Fluorosequencing: Concept, Features and Benefits

Jagannath Swaminathan, Cofounder of Erisyon Inc. – jagannath@erisyon.com

- **Erisyon introduces the platform technology, fluorosequencing, for single molecule protein sequencing.**
- **Technology is geared to identify and quantify peptides and proteins in complex biological mixtures at an unprecedented sensitivity.**
- **Fundamental concept: Positional information of a few select amino acids on a peptide is generated and provides sufficient information to pattern match against proteome database to identify the source proteins.**
- **Platform architecture: Highly parallelized in-place sequencing of peptides (up to 1 billion) tethered to a glass slide and measuring the changes in fluorescence of individual molecules using microscopy.**
- **Technical features: (a) Single molecule sensitivity (b) Throughput scaleable to a billion peptides (c) characterize highly heterogeneous samples and (d) Digital quantification.**

Concept of the technology

Fluorosequencing, the single molecule protein sequencing technology is a new, alternative method that can address limitations of sensitivity and quantification of current proteomics technologies, namely mass-spectrometry and affinity based methods. The concept of fluorosequencing takes inspiration from advances in DNA and RNA sequencing technologies, which sequence millions to billions of oligonucleotides in parallel to identify and digitally quantify specific nucleic acids in complex mixtures. By analogy, a successful single molecule protein sequencing technology would offer more than a million-fold improvement in sensitivity over conventional technologies while also allowing millions of distinct peptide molecules to be sequenced in parallel. This would enable peptides or proteins in the sample to be identified and digitally quantified by direct counting, opening the door to many new proteomics applications.

The key concept is shown in **Figure 1** (described in Nature Biotechnology article[1]) is as follows: First, one or more amino acid types on the peptides are selectively chemically labeled with specific identifier fluorophores. Millions of individual peptides are then immobilized on a glass cover slip. Next, each molecule's fluorescence is monitored using total internal reflection fluorescence (TIRF) microscopy following consecutive rounds of amino-terminal (N-terminal) amino acid removal by Edman chemistry. The sequence positions of the labeled amino acids are thus identified for each peptide molecule, providing a partial sequence for each molecule. These fluorescence signatures ("fluorosequences") can then be compared to a reference proteome for assignment to their proteins of origin.

Positional information of few select amino acids is sufficient to identify source protein

Since the advent of next-generation sequencing of nucleic

acids, almost all proteins have reference sequences produced from the cell's DNA or RNA. In fact, in nearly all proteomics experiments performed today, proteins are identified/quantified by either matching to these reference databases (in the case of mass-spectrometry) or the a priori knowledge of the antibody/reagent specificity to a protein (in the case of affinity based methods). Thus instead of having to identify every single amino acid in the sequence, we leverage the database of potential protein sequences to infer identity. In our technology, knowing the positions of only a few amino acid types generates a partial sequence that is matched to a reference database for protein identification. To further clarify on the power of positional information, an analogy can be made with the "Wheel of Fortune" game show, where knowing the positions of a few letters can be sufficient to identify the words and phrases. In our theoretical paper simulating the technology[2], we observed that, under ideal conditions, **labeling 4 different amino acid types was sufficient to identify up to 95% of the proteins in the human proteome** (see **Figure 2**).

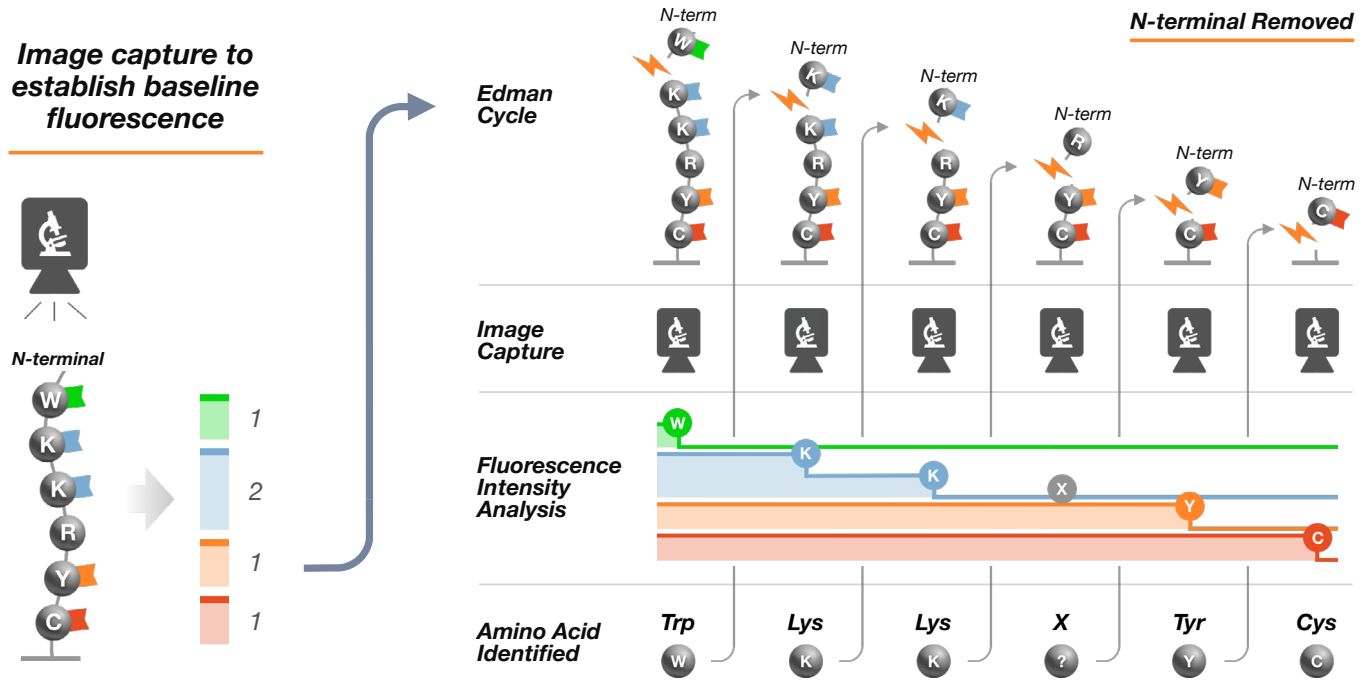

Figure 1. Overview of fluorosequencing – Peptides are covalently labeled with amino acid–specific fluorescent dyes and immobilized in a TIRF single-molecule-microscope stage perfusion chamber. Through TIRF, each peptide is imaged, and its N-terminal amino acid is chemically removed via Edman degradation, thus leaving each peptide one amino acid shorter and regenerating its free N terminus. Alternating cycles of chemistry (each removing one amino acid) with imaging reveals the amino acid positions of fluorescent labels within each molecule. Millions of individual peptide molecules can be analyzed in parallel at reasonable attachment densities, as shown for approximately 3 million peptides in an approximately 1.3 × 5 mm area of the coverslip. (Bullet indicates TMR conjugated to cysteine; diamond indicates Atto647N conjugated to cysteine; gold nanowires serve as fiducial markers.)

Technical features and benefits

The fluorosequencing technology offers four major features/benefits -

1. **Single molecule sensitivity:** The ability to image and sequence individual peptide molecules makes the technology immediately 1 million times more sensitive than the current peptide mass-spectrometers. This offers the benefit of using extremely low sample

amounts, such as clinical biopsies. This smaller sample requirements, assuming a factor of at least 1000, roughly translates to a difference between needing 1/4th of a pancreas or sampling with a needle for biopsy material.

2. **Massively parallel architecture:** The ability to sequence and independently measure fluorescence from individual peptide molecules tethered to an imaging surface, provides a scalable architecture to achieve a throughput to identify billions of peptide molecules on a single glass slide.
3. **Characterize highly heterogeneous samples:** Since every single molecule measurement is independent of another, fluorosequencing platform can discriminate peptides and proteins across a large range of heterogeneity and abundances (upto 106 range). This means that if there is a single peptide, mixed with 1 million copies of a different peptide, the two peptides can both be identified. This helps identifying peptides existing in an extremely diverse backgrounds, such as antigens on tumor surfaces or discriminating low abundant phosphorylation event on a protein.
4. **Digital quantification:** Unlike other proteomic technologies, where peptides and proteins are quantified relative to a standard (often using known concentrations of itself), quantification in fluorosequencing is done by counting molecular observations. This type of counting statistics makes it possible to compare abundances between proteins in the same and across experiments without need for external calibrants.

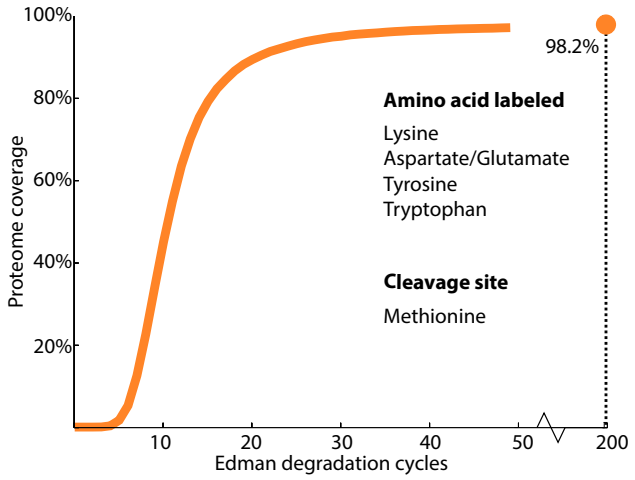

Figure 2. Positional information of amino acids can identify majority of proteins in a mixture – *Simulation indicates that labeling 4 amino acid types (Lys, Asp/Glu, Tyr and Trp) and use of Cyanogen bromide (Methionine cutter) would identify 95% of all proteins in a human proteome from ~20 Edman cycles.*

Conclusions

Fluorosequencing is a single molecule protein sequencing technology that offers the capability to catalogue and quantify every single peptide and protein molecule present in a biological sample. The technology opens up many applications in biology, such as characterizing the antigenic peptides on clinical tumor biopsies for cancer vaccine therapies, single cell proteomics or early detection of biomarkers in neurodegenerative diseases.

References

1. Swaminathan, J. *et al.* Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* (2019).
2. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. *PLoS Comput. Biol.* 11, e1004080 (2015).