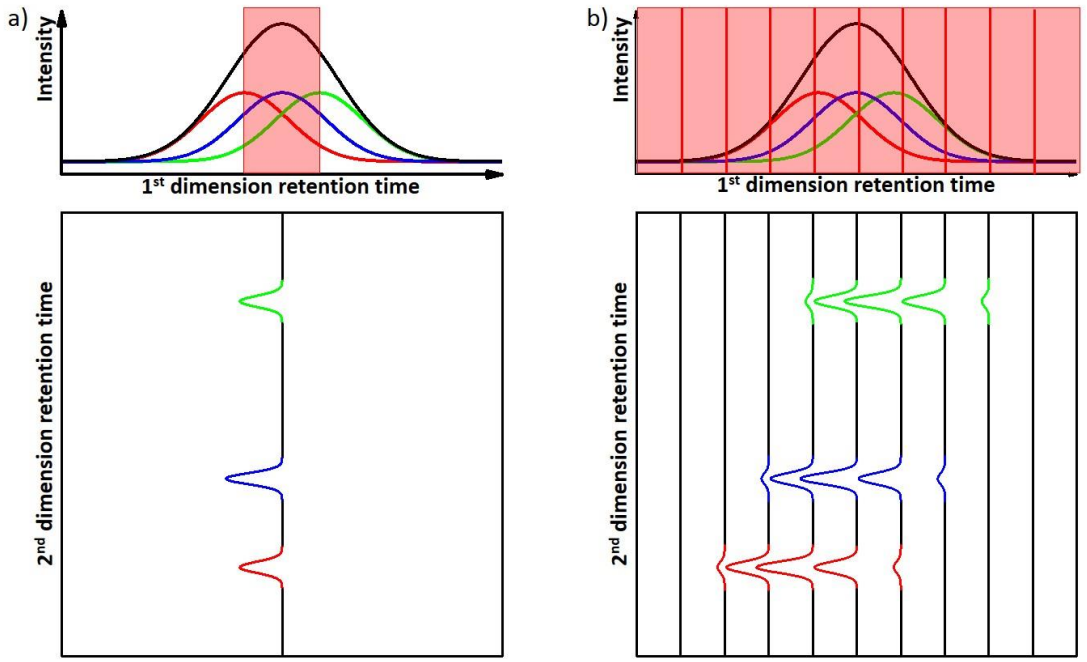


Figure 4-1. (a) Principle of heart-cutting 2D-LC (LC-LC); (b) Principle of comprehensive 2D-LC (LC x LC). (Reprinted from ref[103])

The current LC x LC could be further divided into two categories: online and offline fractionation. When performing an online 2D-LC, the fractions from 1D was directly sent to 2D. Therefore, the 2D is very fast relative to 1D, usually taking 20 s to 2 mins. To make sure the sample dilution and peak width are within reasonable limits, the inner diameter of the 2D column is designed to be much larger than that of the 1D column.[106] However, it requires good compatibility of mobile phase buffers of the 1D column and 2D column. Another choice is offline fractionation, in which the analyte was trapped by an enrichment column before being sent to the 2D column.[106] It is a relatively simple method, and the system could be easily automated. Besides, the most signification advantage is that the samples can be collected at whatever frequency is appropriate for the first dimension, and then run individually on the second dimension at any speed. In this case, higher peak capacity can be obtained because more time can be devoted to 2D separation.[109] In addition, since the capillary column could be used as 2D column, offline fractionation is more frequently used in proteomics. However, due to the undersampling caused by the slower separation speed of the second dimension, the contribution of the first dimension separation to the overall resolution is significantly limited.[106]

2D-LC has already been successfully applied in proteomics for more than 2 decades. Although 2D-LC is a powerful technique, the combination is often complicated. Many factors require consideration, such as the compatibility of the 1D column and 2D column, and the compatibility of the LC method with MS (Table 4-1).

The research objective in this chapter is to develop and improve protocols for LC-MS/MS analysis of data-bearing peptides. In this study, we applied 2D-LC-MS/MS in

peptide-based data storage system for peptide sequencing. In proteomics, the proteins are identified with peptide fingerprinting. Therefore, there is no need for pursuing high coverage. However, this strategy could not be used to sequence data-bearing peptides. Therefore, to achieve a better signal, we chose the offline 2D-LC for peptide separation, then compared it with UPLC and nano-LC to find the suitable method for high-capacity data retrieval.

Table 4-1. Different LC separation modes commonly used in protein analysis. (Reprinted from ref[109])

Mode of separation	**Suitable analytes**	**Common buffer**	**MS compatibility**
RPLC	Proteins, peptides, amino acids	Low concentration volatile buffer	Excellent
HILIC	Peptides, glycans	Low concentration volatile buffer	Excellent
SEC	Proteins, peptides,	Concentrated buff	Poor
Ion exchange (IEX)	Proteins, peptides,	Concentrated buff	Poor
Affinity chromatography	Proteins	Concentrated buff	Poor

4.2 Methods

4.2.1 Materials and chemicals

Based on the study in Chapter 2, we have generated 40 peptides transcoded from dataset D following the new encoding scheme. The peptides used in this study were purchased from Synpeptide Co. Ltd. (Shanghai, China). The peptides varied from 16 to 20 amino acids length without any modification. The sequence of the peptides are listed in Table A4 of the Appendix section. Sodium chloride (NaCl) was purchased from Sigma-Aldrich LLC. (U.S.A.). FA was purchased from VWR LLC. (France). HPLC grade ACN was purchased from Duksan Inc. (South Korea). Water was purified by the MilliQ Direct Laboratory Water Purification system.

4.2.2 Sample preparation

The peptides synthesized were dissolved in DMSO (10 µg/mL) and diluted with 50% ACN with 0.2% FA to 5 nmol/ml for LC-MS/MS analysis. A blank sample containing 50% ACN with 0.2% FA was prepared at the same time. The samples were freshly prepared before analysis.

4.2.3 UPLC

The experiment was performed on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher, MA, USA) coupled with a 2D UPLC (UltiMate 3000 DGLC) system (Thermo Fisher, MA, USA). The scan range was set between 400-1500 *m/z*. The ionspray voltage was set at 2.3 kV for positive ions. The ion transfer tube temperature was set at 280℃. The fragmentation method was HCD with stepped collision energy: 27, 30, and 34. 10 µL of sample solution was injected into the UPLC system, and the peptide mixtures were separated with a C18 column (Thermo Fisher Hypersil GOLD

AQ, 100×2.1 mm, 1.9 μm particle size). The LC parameters were summarized in Table 4-2.

Table 4-2. The UPLC and nano-LC parameters.

	UPLC	**Nano-LC**
Solvent A	0.2% FA in water	0.1% FA in water
Solvent B	0.2% FA in ACN	0.1% FA in ACN
Flow rate	0.3 mL/min	0.3 μL/min
Temperature	55℃	50℃
Gradient	0-3 min: 5% B	0-10 min: 2% B
	3-27 min: 5-50% B	10-12 min: 2-6% B
	27-30 min: 50-95% B	12-47 min: 6-20% B
	30-35 min: 95% B	47-52 min: 20-30% B
	35-35.1 min: 95-5% B	52-56 min: 30-90% B
	35.1-40 min: 5% B	56-61 min: 90% B
		61-61.1 min: 90-2% B
		61.1-66 min: 2% B

4.2.4 Nano-LC

The experiment was performed on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher, MA, USA) coupled with a nano UPLC (Dionex UltiMate 3000 RSLC) system (Thermo Fisher, MA, USA). The scan range was set between 400-1500 *m/z*. The ionspray voltage was set at 2.3 kV for positive ions. The ion transfer tube temperature was set at 300℃. The fragmentation method was HCD with stepped

collision energy: 27, 30, and 34. With 10 µL of sample solution injected into the nano-LC system, the peptide mixtures were separated with a C18 column (Thermo Fisher EASY-Spray PepMap, 15 cm×150 µm, 2 µm particle size). The LC parameters were summarized in Table 4-2.

4.2.5 2D-LC

The experiment was performed on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher, MA, USA) coupled with a 2D UPLC (UltiMate 3000 DGLC) system (Thermo Fisher, MA, USA). The scan range was set between 400-1500 *m/z*. The ionspray voltage was set at 2.3 kV for positive ions. The ion transfer tube temperature and vaporizer temperature were set at 280℃. The fragmentation method was HCD with stepped collision energy: 27, 30, and 34. With 40 µL of sample solution injected into the UPLC system, the peptide mixtures were separated. The first-dimension column was an SCX column (Thermo Fisher BioBasic SCX, 100×2.1 mm, 5 µm particle size). The second-dimension column is a C18 column (Thermo Fisher Hypersil GOLD AQ, 100×2.1 mm, 1.9 µm particle size). The analyte from the 1D column was trapped by a C18 enrichment column (ACQUITY CSH C18 VanGuard pre-column, 5×2.1 mm, 1.7 µm particle size). The scheme of the ten-port two positions valve-based configuration for the 2D-LC experiment was shown in Figure 4-2. The experiment was divided into two phases: the loading phase and the cycle phase. In the loading phase, the sample was loaded onto the SCX column. In the cycle phase, while the valve was at position A, the analyte loaded on the 1D column was eluted by salt plug and trapped by the enrichment column. While the valve was at position B, the analyte trapped by the enrichment column was eluted and further separated by the 2D column. The position of the valve would change several times until all the analyte loaded on the 1D column was eluted.

The valve switched position at 15 min both in the loading phase and the cycle phase. The 2D-LC parameters were summarized in Table 4-3. The above process was repeated for 5 cycles. In each cycle, a salt plug was loaded onto the 1D column for peptide elution. The concentrations of salt plugs were designed to be at 200 mM, 400 mM, 600 mM, 800 mM, and 2000 mM.

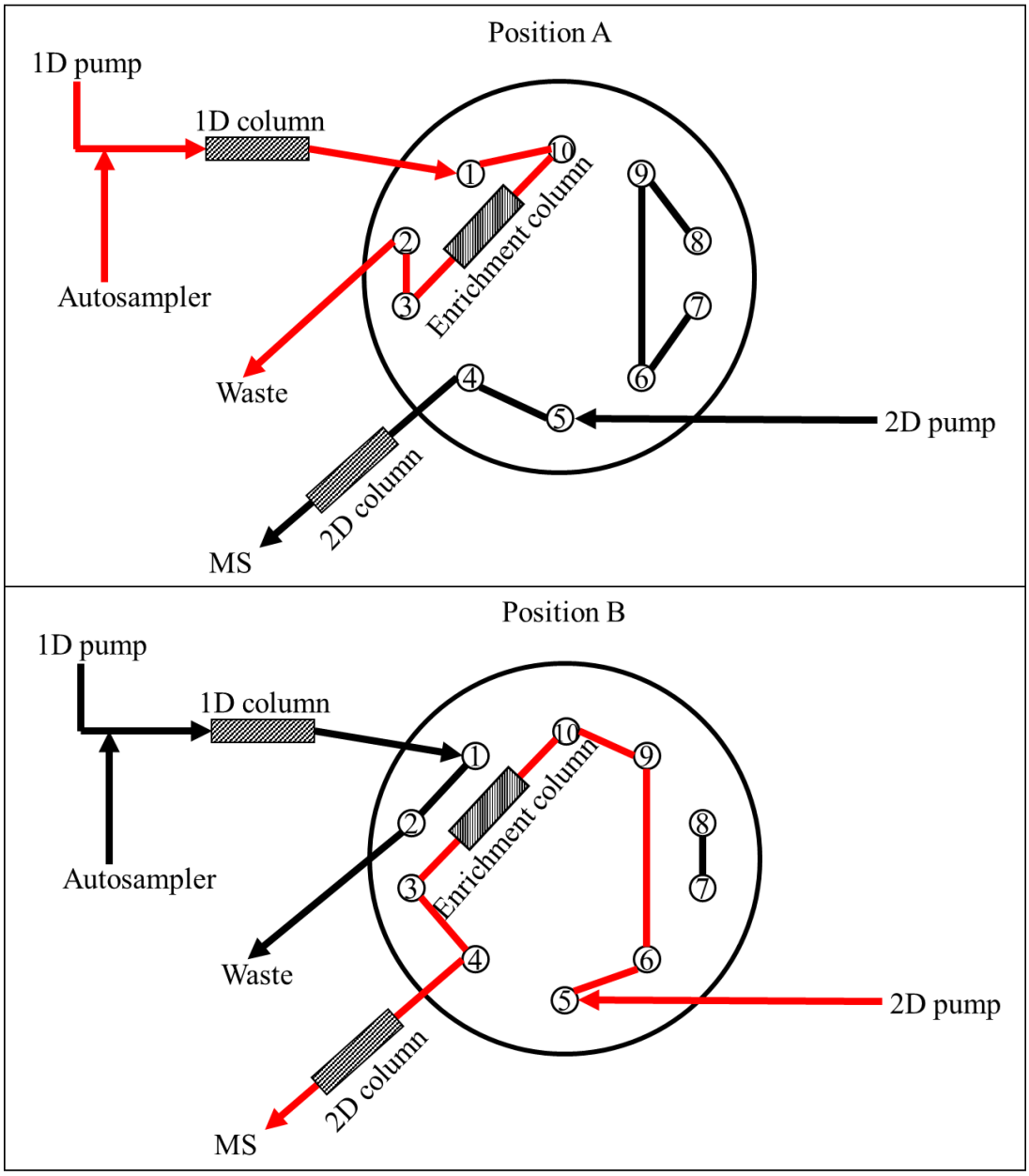


Figure 4-2. 2D-LC configuration.

Table 4-3. 2D-LC parameters.

	1D	**2D**
Solvent A	0.2% FA in water	0.2% FA in water
Solvent B	0.2% FA in ACN	0.2% FA in ACN
Flow rate	0.3 mL/min	0.3 mL/min
Temperature	55℃	55℃
Gradient (loading phase)	0-40 min: 5% B	0-15 min: 5% B
		15-25 min: 5-60% B
		25-30 min: 60-95% B
		30-33 min: 95% B
		33-33.1 min: 95-5% B
		33.1-40 min: 5% B
Gradient (cycle phase)	0-40 min: 5% B	0-23 min: 5% B
		23-35 min: 5-95% B
		35-35.1 min: 95-5% B
		35.1-40 min: 5% B

4.3 Results and discussion

4.3.1 Peptide sequencing

The 40 peptides generated with dataset D were separated with three different LC methods: UPLC, Nano-LC, and 2D-LC. The sequencing recovery was shown in Table 4-4. The result showed that the recovery of peptides separated by UPLC was the highest, which was 95.63%; with those separated by Nano-LC, the recovery was 91.25%, slightly lower than that of those separated with UPLC. Since the error correction code allowed 10% of sequence missing or error, dataset D could be fully retrieved with these two methods. However, only 78.91% of data was correctly retrieved without the error correction code in the 2D-LC group. With UPLC and Nano-LC, all the peptides could be detected, and only 3 and 4 peptides were misread with more than 3 amino acids. The error was mostly caused by missed fragmentation. Since the error rate caused by missed fragmentation normally did not increase as the dataset became larger, it's acceptable with the error correction code. However, with 2D-LC, a total of 11 peptides had more than 3 amino acids misread, and among them, two peptides were not detected at all. This might be due to signal attenuation caused by serial dilution of the sample between the two separations. Additionally, since salt was introduced into the sample solution during 1D elution, more noise could be detected in MS spectra. Therefore, the lower signal-to-noise ratio caused data missing. In proteomic, there was no need to achieve high sequencing coverage for protein identification. However, since the peptides bear information in this study, we should guarantee the data recovery.

In the experiments, 1D column elution was performed using five salt plugs, each of which resulted in an LC-MS/MS analysis. The sequencing result was shown in Table 4-5. The peptide No. 19 and No. 29 were not detected in all five groups, and most

peptides were detected in more than one group, e.g., No. 4, No. 10, No. 12, No. 21, No. 32, and No. 33 were detected in all five groups. 33 peptides were detected in the first group, but only 12 peptides were detected in the last group, which suggested signal attenuation caused by successive dilution.

Table 4-4. The sequencing recovery of dataset D.

No. of correct amino acids	UPLC	Nano-LC	2D-LC
0	0	0	2
1	0	0	0
2	0	0	0
3	0	0	1
4	0	1	3
5	0	1	1
6	0	0	0
7	0	0	2
8	0	0	0
9	0	0	0
10	0	0	0
11	2	0	0
12	1	0	1
13	1	2	3
14	10	15	12
15	8	7	5
16	6	7	5
17	7	4	3
18	5	3	2

Table 4-5. The number of amino acids correctly retrieved with 2D-LC.

Peptide	200 mM	400 mM	600 mM	800 mM	2000 mM	In total
No. 1	14	0	0	0	0	14
No. 2	13	13	0	11	13	13
No. 3	14	14	0	13	0	14
No. 4	15	17	15	11	11	15
No. 5	7	0	3	0	0	7
No. 6	14	12	0	0	0	14
No. 7	13	13	4	0	0	13
No. 8	4	4	2	4	0	4
No. 9	14	11	14	14	0	14
No. 10	18	18	18	16	15	18
No. 11	14	0	0	0	0	14
No. 12	15	15	15	15	15	15
No. 13	0	0	0	3	0	3
No. 14	5	0	0	0	0	14
No. 15	0	18	14	16	18	16
No. 16	0	14	14	14	14	14
No. 17	5	0	0	0	0	5
No. 18	14	3	13	0	0	14
No. 19	0	0	0	0	0	0
No. 20	18	0	0	0	0	18
No. 21	14	14	14	14	14	14
No. 22	15	15	0	15	0	15
No. 23	16	14	0	0	0	16
No. 24	17	17	0	15	0	17
No. 25	4	0	13	0	0	13
No. 26	14	14	0	0	0	14
No. 27	15	0	0	0	0	15
No. 28	16	13	14	0	0	16
No. 29	0	0	0	0	0	0
No. 30	4	3	0	0	0	4

(To be continued)

No. 31	14	0	0	2	3	14
No. 32	15	15	12	15	15	15
No. 33	16	14	16	16	16	16
No. 34	17	12	17	15	0	17
No. 35	14	0	0	0	0	14
No. 36	0	0	14	12	14	14
No. 37	0	0	12	0	0	12
No. 38	7	0	0	0	0	7
No. 39	17	0	6	0	0	17
No. 40	4	0	16	18	16	16
In total	416	283	246	239	164	505

4.3.2 LC chromatograms

LC chromatograms generated with three LC methods have also been compared as shown in Figure 4-3. In this study, with UPLC, Nano-LC, and 2D-LC, 10 µL, 10 µL, and 40 µL sample solution was consumed respectively, and time consumption was 40 min, 66 min, and 240 min respectively. Typically, 2D-LC takes longer than 1D-LC, and 2D-LC consumes more samples than 1D-LC, which are the limitations of 2D-LC. In this study, these limitations would cause lower data density and slower data read speed. Among these three methods, separation by UPLC could achieve the highest data density and fastest data read speed. However, the nano-LC separation appeared to have a 10-fold higher signal intensity than the UPLC separation. Since the column capacity was limited, which meant the signal would be lower when more peptides were separated, Nano-LC might be more suitable for high-capacity data retrieval.

Another limitation of 2D-LC was that coelution has diminished but still existed. As shown in Figure 4-4, peptide No. 2 and peptide No. 21 could not be separated with UPLC, while with 2D-LC, these two peptides still could be separated. It was because the N-terminal and C-terminal amino acids of peptide encoding data were fixed, and the amino acid composition was similar among these peptides. Therefore, to better separate the peptides, the peptide design should be varied in future studies, e.g., incorporating more amino acids.

Additionally, the decreased detection sensitivity and compatibility issues also appeared in the result. The signal intensity of peptide No. 2 decreased from 2.74E8 to 9.32E6 due to successive dilution, and it was also the reason that the two peptides were not sequenced at all. Peptide No. 19 and peptide No. 29 could be detected in the LC-MS

spectra, but the signal was too low for MS/MS analysis. In 2D-LC, the 1D eluent was the injection solvent of the 2D, and therefore there was still a certain concentration of salt remaining in the enrichment column. It not only caused the serious peak tailing of the LC-MS chromatogram but also induced an impurity peak at 18 min.

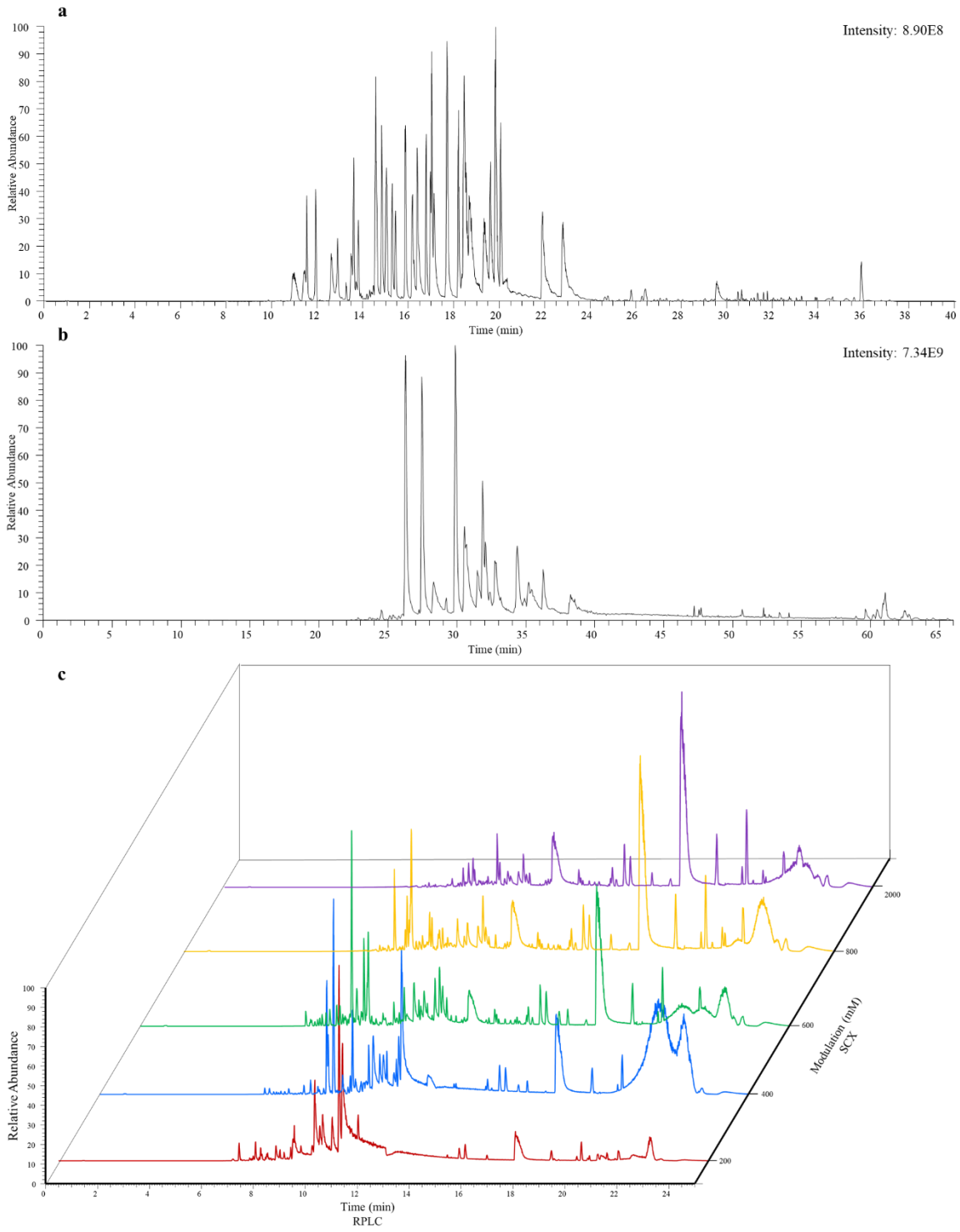
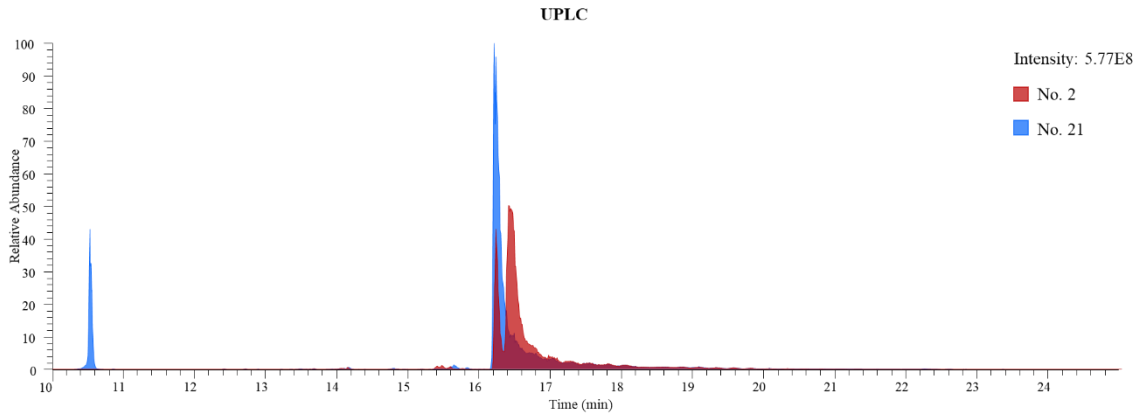
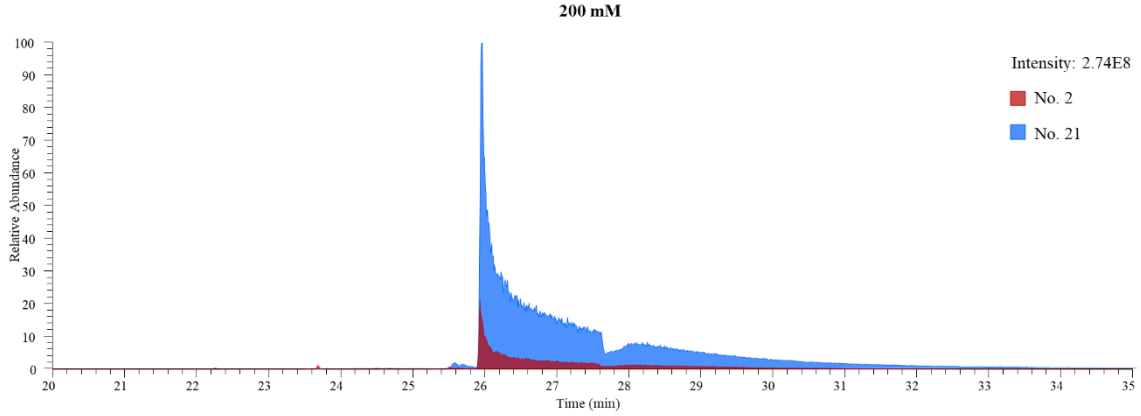
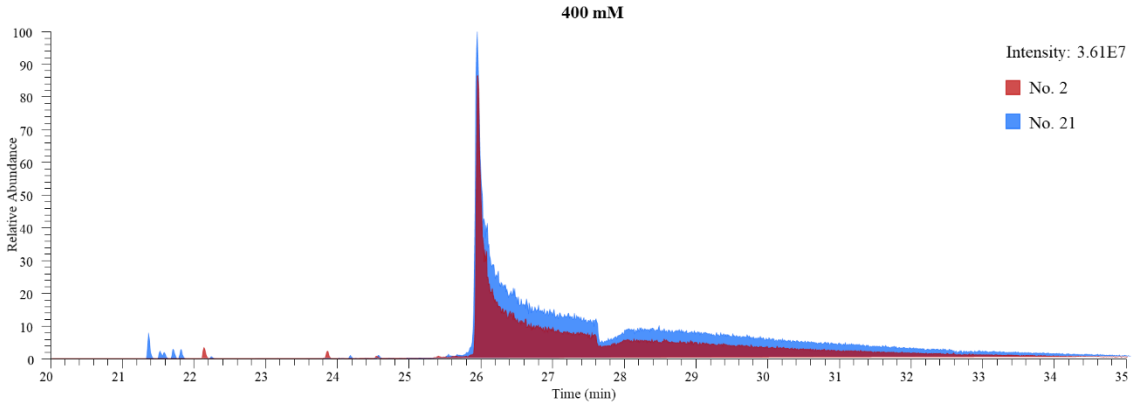
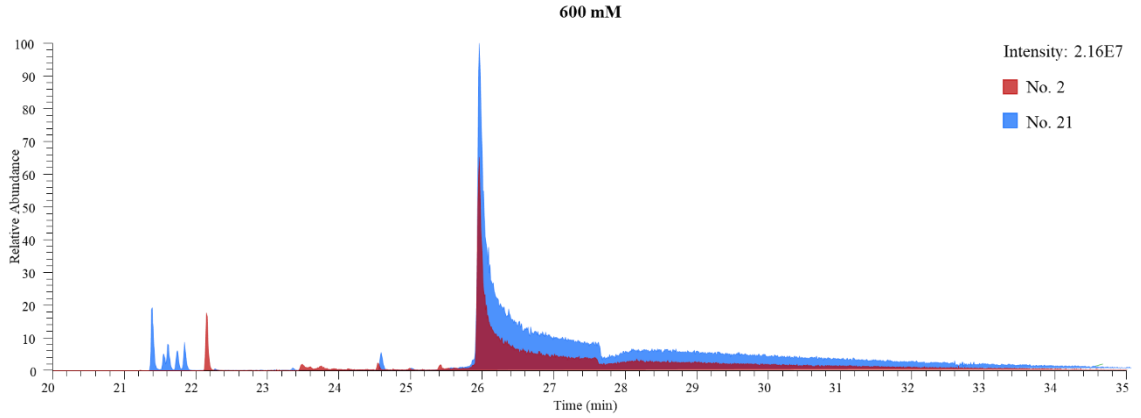


Figure 4-3. LC-MS chromatograms generated with (a) UPLC, (b) Nano-LC, and (c) 2D-LC.









(To be continued)

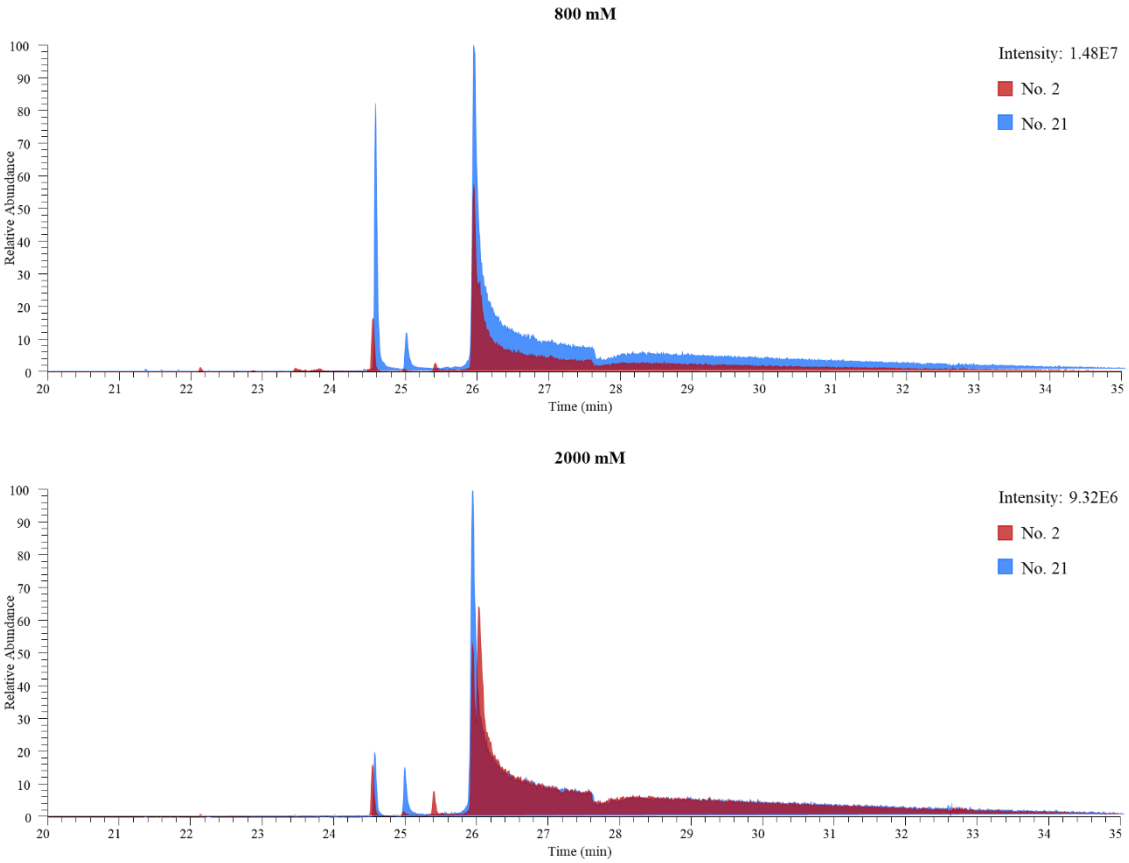


Figure 4-4. The LC-MS chromatograms of peptide No. 2 and peptide No. 21.

4.4 Conclusions

The objective of this study was to develop and improve protocols of LC-MS/MS analysis for high-capacity data retrieval. In this study, we applied 2D-LC-MS/MS in peptide-based data storage system for peptide sequencing, and the result was compared with that of UPLC and nano-LC. The SCX column was chosen as the 1D column, and the RP column worked as the 2D column. The 2D-LC salt plug was chosen for peptide separation to achieve a better signal and avoid poor compatibility of the SCX LC method with MS. 40 peptides generated with dataset D following encoding scheme were chosen for the test. The result showed that peptides could be sequenced with the highest recovery in the shortest separation time with UPLC. The low signal-to-noise ratio of the 2D-LC chromatograms was caused by two reasons: the signal attenuation caused by successive dilution, and the serious peak tailing and induced impurity peak caused by salt residue. Therefore, in the future study, we will optimize the current 2D-LC method, including trying different combinations, e.g., RP-RP, or using the combination of UPLC and nano-LC. Additionally, although nano-LC took longer than UPLC, the signal intensity generated with Nano-LC was 10 times that of UPLC with the same volume of sample solution. Therefore, nano-LC might be suitable for the separation of peptides for high-capacity data retrieval. It will be explored in future studies.

Chapter 5: Overall conclusions and prospects

In the era of big data, global data creation is growing at an unprecedented rate. Due to the limitations of current data storage methods, next-generation data storage technology is urgently needed. In this study, we improved peptide-based data storage system based on the previous study and further discussed the possibility of using peptides as data storage media.

The work in this thesis involved three parts of peptide-based data storage system: data encoding, storage longevity, and data retrieval. The peptide design and encoding scheme were improved to highly reduce the situation of redundant masses. Dataset D was retrieved, and the feasibility of the new peptide design was proved. Then, the kinetic stability of the peptide was explored in this study, and the result showed that the half-life of peptide at -20℃ could be more than 200 years, and the stability could be further improved, which proved the great potential of peptides as storage material. At last, the LC-MS/MS method for peptide sequencing was improved to achieve high-capacity data retrieval. The 2D-LC salt plug was applied for peptide sequencing and the result was compared with that of UPLC and nano-LC. Although the result suggested that 2D-LC might lower data recovery caused by successive dilution and salt residue, it provided valuable information about peptide sequencing.

However, there were several limitations of this study. First, redundant masses of peptides are still a big problem for data retrieving, and since the peptides for data storage were similar, the separation ability of LC was limited. Therefore, the peptide design should be further optimized in the future. Second, the reason for the unusual stability of FE9 was still unknown, and the relationship between the stability and

structure of peptides should be further explored. Third, the current 2D-LC method needs to be further optimized for *de novo* peptide sequencing because of the lower recovery.

Based on the findings of this study, there may be a need for the development of a peptide storage system. Since the existence of a critical threshold, putting all the peptides in one mixture will reduce encoding efficiency. Therefore, a reasonable solution is dividing peptides into several mixtures and stored in a specific order, the order of peptides will be determined by both the address code and position of the mixture. With the development and integration of microfabricated liquid chromatography, microfluidics chips can implement multiple functions, including detection, storage, and preliminary separation. Therefore, a peptide storage system based on a microfluidics chip might be an option. Another direction for future study is to further explore the stability of peptides, including the relationship between peptide stability and peptide structure, and the stability of peptides in space. Finally, the usage of peptide-based data storage system should be reconsidered. Although peptides have shown great potential for high-capacity data storage, the cost to synthesize peptides is still very expensive based on current technologies. There is still a long way to go to apply it for high-capacity data storage. Therefore, other usages of peptide-based data storage system should be considered, e.g., steganography.

References

(1) Prainsack, B. The political economy of digital data: Introduction to the special issue. *Policy Studies* **2020**, *41* (5), 439-446.

(2) Hoffmann, D. L.; Angelucci, D. E.; Villaverde, V.; Zapata, J.; Zilhão, J. Symbolic use of marine shells and mineral pigments by Iberian Neandertals 115,000 years ago. *Science Advances* **2018**, *4* (2), eaar5255.

(3) Holst, 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: May. 2022.

(4) Zhirnov, V.; Zadegan, R. M.; Sandhu, G. S.; Church, G. M.; Hughes, W. L. Nucleic acid memory. *Nature Materials* **2016**, *15* (4), 366-370.

(5) Somasundaram, G.; Shrivastava, A. *Information storage and management*; Wiley Publishing, Inc., 2009.

(6) Wu, C.; Buyya, R. Chapter 12 - Cloud Storage Basics. In *Cloud Data Centers and Cost Modeling*, Wu, C., Buyya, R. Eds.; Morgan Kaufmann, 2015; pp 425-495.

(7) O'Grady, K.; Laidler, H. The limits to magnetic recording — media considerations. *Journal of Magnetism and Magnetic Materials* **1999**, *200* (1), 616-633.

(8) Daniel, E. D.; Mee, C. D.; Clark, M. H. *Magnetic recording: the first 100 years*; John Wiley & Sons, 1998.

(9) Sony. Sony develops magnetic tape technology with the world's highest recording density.http://www.sony.net/SonyInfo/News/Press/201404/14-044E/. 2014.

(10) Tilmanis, P. Intro to Computer Systems https://www.dlsweb.rmit.edu.au/set/Courses/Content/CSIT/oua/cpt160/2014sp4/chapter/08/OpticalMedia.html. Accessed: May. 2022.

(11) Miyagawa, N. Overview of Blu-Ray Disc™ recordable/rewritable media technology. *Frontiers of Optoelectronics* **2014**, *7* (4), 409-424.

(12) Jin, Y.; Lee, B. Chapter One - A comprehensive survey of issues in solid state drives. In *Advances in Computers*, Hurson, A. R. Ed.; Vol. 114; Elsevier, 2019; pp 1-69.

(13) O'Reilly, J. Chapter 3 - Network Infrastructure Today. In *Network Storage*, O'Reilly, J. Ed.; Morgan Kaufmann, 2017; pp 15-58.

(14) Iraci, J. Caring for audio, video and data recording media https://www.canada.ca/en/conservation-institute/services/preventive-conservation/guidelines-collections/caring-audio-video-data-recording-media.html#a2. Accessed: May. 2022.

(15) Cheng, Y.; Li, S.; Liu, J. Abnormal deformation and negative pressure of a hard magnetic disc under the action of a magnet. *Sensors and Actuators A: Physical* **2021**, *332*, 113065.

(16) Rothenberg, J. Ensuring the longevity of digital documents. *Scientific American* **1995**, *272* (1), 42-47.

(17) Runardotter, M.; Quisbert, H.; Nilsson, J.; Hägerfors, A.; Mirijamdotter, A. The information life cycle: issues in long-term digital preservation. In *Information Systems Research Seminars in Scandinavia: 06/08/2005-10/08/2005*, 2005.

(18) Extance, A. How DNA could store all the world's data. *Nature News* **2016**, *537* (7618), 22.

(19) Wetterstrand, K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) http://www.genome.gov/sequencingcostsdata. Accessed: May. 2022.

(20) Ma, S.; Tang, N.; Tian, J. DNA synthesis, assembly and applications in synthetic biology. *Current Opinion in Chemical Biology* **2012**, *16* (3), 260-267.

(21) Church, G. M.; Gao, Y.; Kosuri, S. Next-generation digital information storage in DNA. *Science* **2012**, *337* (6102), 1628-1628.

(22) Goldman, N.; Bertone, P.; Chen, S.; Dessimoz, C.; Leproust, E. M.; Sipos, B.; Birney, E. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **2013**, *494* (7435), 77-80.

(23) Sennels, L.; Bentin, T. To DNA, all information is equal. *Artif DNA PNA XNA* **2012**, *3* (3), 109-111.

(24) Allentoft, M. E.; Collins, M.; Harker, D.; Haile, J.; Oskam, C. L.; Hale, M. L.; Campos, P. F.; Samaniego, J. A.; Gilbert, M. T. P.; Willerslev, E.; et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences* **2012**, *279* (1748), 4724-4733.

(25) Orlando, L.; Ginolhac, A.; Zhang, G.; Froese, D.; Albrechtsen, A.; Stiller, M.; Schubert, M.; Cappellini, E.; Petersen, B.; Moltke, I. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **2013**, *499* (7456), 74-78.

(26) Quail, M.; Smith, M. E.; Coupland, P.; Otto, T. D.; Harris, S. R.; Connor, T. R.; Bertoni, A.; Swerdlow, H. P.; Gu, Y. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* **2012**, *13* (1), 1-13.

(27) Erlich, Y.; Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **2017**, *355* (6328), 950-954.

(28) 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. *Angewandte Chemie International Edition* **2015**, *54* (8), 2552-2555.

(29) Tabatabaei Yazdi, S. M. H.; Yuan, Y.; Ma, J.; Zhao, H.; Milenkovic, O. A rewritable, random-access DNA-based storage system. *Scientific Reports* **2015**, *5* (1), 1-10.

(30) Bornholt, J.; Lopez, R.; Carmean, D. M.; Ceze, L.; Seelig, G.; Strauss, K. A DNA-based archival storage system. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, 2016; pp 637-649.

(31) Blawat, M.; Gaedke, K.; Huetter, I.; Chen, X.-M.; Turczyk, B.; Inverso, S.; Pruitt, B. W.; Church, G. M. Forward error correction for DNA data storage. *Procedia Computer Science* **2016**, *80*, 1011-1022.

(32) Yazdi, S. M. H. T.; Gabrys, R.; Milenkovic, O. Portable and error-free DNA-based data storage. *Scientific Reports* **2017**, *7* (1), 1-6.

(33) Koch, J.; Gantenbein, S.; Masania, K.; Stark, W. J.; Erlich, Y.; Grass, R. N. A DNA-of-things storage architecture to create materials with embedded memory. *Nature Biotechnology* **2020**, *38* (1), 39-43.

(34) Ng, C. C. A.; Tam, W. M.; Yin, H.; Wu, Q.; So, P.-K.; Wong, M. Y.-M.; Lau, F. C. M.; Yao, Z.-P. Data storage using peptide sequences. *Nature Communications* **2021**, *12* (1), 1-10.

(35) Fischer, E.; Fourneau, E. Ueber einige derivate des glykocolls. In *Untersuchungen über Aminosäuren, Polypeptide und Proteïne (1899–1906)*, Springer, 1906; pp 279-289.

(36) Curtius, T. Ueber einige neue der Hippursäure analog constituirte, synthetisch dargestellte Amidosäuren. *Journal für praktische Chemie* **1882**, *26* (1), 145-208.

(37) Curtius, T. Verkettung von amidosäuren I. Abhandlung. *Journal für praktische Chemie* **1904**, *70* (1), 57-72.

(38) Bergmann, M.; Zervas, L. Über ein allgemeines Verfahren der Peptid‐synthese. *Berichte der deutschen chemischen Gesellschaft (A and B Series)* **1932**, *65* (7), 1192-1201.

(39) Merrifield, R. B. Solid phase peptide synthesis. I. The synthesis of a tetrapeptide. *Journal of the American Chemical Society* **1963**, *85* (14), 2149-2154.

(40) Stawikowski, M.; Fields, G. B. Introduction to peptide synthesis. *Current Protocols in Protein Science* **2012**, *69* (1), 18.11.11-18.11.13.

(41) Ryle, A.; Sanger, F.; Smith, L.; Kitai, R. The disulphide bonds of insulin. *Biochemical Journal* **1955**, *60* (4), 541-556.

(42) Hughes, J.; Smith, T. W.; Kosterlitz, H. W.; Fothergill, L. A.; Morgan, B. A.; Morris, H. R. Identification of two related pentapeptides from the brain with potent opiate agonist activity. *Nature* **1975**, *258* (5536), 577-579.

(43) Seidler, J.; Zinn, N.; Boehm, M. E.; Lehmann, W. D. De novo sequencing of peptides by MS/MS. *Proteomics* **2010**, *10* (4), 634-649.

(44) Muth, T.; Hartkopf, F.; Vaudel, M.; Renard, B. Y. A potential golden age to come—current tools, recent use cases, and future avenues for de novo sequencing in proteomics. *Proteomics* **2018**, *18* (18), 1700150.

(45) Kaufmann, A. Analytical performance of the various acquisition modes in Orbitrap MS and MS/MS. *Journal of Mass Spectrometry* **2018**, *53* (8), 725-738.

(46) Dole, M.; Mack, L. L.; Hines, R. L.; Mobley, R. C.; Ferguson, L. D.; Alice, M. B. Molecular beams of macroions. *The Journal of Chemical Physics* **1968**, *49* (5), 2240-2249.

(47) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989**, *246* (4926), 64-71.

(48) Rayleigh, L. XX. On the equilibrium of liquid conducting masses charged with electricity. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1882**, *14* (87), 184-186.

(49) Iribarne, J.; Thomson, B. On the evaporation of small ions from charged droplets. *The Journal of Chemical Physics* **1976**, *64* (6), 2287-2294.

(50) Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S. Unraveling the mechanism of electrospray ionization. *Analytical Chemistry* **2013**, *85* (1), 2-9.

(51) Ahadi, E.; Konermann, L. Ejection of solvated ions from electrosprayed methanol/water nanodroplets studied by molecular dynamics simulations. *Journal of the American Chemical Society* **2011**, *133* (24), 9354-9363.

(52) Ahadi, E.; Konermann, L. Modeling the behavior of coarse-grained polymer chains in charged water droplets: Implications for the mechanism of electrospray ionization. *The Journal of Physical Chemistry B* **2012**, *116* (1), 104-112.

(53) Douglas, D. Linear quadrupoles in mass spectrometry. *Mass Spectrometry Reviews* **2009**, *28* (6), 937-960.

(54) Mellon, F. A. Mass spectrometry| Principles and Instrumentation. In *Encyclopedia of Food Sciences and Nutrition (Second Edition)*, Caballero, B. Ed.; Academic Press, 2003; pp 3739-3749.

(55) Perchalski, R. J.; Yost, R. A.; Wilder, B. Structural elucidation of drug metabolites by triple-quadrupole mass spectrometry. *Analytical Chemistry* **1982**, *54* (9), 1466-1471.

(56) Zeller, M.; König, S. The impact of chromatography and mass spectrometry on the analysis of protein phosphorylation sites. *Analytical and Bioanalytical Chemistry* **2004**, *378* (4), 898-909.

(57) Faktor, J.; Dvorakova, M.; Maryas, J.; Procházková, I.; Bouchal, P. Identification and characterisation of pro-metastatic targets, pathways and molecular complexes using a toolbox of proteomic technologies. *Klinická onkologie : casopis Ceské a Slovenské onkologické spolecnosti* **2012**, *25* (Suppl 2), 2S70-72S77.

(58) Zubarev, R. A.; Makarov, A. Orbitrap mass spectrometry. *Analytical Chemistry* **2013**, *85* (11), 5288-5296.

(59) Scigelova, M.; Makarov, A. Orbitrap mass analyzer – Overview and applications in proteomics. *Proteomics* **2006**, *6* (S2), 16-21.

(60) 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. *Analytical Chemistry* **2003**, *75* (9), 1963-1971.

(61) Tabb, D. L.; Huang, Y.; Wysocki, V. H.; Yates, J. R. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Analytical chemistry* **2004**, *76* (5), 1243-1248.

(62) Clark Jr, G. C.; Cain, J. B. *Error-correction coding for digital communications*; Springer Science & Business Media, 2013.

(63) Gallager, R. Low-density parity-check codes. *IRE Transactions on Information Theory* **1962**, *8* (1), 21-28.

(64) Reed, I. S.; Solomon, G. Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics* **1960**, *8* (2), 300-304.

(65) Henninot, A.; Collins, J. C.; Nuss, J. M. The current state of peptide drug discovery: back to the future? *Journal of Medicinal Chemistry* **2018**, *61* (4), 1382-1414.

(66) Lazarski, C. A.; Chaves, F. A.; Jenks, S. A.; Wu, S.; Richards, K. A.; Weaver, J.; Sant, A. J. The kinetic stability of MHC class II: peptide complexes is a key parameter that dictates immunodominance. *Immunity* **2005**, *23* (1), 29-40.

(67) Oliyai, C.; Patel, J. P.; Carr, L.; Borchardt, R. T. Chemical pathways of peptide degradation. VII. Solid state chemical instability of an aspartyl residue in a model hexapeptide. *Pharmaceutical Research* **1994**, *11* (6), 901-908.

(68) Lai, M.; Topp, E. Solid-state chemical stability of proteins and peptides. *Journal of Pharmaceutical Sciences* **1999**, *88* (5), 489-500.

(69) Chang, L. L.; Pikal, M. J. Mechanisms of protein stabilization in the solid state. *Journal of Pharmaceutical Sciences* **2009**, *98* (9), 2886-2908. Kertscher, U.; Bienert, M.; Krause, E.; Sepetov, N.; Mehlis, B. Spontaneous chemical degradation of substance P in the solid phase and in solution. *International Journal of Peptide and Protein Research* **1993**, *41* (3), 207-211.

(70) Patel, K. B., Ronald T. Chemical pathways of peptide degradation. VII. Kinetics of deamidation of an aspartyl residue in a model hexapeptide. *Pharmaceutical Research* **1990**, *7* (7), 703-711.

(71) Li, S.; Patapoff, T. W.; Overcashier, D.; Hsu, C.; Nguyen, T. H.; Borchardt, R. T. Effects of reducing sugars on the chemical stability of human relaxin in the lyophilized state. *Journal of Pharmaceutical Sciences* **1996**, *85* (8), 873-877.

(72) Manning, M. C.; Patel, K.; Borchardt, R. T. Stability of protein pharmaceuticals. *Pharmaceutical Research* **1989**, *6* (11), 903-918.

(73) Fransson, J.; FIorin-Robertsson, E.; Axelsson, K.; Nyhlén, C. Oxidation of human insulin-like growth factor I in formulation studies: kinetics of methionine oxidation in aqueous solution and in solid state. *Pharmaceutical Research* **1996**, *13* (8), 1252-1257.

(74) Pikal, M.; Dellerman, K.; Roy, M. Formulation and stability of freeze-dried proteins: effects of moisture and oxygen on the stability of freeze-dried formulations of human growth hormone. *Developments in Biological Standardization* **1992**, *74*, 21-37.

(75) Ye, W. P.; Xu, Y.; Du, F. S. The properties of polyesteramide and the effects on the stability of bovine serum albumin. *Pharmaceutical Development and Technology* **1999**, *4* (1), 97-106.

(76) Chang, B. S.; Randall, C. S.; Lee, Y. S. Stabilization of lyophilized porcine pancreatic elastase. *Pharmaceutical Research* **1993**, *10* (10), 1478-1483.

(77) Li, S.; Schöneich, C.; Wilson, G. S.; Borchardt, R. T. Chemical pathways of peptide degradation. V. Ascorbic acid promotes rather than inhibits the oxidation of methionine to methionine sulfoxide in small model peptides. *Pharmaceutical Research* **1993**, *10* (11), 1572-1579.

(78) Carpenter, J. F.; Pikal, M. J.; Chang, B. S.; Randolph, T. W. Rational design of stable lyophilized protein formulations: some practical advice. *Pharmaceutical Research* **1997**, *14* (8), 969-975.

(79) Bansal, A.; Lale, S.; Goyal, M. Development of lyophilization cycle and effect of excipients on the stability of catalase during lyophilization. *International Journal of Pharmaceutical Investigation* **2011**, *1* (4), 214.

(80) Tang, X. C.; Pikal, M. J. Design of freeze-drying processes for pharmaceuticals: practical advice. *Pharmaceutical Research* **2004**, *21* (2), 191-200.

(81) Franks, F. Freeze-drying of bioproducts: Putting principles into practice. *European journal of Pharmaceutics and BioPharmaceutics* **1998**, *45* (3), 221-229.

(82) Kasper, J. C.; Friess, W. The freezing step in lyophilization: Physico-chemical fundamentals, freezing methods and consequences on process performance and quality attributes of biopharmaceuticals. *European Journal of Pharmaceutics and Biopharmaceutics* **2011**, *78* (2), 248-263.

(83) Kasper, J. C.; Winter, G.; Friess, W. Recent advances and further challenges in lyophilization. *European Journal of Pharmaceutics and Biopharmaceutics* **2013**, *85* (2), 162-169.

(84) Wang, W. Lyophilization and development of solid protein pharmaceuticals. *International Journal of Pharmaceutics* **2000**, *203* (1), 1-60.

(85) Hottot, A.; Vessot, S.; Andrieu, J. Freeze drying of pharmaceuticals in vials: Influence of freezing protocol and sample configuration on ice morphology and freeze-dried cake texture. *Chemical Engineering and Processing: Process Intensification* **2007**, *46* (7), 666-674.

(86) Lale, S. V.; Goyal, M.; Bansal, A. K. Development of lyophilization cycle and effect of excipients on the stability of catalase during lyophilization. *International Journal of pharmaceutical investigation* **2011**, *1* (4), 214-221.

(87) Fox, K. C. Putting proteins under glass. *Science* **1995**, *267* (5206), 1922-1924.

(88) Hagen, S. J.; Hofrichter, J.; Eaton, W. A. Geminate rebinding and conformational dynamics of myoglobin embedded in a glass at room temperature. *The Journal of Physical Chemistry* **1996**, *100* (29), 12008-12021.

(89) Pikal, M. J.; Rigsbee, D. R. The stability of insulin in crystalline and amorphous solids: observation of greater stability for the amorphous form. *Pharmaceutical Research* **1997**, *14* (10), 1379-1387.

(90) Allison, S. D.; Randolph, T. W.; Manning, M. C.; Middleton, K.; Davis, A.; Carpenter, J. F. Effects of drying methods and additives on structure and function of

actin: mechanisms of dehydration-induced damage and its inhibition. *Archives of Biochemistry and Biophysics* **1998**, *358* (1), 171-181.

(91) Leslie, S. B.; Israeli, E.; Lighthart, B.; Crowe, J. H.; Crowe, L. M. Trehalose and sucrose protect both membranes and proteins in intact bacteria during drying. *Applied and Environmental Microbiology* **1995**, *61* (10), 3592-3597.

(92) Abd-Elrahman, M. I.; Ahmed, M. O.; Ahmed, S. M.; aboul-Fadl, T.; El-Shorbagi, A. Kinetics of solid state stability of glycine derivatives as a model for peptides using differential scanning calorimetry. *Biophysical Chemistry* **2002**, *97* (2), 113-120.

(93) Stewart, J. M. Solid phase peptide synthesis. *Journal of Macromolecular Science: Part A - Chemistry* **2006**, *10* (1-2), 259-288. Mosquera, J.; Sánchez, M. I.; Mascareñas, J. L.; Vazquez, M. E. Synthetic peptides caged on histidine residues with a bisbipyridyl ruthenium (ii) complex that can be photolyzed by visible light. *Chemical Communications* **2015**, *51* (25), 5501-5504.

(94) Neubauer, K. Calibration: Effects on accuracy and detection limits in atomic spectroscopy. *Spectroscopy* **2021**, *36* (8), 14-16.

(95) Patel, K.; Borchardt, R. T. Chemical pathways of peptide degradation. II. Kinetics of deamidation of an asparaginyl residue in a model hexapeptide. *Pharmaceutical Research* **1990**, *7* (7), 703-711.

(96) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583-589.

(97) Papageorgiou, A. C.; Poudel, N.; Mattsson, J. Protein structure analysis and validation with X-Ray crystallography. In *Protein Downstream Processing*, Springer, 2021; pp 377-404.

(98) Roy, I.; Munishwar. Freeze-drying of proteins: some emerging concerns. *Biotechnology and Applied Biochemistry* **2004**, *39* (2), 165-177.

(99) Fukuda, Y.; Miura, Y.; Mizohata, E.; Inoue, T. Structural insights into a secretory abundant heat-soluble protein from an anhydrobiotic tardigrade, Ramazzottius varieornatus. *FEBS Letters* **2017**, *591* (16), 2458-2469.

(100) Gulsevin, A.; Meiler, J. Benchmarking peptide structure prediction with AlphaFold2. *bioRxiv* **2022**.

(101) Villegas, V.; Viguera, A. R.; Avilés, F. X.; Serrano, L. Stabilization of proteins by rational design of α-helix stability using helix/coil transition theory. *Folding and Design* **1996**, *1* (1), 29-34.

(102) Giddings, J. Concepts and comparisons in multidimensional separation. *Journal of High Resolution Chromatography* **1987**, *10* (5), 319-323.

(103) Teutenberg, T. Basics of 2D-LC. https://analyticalscience.wiley.com/do/10.1002/gitlab.16194. 2018.

(104) Stoll, D. R.; Carr, P. W. Two-dimensional liquid chromatography: A state of the art tutorial. *Analytical Chemistry* **2017**, *89* (1), 519-531.

(105) Delahunty, C.; Yates Iii, J. R. Protein identification using 2D-LC-MS/MS. *Methods* **2005**, *35* (3), 248-255.

(106) Wilson, S. R.; Jankowski, M.; Pepaj, M.; Mihailova, A.; Boix, F.; Vivo Truyols, G.; Lundanes, E.; Greibrokk, T. 2D-LC separation and determination of bradykinin in rat muscle tissue dialysate with on-line SPE-HILIC-SPE-RP-MS. *Chromatographia* **2007**, *66* (7-8), 469-474.

(107) Sandra, K.; Steenbeke, M.; Vandenheede, I.; Vanhoenacker, G.; Sandra, P. The versatility of heart-cutting and comprehensive two-dimensional liquid chromatography

in monoclonal antibody clone selection. *Journal of Chromatography A* **2017**, *1523*, 283-292.

(108) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. Orthogonality of separation in two-dimensional liquid chromatography. *Analytical Chemistry* **2005**, *77* (19), 6426-6434.

(109) Wang, X.; Buckenmaier, S.; Stoll, D. The growing role of two-dimensional LC in the biopharmaceutical industry. *Journal of Applied Bioanalysis* **2017**, *3* (5), 120-126.

Appendices

Table A1. The sequences and masses of 40 peptides encoding dataset D in Chapter 2.

Number	Sequence	Mass
1	HYVTTVEYYVGAEAER	1885.885
2	HVETAYEYTFFFSGATR	2024.927
3	HVYYEFGYESVEATGYVR	2167.985
4	HYETYFAYVTGSATGTYYR	2249.007
5	HVEVAFEYAAEYYVEESYVR	2452.122
6	HYTVGYVYSESEAFFR	1953.890
7	HVYESFEYGAFEYGVTR	2052.922
8	HTYVFVFYFAGEYVGYYR	2280.068
9	HTEVEYETYVFSFETTFYR	2447.096
10	HTYVFVVTTSGSSVATFGVR	2114.080
11	HYYEVAYTESYVGVSR	1921.885
12	HEVYAFATVYETVEVAR	1982.974
13	HVTTGTYTAAESGAYTER	1913.876
14	HYTEEYETSESSFEGYVVR	2310.992
15	HTYVTFTTGAGTGFGAFFYR	2200.038
16	HEYEEAFTFSYSEEFR	2069.864
17	HTEYATEEYVTGGEAGR	1868.818
18	HVEVFTTETGGSFGSGTR	1867.870
19	HEVYVFTEEVFEVTVAVVR	2251.153
20	HVVYVGYEVSGSAFYTTSYR	2284.080
21	HEVVVFTEAYEESAFR	1911.900
22	HYTYSSVESVYSFYEGR	2072.912
23	HVVYSFVEGAFSGTAGYR	1945.932
24	HTTEYYEEFFVASTTTESR	2297.012
25	HTETALAVYEGYTTVGTGGR	2082.002
26	HTETEAYVTVGVFATR	1779.879
27	HVETVGSVEAGYVGFTR	1806.890

(To be continued)

28	HYEYESAVVTESVYTVSR	2117.991
29	HEYTAYSVAGEYTAETTGR	2104.934
30	HYTYSAFVSVAGEFTFYFGR	2348.090
31	HYYEYGTVGSGTEFER	1893.817
32	HTYESYYVFSVETEEFR	2184.964
33	HTEVETAAYVVAEFYEGR	2069.969
34	HTYVFSEATSGVGTTYAYR	2108.980
35	HEVEATYAETYYVEFTAASR	2336.060
36	HEVYYYEAVYFVSTTR	2025.947
37	HETTVEYAAGAEATAVR	1774.849
38	HYTYFVSASVYVVYYYER	2308.084
39	HYEVTYYAGYSEYSYEEYR	2471.023
40	HVYTFGGAFAYEAFGVFTGR	2196.043

Table A2. Peptide sequencing results of dataset D in Chapter 3, Section 3.3.7.

Peptide	No. of incorrect or missed amino acids			
	0 weeks	4 weeks	8 weeks	12 weeks
No.1	0	0	0	0
No.2	0	0	2	2
No.3	0	3	2	3
No.4	0	0	2	2
No.5	0	0	0	0
No.6	0	0	0	0
No.7	2	2	2	2
No.8	0	0	2	2
No.9	0	0	0	0
No.10	0	0	0	0
No.11	0	0	2	2
No.12	0	0	0	0
No.13	0	0	0	0
No.14	0	0	0	0
No.15	0	0	0	2
No.16	0	0	0	0
No.17	0	0	0	0
No.18	0	0	2	2
No.19	0	14	14	14
No.20	0	0	18	18
No.21	0	0	0	0
No.22	0	0	0	0
No.23	0	0	2	2
No.24	0	0	0	0
No.25	0	0	6	5
No.26	0	0	0	0
No.27	0	0	2	2
No.28	0	0	0	0
No.29	0	0	0	0

(To be continued)

No.30	7	12	14	18
No.31	0	0	0	0
No.32	0	0	0	0
No.33	0	0	0	0
No.34	0	0	0	0
No.35	2	0	4	6
No.36	0	0	0	0
No.37	0	0	0	0
No.38	0	0	0	0
No.39	0	0	0	0
No.40	0	0	0	2

Table A3. The sequences of the 100 peptides used in Chapter 3, Section 3.3.8.

Number	Experimental group	Control group
1	FYYYLVLFAYLAYLVYFR	EDILSMISFSLTFLHWKY
2	FYYTAFLEAYSSTEEVSR	FQLCKDCTRYCLCATLCQ
3	FYYESFFFAETSEELVER	IHNHICNEHGTKAVSNRM
4	FYYVSSFTTSFSYLTSAR	MPRFTKNKQDIVNKMCII
5	FYYAAETYVALFYYFTER	GMPWKKNKTREINHSIER
6	FYYSSEAVFLLESFSFFR	YLYKVKQLMCFNWVFRHI
7	FYYLSLLSTLLYAVYAVR	PNMNCMCFYPHKTKHQIW
8	FYYFSSFEELSSYLLFVR	QWFMCIDCNKSAKTLPGS
9	FYTYVVFFSEELASFLER	PYDWVYFCLYHLVVPPQC
10	FYTTVEVEETELLSFSTR	TPNCIGMTHIGMMADQED
11	FYTETVLFATYFSVSLER	CIPTKFQNCIHCNAFWHE
12	FYTVTESSVEFYTAEVYR	VNPYSENIDMVTTPRSHI
13	FYTAYTASYSFSYSTLFR	FPPSMIEFYRWIKIMKSV
14	FYTSTAFLVAELYVALLR	DCLSIQMLRRHVKTMVQA
15	FYTLTFVAEVLAAVSFSR	RVDDGVILIRAAHVLPQT
16	FYTFYLFFATAALLETSR	PWYVYLPFNEHVHFWSSK
17	FYEYEFYYYLVATAAALR	KYFSVEEKYVLEFWGTIK
18	FYETVLVSASFTFASFTR	MGFTQNPQWGCAANVKYI
19	FYEEVEFEFSVFFALAER	DCSREYQADSEVDDCLGH
20	FYEVTVFAEVLYFEYETR	QAPIAAFNEYEVSAGGTE
21	FYEATFYSYALSYTTEYR	WMRGPSTHADRDCKVVNM
22	FYESYSLFTVFEVSSTFR	ILCMLYVYGNGKTRDSFC
23	FYELTLYYFTVLESYFFR	VPDIRTMYDVCAQEENES
24	FYEFYYTAFVTLAVSTLR	DRSNFYYGVNDPSMDNGS
25	FYVYEAETSSAYAFEEAR	IKINGGIAFDAPTTNMHF
26	FYVTVSSSEAELSSESAR	NCMMSEGELDWLNMHVRQ
27	FYVEVSVVASTYYVALVR	SDSHDHTVLCTDRGHRND
28	FYVVETTFSATLVYFALR	QDIPSVMWREIKNCMCMP
29	FYVATSVYVEYSYTTFTR	WNIDPKMKIEWTDCYVDT
30	FYVSTAFVEVAFYLLSVR	RNFEKMAHVKWQHQCDSN

(To be continued)

31	FYVLYFATAEELLFTASR	IGCPMYGYTCRYLHMETW
32	FYVFYFFYYVEFFSAAFR	ACDARGNREHIYTSSSDM
33	FYAYVVFYTYAFSYYSYR	YKCVWMTHLYLQCKSGLF
34	FYATEVSAVYEVAESYFR	RTTADWFTPDYFMYTIVL
35	FYAEELTEEYVYASEEFR	AVTSSVMKRHRGMPHQTQ
36	FYAVELELFFTELYVSLR	HMLNSYQVCSFLNVNHSM
37	FYAAESTSVVAYVVTTFR	HNMDFIPKLLGHDIYARE
38	FYASTATAELTLVAFLAR	DAGQTQHDFICKWIYNRC
39	FYALTASSAVSALFSSTR	YNPAPYFVICTLNDKFFA
40	FYAFTFSAASFSTFFAYR	MRQLNKLKWYPCSQYKQG
41	FYSYELLTTVATSTEVSR	SMNSMHLWVKNDDLPFHI
42	FYSTEFSYAEVYLETYSR	ISSNQTLTIHQWTWCFRN
43	FYSEVVFVSYVFEFELER	LPKTCSYPGGQCIGFNMH
44	FYSVVEELTTFEFFTLYR	DQCFATKQFSTYWLKRQM
45	FYSAEVVLFEFASEVTAR	FRNIEPHQTVFQMKENTD
46	FYSSEFTSSEVAFTEEVR	HPWETIMWYGHEYHRRTI
47	FYSLYLSAVASTVTSALR	AAFYRGQDDRVLDTGEVW
48	FYSFTTEEAFETSAAFSR	IREDKAIDPTTSFVRYII
49	FYLYEVSESETTLEFVAR	EYYFSMLQRKIAYQDDIV
50	FYLTETSAFSFYSFTVFR	YIEMDENREPMDIKPNEH

Table A4. The sequences and masses of 40 peptides encoding dataset D in Chapters 2, 3, and 4.

Number	Sequence	Mass
1	FYVTTVEYYVGAEAER	1885.885
2	FVETAYEYTFFFSGATR	2024.927
3	FVYYEFGYESVEATGYVR	2167.985
4	FYETYFAYVTGSATGTYYR	2249.007
5	FVEVAFEYAAEYYVEESYVR	2452.122
6	FYTVGYVYSESEAFFR	1953.890
7	FVYESFEYGAFEYGVTR	2052.922
8	FTYVFVFYFAGEYVGYYR	2280.068
9	FTEVEYETYVFSFETTFYR	2447.096
10	FTYVFVVTTSGSSVATFGVR	2114.080
11	FYYEVAYTESYVGVSR	1921.885
12	FEVYAFATVYETVEVAR	1982.974
13	FVTTGTYTAAESGAYTER	1913.876
14	FYTEEYETSESSFEGYVVR	2310.992
15	FTYVTFTTGAGTGFGAFFYR	2200.038
16	FEYEEAFTFSYSEEFR	2069.864
17	FTEYATEEYVTGGEAGR	1868.818
18	FVEVFTTETGGSFGSGTR	1867.870
19	FEVYVFTEEVFEVTVAVVR	2251.153
20	FVVYVGYEVSGSAFYTTSYR	2284.080
21	FEVVVFTEAYEESAFR	1911.900
22	FYTYSSVESVYSFYEGR	2072.912
23	FVVYSFVEGAFSGTAGYR	1945.932
24	FTTEYYEEFFVASTTTESR	2297.012
25	FTETALAVYEGYTTVGTGGR	2082.002
26	FTETEAYVTVGVFATR	1779.879
27	FVETVGSVEAGYVGFTR	1806.890
28	FYEYESAVVTESVYTVSR	2117.991

(To be continued)

29	FEYTAYSVAGEYTAETTGR	2104.934
30	FYTYSAFVSVAGEFTFYFGR	2348.090
31	FYYEYGTVGSGTEFER	1893.817
32	FTYESYYVFSVETEEFR	2184.964
33	FTEVETAAYVVAEFYEGR	2069.969
34	FTYVFSEATSGVGTTYAYR	2108.980
35	FEVEATYAETYYVEFTAASR	2336.060
36	FEVYYYEAVYFVSTTR	2025.947
37	FETTVEYAAGAEATAVR	1774.849
38	FYTYFVSASVYVVYYYER	2308.084
39	FYEVTYYAGYSEYSYEEYR	2471.023
40	FVYTFGGAFAYEAFGVFTGR	2196.043