

The Emerging Landscape and Application of Protein Sequencing

Hao Xu *

Pharmaceutical College, China Pharmaceutical University, Wuhan, China

* Corresponding author: 2020202272@stu.cpu.edu.cn

Abstract. Proteins play an indispensable role in all cells and tissues in all living creatures especially the human body. To determine the protein primary structure, also known as protein sequencing, is an important subject in life science and medicine. Scientists have tried chemical reactions such as Edman degradation, as well as instrument analysis like mass spectrometry. These mainstream methods can effectively determine the sequence of protein and are widely used. On the other hand, the new method like nanopore sequencing shows the single molecule level sequencing sensitivity and has great emerging landscape. The development of protein sequencing technology will help people better understand the laws of life activities and achieve early diagnosis and precise treatment of diseases. This paper will briefly summarize the traditional protein sequencing methods, focus on the introduction of the new generation of sequencing technology represented by nanopore sequencing. Meanwhile, after comparing those advantages and disadvantages, its future research direction could be found. More cheap, high-throughput and highly sensitive protein sequencing methods and instruments will be discovered and popularized in the near future.

Keywords: Protein sequencing, nanopore, Edman degradation, mass spectrometry.

1. Introduction

Protein is the cornerstone of organism, and determines every level of cell, such as its growth, differentiation and repair. The regular behavior of proteins is a prerequisite for normal activities of living organisms. It was well recognized by the middle of the 20th century that proteins were made of amino acids, but people were still wondering how these were joined together. There were too many mysteries including whether they were randomly mixed together or in repeated patterns. Soon afterwards, through a series of exploration, techniques were devised to resolve these problems, showing that each protein has its own amino acid residues arranged linearly in a different order.

In the field of sequencing technology, DNA sequencing like human genome project is revolutionary. Since single-molecule DNA and next-generation sequencing technology became available, both precision medical diagnostics and revolutionized genetics have been radically changed. Similarly, if advanced protein sequencing technology can emerge, it will also be a transformative wave for proteomics. It will make single-cell and single-molecule protein analysis possible, thus providing early diagnosis and treatment of diseases [1]. In short, protein sequencing is the determination of protein primary structure, which involves determining the number of polypeptide chains and distinguishing the subsequence of amino acids. Protein sequencing is a relevant and important subject in life science and medicine. For instance, both cancer study and embryonic development are related to the regional heterogeneity in proteomes, which suggests a new line of research. Certain sequencing techniques can locate uncommon proteins with just one or more copies, revealing novel cellular regulatory systems in living things. In addition, high single-molecule resolution helps decompose very small number of proteins in bodies, and benefits the application in disease diagnostics [1]. Besides, many diseases are closely related to abnormal protein primary structure. For instance, Alzheimer's disease (AD) is a serious neurodegenerative disease, and the current research reveals that amyloid polypeptide is one of the pathogenic factors of this disease, which can lead to the formation of senile plaques in the brain and the apoptosis of nerve cells. Now researchers are using the nanopore protein sequencing technology to study the initial structure of amyloid polypeptide and distinguish promoters and inhibitors of it [2]. Overall, it is anticipated that

protein sequencing would enhance the foundations of our knowledge of all organisms radically and have substantial applications in the detection of disease.

Recent years, in addition to traditional protein primary structure measurement methods like Edman degradation or mass spectrometry, some new protein sequencing technologies have emerged, such as nanopore protein sequencing. Nanopore single-molecule protein sequencing is developing as a field, providing unheard-of prospects for basic research and diagnostics. It can also be used as a general technique to examine biomolecules, similar to gel electrophoresis. In the field of sequencing analysis, DNA sequencing based on nanopores continues to advance and gradually matures, then this method is extending to additional aspect such as proteins, metabolites and other analytes. Although DNA and RNA sequencing has made remarkable strides, protein sensing technique based on nanopore is still in its infancy and faces difficulties specific to proteins and proteomics.

Traditional sequencing technologies of protein include Edman degradation and Mass spectrometry, which can effectively determine the sequence of proteins and are widely used. In the middle of the 20th century, Edman and others realized protein sequence analysis through a series of chemical reactions, usually called Edman degradation. At the end of the 1960s, automatic sequencers based on this principle began to appear. In the 1990s, mass spectrometry technology began to be used for protein sequence analysis. However, it is still a challenge to use these classical methods to determine protein sequences at very low concentrations.

2. Edman Degradation

Edman degradation is a classical chemical method for protein identification, which can accurately analyze the Nitrogen-terminal sequence of protein. Using chemical reaction, such as adding phenylisothiocyanate (PITC), the peptide can be labeled and cut at the N-terminus, and the cycle of the reaction is repeated for identification of the next amino acid. The whole process is shown in Fig. 1. Under normal circumstances, the first step is to add PITC to the N-terminal of a peptide/protein, and this process is generally carried out under alkaline conditions. Secondly, under acidic conditions, the PITC conjugated N-terminal amino acid cleaves to produce a thiazolinone derivative. Then it is converted to the more stable state called phenyl thiohydantoin derivative of amino acid residue (PTH-AA), and it can be recognized by chromatography. Meanwhile, to identify the following amino acid, the PITC reaction cycle is performed.

This method has low sensitivity and it is difficult to determine the peptide chain with more than 30 kinds of amino acid accurately. Furthermore, the yield of sequence decreases gradually over several cycles, while the amount of stagger raises little by little [3]. At the same time, the amount of background noise increases with the expansion of cyclic reaction, which results in that the sequence will become unintelligible when the sequence signal equals the background noise level.

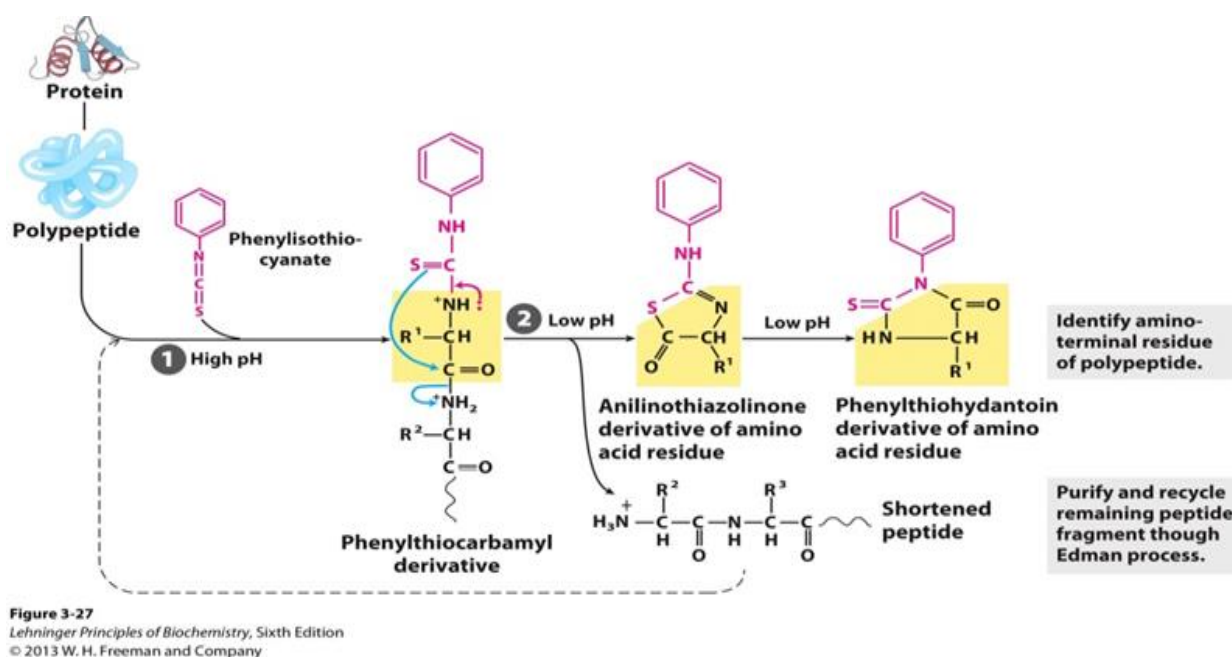


Figure 1. The process of Edman degradation [4]

3. Mass Spectrometry

At present, mass spectrometry is the most popular method used to determine the primary structure of proteins. There are two main ideas: bottom-up proteomics and top-down proteomics analysis method [5]. The former is the mainstream technology of mass spectrometry sequencing, where protein is broken into small polypeptide fragments by digestion or enzymatic hydrolysis then analyzed by Mass spectrometry. Order information of polypeptide chain can be identified after comparing with the known protein sequence. The top-down mass spectrometry principle starts with the whole protein molecule. By the use of electrospray or matrix-assisted laser desorption ionization and other technologies, the whole protein molecule can be led into the mass spectrometry for protein dissociation and tandem analysis.

Mass spectrometry is difficult to identify proteins with low content, and each fragment may be indistinguishable due to close quality [6-8], reducing the accuracy of analysis. Moreover, current mass spectrometers still have some drawbacks, for example, there will be a defect when reading sequenced peptide length, and deviation of analysis caused by ionization mechanism is a big problem as well.

4. Nanopore

With the development of science, researchers find some new technology. Nanopore single molecule detection and analysis technology developed from the mid-1990s. It is easier to use and label-free. At the same time, it has higher sensitivity and faster detection speed than previous technologies. This method has great application value and development potential in protein detection, gene sequencing and characterization of nanoparticles. In this paper, the electrical nanopore sequencing will be introduced in detail. Nanopore is a small pore structure with nano scale aperture. From the development history of sequencing technology, it is evident that nanopore single molecule detection technology is the future development trend. Furthermore, in comparison with mainstream sequencing platforms, nanopore protein sequencing has irreplaceable advantages, which will continuously promote the development of biomolecular detection technology towards high-precision, high-throughput, low-cost.

According to various carriers, there are two different types of nanopore. One of them uses protein with micropores such as α -Hemolysin [9], which is called biological nanopore. The other one is solid nanopore [6], it is generally prepared by drilling holes in ultra-thin materials through special processing technology, and the selected materials are two-dimensional materials such as silicon nitride or graphene [10,11]. Compared with biological nanopore, it is challenging to precisely manage the solid nanopore's pore diameter, and the preparation cost is high. However, it is more stable.

Electrical sensing is the most common and popular method in nanopore detection technology, the detection principle of it is mainly the ion blocking current principle. The whole process is as followed. As shown in Fig. 2, although the measured molecule is DNA, the process is the same with protein sequencing. When supplying two sides of the nanopore with a transmembrane voltage, charged nanoparticles (such as ions) will form ionic currents through the nanopore under the action of electrophoresis. When the analyte is put into the cis compartment, the electric current goes through the nanopore one after another under electric field, and will produce a transient blocking current signal that changes with time. The blocking degree of ion current depends on the number or volume of ions which was squeezed out of the hole by measured biomolecules. Each perforation can be characterized by the duration and blocking degree of the current, so the characteristic signal of the protein can be obtained, which means, the size and spatial configuration of the analyte can be inferred by statistical analysis [2,12].

However, proteins are more complex, resulting in many improvements. Compared with nucleic acids, peptides are not easy to generate electrical signals. And the components of proteins are diverse amino acids instead of four nucleobases. In particular, proteins are constructed by a steady three-dimensional folded formation and multifarious combination mode, making the assignment of linking the ionic signals with the amino acid order more sophisticated. At the same time, for the determination of protein primary sequence, it is often necessary to expand the high structure of the protein and cut off the peptide chain appropriately. In response to these problems, many research teams have proposed solutions. For example, Aksimentiev and Oukhaled team researched out aeromonas lysin protein nanopore channel as a peptide sequencing method, which can realize the determination of 20 amino acids [13]. In this method, the core point is to use polycation carrier (arginine hexapeptide) to load a single target amino acid. This design can make the target molecule stay in the sensing region of the nanopore for enough time, helping produce obvious changes in the cross-pore current for corresponding analysis. Meanwhile, it will have better effect to combine with the sequential cutting technology of peptide bonds. At present, nanopore technology has significant progress and important application in the field of biomolecular detection. Research proofs that Aerolysin nanopore has ultra-sensitive measurement ability for phosphorylated peptides, and it can complete the detection of single molecule activity of PKA kinase [14]. Aerolysin nanopores can also be used to explore the perforation of non-folded proteins [15]. In addition, Mayer's team proposed a measurement technology based on phospholipid membrane nanopore to analyze the whole protein molecule [16]. The size of this phospholipid nanopore is approximately 30 nm in diameter and 300 nm thick. When a single protein molecule passes through the nanopore, the volume, shape, charge, spin speed and ability to interact with specific molecules of it will have an impact on the size of the blocking current across the pore. By capturing this information, the characteristic signal of the protein can be obtained. Because each molecule needs to be measured separately, the throughput of this method is extremely low.

As a consequence, breakthroughs can be made in the following aspects. First of all, achieving high throughput is a big challenge. There is no doubt that each cell unquestionably includes billions of proteins. Consequently, in order to resolve all of the proteins and enable comprehensive analysis in all aspects of cells within a realistic timeframe, future technology must achieve an extraordinarily high sensing throughput [1]. In addition, accurate recognition is also important, some documents explored that modifying the nanopore to make it react with the amino acid residues may enrich the information of the electrical signals, and then realize the recognition of residues through artificial intelligence [17,18]. In the future, it is expected to achieve these improvements by nanopore arrays and other technologies.

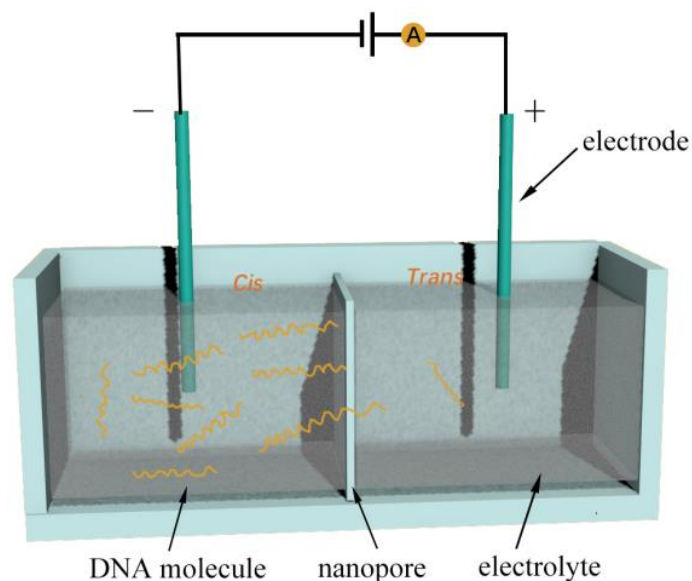


Figure 2. The process of ionic current blockage method in electrical nanopore detection [19]

5. Conclusion

At present, the mainstream protein determination technology still has some limitations such as low sensitivity and database dependence. A new generation of protein sequencing methods need to be developed. This paper briefly introduces two common classical protein sequencing methods, and focuses on the advanced topic of nanopore sequencing. Despite the disadvantages, the technique of single molecule nanopore for proteins has made great progress in the respect of detection, identification and sequencing. And it is acknowledged that this technology will develop into a highly active and prosperous research field, which also demonstrates strong potential in medical diagnostics and fundamental biosciences. In view of the current challenges, some scholars assert that the selectivity and sensitivity of channels can be improved by controlling the movement of biomolecules and modifying the surface of nanopores. Altogether, it is hopeful that in the near future, low-cost, high-throughput and highly sensitive protein sequencing methods and instruments can be found and popularized, providing new power for the development of protein related scientific research and medical treatment.

References

- [1] Alfaro, J. A., Bohländer, P., Dai, M. et al. The emerging landscape of single-molecule protein sequencing technologies. *Nat Methods* 18, 604 – 617 (2021).
- [2] Lin Z H, Chen H, Dong J W, Zhao D H, Li L B. Nanopore-Based Biomolecular Detection. *Progress in Chemistry*, 2020, 32 (5): 562 - 580.
- [3] Smith, John Bryan. *Peptide Sequencing by Edman Degradation*, 2001.
- [4] Creative-proteomics Blog, <https://www.creative-proteomics.com/blog/index.php/protein-sequencing-of-edman-degradation/>, January 25, 2018.
- [5] Chait B T. Mass spectrometry: Bottom-up or top-down? *Science*, 2006, 314: 65 - 66.
- [6] Chen H K, Hao R, Tian R J. Advances in Protein Sequencing Technology. *Chinese Science Bulletin*, 2021, 66 (25): 3309 - 3317.
- [7] Siuti N, Kelleher N L. Decoding protein modifications using top-down mass spectrometry. *Nat Methods*, 2007, 4: 817 - 821.
- [8] Han X, Jin M, Breuker K, et al. Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science*, 2006, 314: 109 - 112.

- [9] Branton D, Deamer D W, Marziali A, et al. The potential and challenges of nanopore sequencing. *Nat Biotechnol*, 2008, 26: 1146 - 1153.
- [10] Deng T, Li M, Wang Y, et al. Development of solid-state nanopore fabrication technologies. *Sci Bull*, 2015, 60: 304 - 319.
- [11] Schneider G F, Kowalczyk S W, Calado V E, et al. DNA translocation through graphene nanopores. *Nano Lett*, 2010, 10: 3163 - 3167.
- [12] Venkatesan B M, Dorvel B, Yemenicioglu S, Watkins N, Petrov I, Bashir R. *Adv. Mater.*, 2009, 21: 2771.
- [13] Ensslen T. Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat Biotechnol*, 2019, 38: 176 – 181.
- [14] Meng F N, Ying Y L, Yang J, Long Y T. *Anal. Chem.*, 2019, 1: 9910.
- [15] Pastoriza-Gallego M, Rabah L, Gibrat G, Thiebot B, van der Goot F G, Auvray L, Betton J M, Pelta J. J. *Am. Chem. Soc.*, 2011, 133: 2923.
- [16] Yusko E C, Bruhn B R, Eggenberger O M, et al. Real-time shape approximation and fingerprinting of single proteins using a nanopore. *Nat Nanotechnol*, 2017, 12: 360 - 367.
- [17] Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature*, 2016, 537: 347 - 355.
- [18] Tyanova S, Temu T, Sinitcyn P, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*, 2016, 13: 731 - 740.
- [19] Feng Y, Zhang Y, Ying C, Wang D, Du C. *Genomics Proteomics Bioinformatics*, 2015, 13: 4.