SEQUENCING

Next-generation peptide sequencing

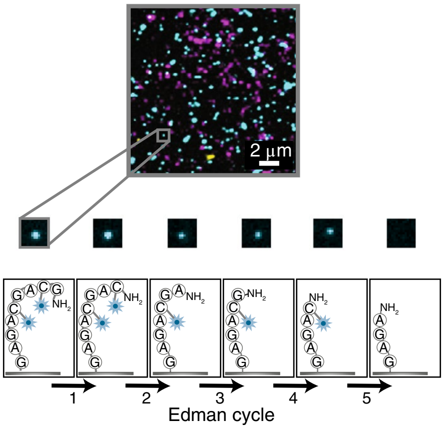
The concept of massively parallel single-molecule protein sequencing emerges.

DNA-sequencing technology has been rapidly expanding the horizons in genomics and transcriptomics space. The development of an equivalent protein-sequencing technology, however, has been a challenge. Proteins often exist in complex mixtures with dynamic abundance ranges and contain post-translational modifications, which can critically obstruct the identification and quantification of low-abundance proteins.

To date, shotgun proteomics coupling mass spectrometry with high-performance liquid chromatography has been the primary technique for high-throughput protein identification. However, mass spectrometry often fails to identify rare proteins present in a complex sample because of its low sensitivity. “There is no PCR for protein amplifications,” says Edward Marcotte. His group at the University of Texas at Austin has had a long-standing interest in proteomics. “We would like to have a much more sensitive technology and much deeper quantitation,” Marcotte says.

In 2014, Marcotte and his PhD students Jagannath Swaminathan and Alexander Boulgakov computationally explored the feasibility of single-molecule-based peptide sequencing (Swaminathan et al., 2015). They proposed peptide sequencing by stepwise degradation, in contrast to DNA sequencing by synthesis. In their method, proteolytically digested peptides are fluorescently labeled at designated amino acid positions and then subjected to sequential removal of amino acids from the N terminus.

When peptides lose amino acids one by one, such as by Edman degradation, a total internal reflection fluorescence (TIRF) microscope can be used to monitor the consecutive fluorescence changes for millions of peptide molecules in parallel. This allows one to determine the positions of the fluorescent amino acids on the individual peptides. This partial position information can be matched to a sequence in a database reference to infer the identity of the peptide, and ultimately its parent protein. The key question was how sparse the partial position information can be while still uniquely identifying a peptide from a mixture. Using Monte Carlo simulations, Marcotte and his students demonstrated that a four-color code would be able to theoretically identify more than 98% of the total human proteome.



TIRF imaging monitors fluorescent changes as amino acids are sequentially removed via Edman degradation. Reproduced with permission from Swaminathan et al. (2018), Springer Nature.

Notwithstanding the promising simulation results, the implementation of this single-molecule fluorosequencing platform took “three PhD dissertations,” notes Marcotte. “A lot of the difficulties are labeling chemistry, peptide surface immobilization and image processing,” he recalls. A collaboration with Eric Anslyn, whose expertise is in organic chemistry, eventually transformed the technology from an in silico result to an experimental platform. Anslyn and his PhD student Erik Hernandez established how to efficiently label amino acids such as cysteine and lysine, and with Swaminathan they pinpointed the fluorescent dyes that can survive the harsh reagents used for Edman degradation.

In their recent *Nature Biotechnology* paper (Swaminathan et al., 2018), they describe the method of single-molecule fluorosequencing with a two-color code. The method shares upstream peptide fragmentations with shotgun mass spectrometry and similarly relies on a database matching scheme. They showcased the sequencing capacity on synthetic and naturally derived peptide molecules, and also demonstrated sequencing of phosphorylated sites through the use of fluorescently labeled phosphoserines.

The key strengths of single-molecule fluorosequencing are its high sensitivity and the potential for digital quantification. “The paper is the first experimental description of, analogous to DNA sequencing, a short-read, quantitative, single-molecule protein sequencing platform,” Marcotte says. The platform currently allows sequencing of millions of peptide molecules in parallel, with a sub-million-times improvement in sensitivity compared with that of mass spectrometry. “It is not unreasonable to scale up to billions of reads per run,” Marcotte notes.

The two-color code is sufficient to uniquely identify most proteins in mixtures comprising several hundred human proteins. Yet, more colors are required for high sequencing coverage of the human proteome. “But we are not going to get all amino acids labeled,” Marcotte remarks. “So the way we have done that is not a full de novo protein sequence approach. It will require a reference database.” Similar to the case for DNA sequencing technology, alternative technologies are under development, such as nanopore-based ‘long-read’ sequencing by recognition tunneling, which may provide a potential method for de novo protein sequencing.

For sequencing of complex proteomic mixtures, further improvements in sample preparation for rare proteins, labeling chemistry, and data interpretation are ongoing. The exciting news is that Marcotte and colleagues plan to commercialize the instrument under the umbrella of a startup company. “We hope that this technology will be accessible to any research lab,” Marcotte says.

Lei Tang

Published online: 30 November 2018
https://doi.org/10.1038/s41592-018-0240-7

Research papers

Swaminathan, J. et al. A theoretical justification for single molecule peptide sequencing. *PLoS Comput. Biol.* **11**, e1004080 (2015).

Swaminathan, J. et al. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* **36**, 1076–1082 (2018).