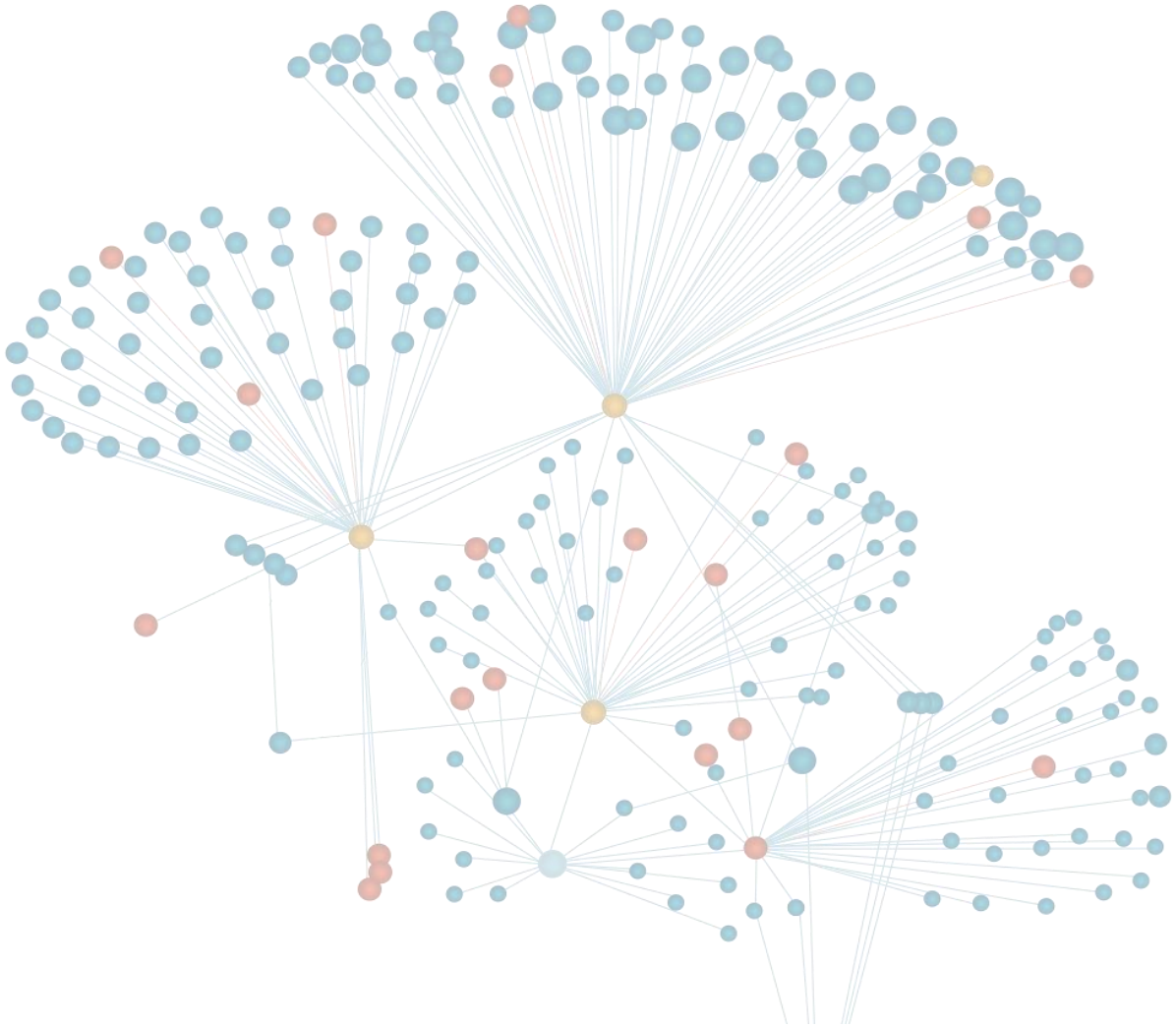
nautilus biotechnology

Proteomics at Scale: Current Approaches and Emerging Technologies



The promise of proteomics

From the day you are born to the day you die, thousands of proteins in your body are controlling and determining the behavior of your cells. While our DNA contains the blueprint for our biology, the proteins are the molecular actors doing the physical “work” within our cells, and it is these proteins that are ultimately responsible for carrying out nearly all molecular functions. The proteome is the complete set of proteins present in our cells, and it can vary widely across the different tissues in our bodies, changing and shifting during both healthy or disease states every second. This makes the human proteome one of the most dynamic and valuable sources of direct biological information available.

Proteomics, or the study of the proteome, allows us to understand how cells perform everyday functions and how they become altered in disease. There lies a unique opportunity in proteomics to uncover novel biology, clinically meaningful biomarkers, and new drug targets beyond what has been possible with genomics alone. Gaining proteomic insight on the countless diseases that cannot yet be treated or cured offers the potential to open up a new era of personalized and predictive medicine.

With more than 95% of FDA-approved drugs[1] currently targeting proteins, proteomics has historically been a common focus of research to develop new therapies and diagnostics. However, the number of new drugs approved each year has failed to increase proportionally to the nearly three-fold surge in global R&D spend since 2002.[2] Fewer than 2% of drugs are projected to return the investment[3] that went into creating them. Additionally, proteomic biomarker discovery has failed to meet expectations.[4] The current industry approach to measuring the proteome must be reevaluated and reimagined.

Proteomics has been limited by several challenging issues. First, unlike DNA and RNA, the biophysical characteristics of proteins (size, charge, hydrophobicity, etc.) are extremely diverse. This diversity makes proteins very difficult to biochemically “read” in the same way we can read DNA or RNA. Second, the physical amount of each unique protein found within samples, such as tissue extracts and biofluids, is present across an incredibly wide range of expression: some proteins are hugely abundant with hundreds of millions of molecules within a sample, while others are exceptionally rare with as few as a handful of molecules. This range of abundances is orders of magnitude wider than either DNA or RNA, and is commonly referred to as the “dynamic range” problem. This problem is further exacerbated by the fact that proteins cannot be amplified, making it more challenging to comprehensively quantify the proteome with the sensitivity and precision that is possible in genomics.

Although decades of research in genomics have provided a wealth of information on the genes relevant to disease onset and progression, a gene’s copy number or transcriptomic expression is not typically well correlated with its protein abundance.[5] Additionally, existing proteomics tools are unable to quantitatively measure the entire proteome, leaving much of it unexplored and inaccessible to biomedical research.

Approaches including mass spectrometry, affinity-based assays, and peptide sequencing have taken initial steps toward quantifying the proteome. Still, only a fraction of the human proteome — which comprises as many as 20 thousand or more proteins[6] — can be routinely measured by current methods. In addition, protein variations, including sequence variations, splice isoforms, and post-translational modifications, create millions of proteoforms[7] that, despite being biologically relevant, are largely invisible to these types of proteomic analysis tools.

We will outline the primary advantages and challenges of these three main approaches for proteomic analysis, as related to scale, depth, sensitivity, reproducibility, and ease of use for the purposes of comprehensively measuring the human proteome and furthermore, discuss the measures needed to enable quantification of the proteome at depths substantially greater than the 8-30% of the proteome that is readily accessible today.

Mass spectrometry-based proteomics

Mass spectrometry is a widely used approach for broad-scale protein analysis and is employed for a variety of basic science, drug development, and diagnostic applications.

However, typical mass spectrometry-based approaches are limited to routinely measuring approximately 8% of the proteome from blood and 30% of the proteome from cells. Mass spectrometric approaches often make trade-offs between ease of use, throughput, coverage, and sensitivity, preventing studies from easily, reproducibly, and rapidly measuring many proteins across a wide range of abundances. Despite being one of the most common approaches to protein analysis, the technology is complex and time-consuming, with workflows and processes that require highly specific skills and training.

Additionally, shotgun mass spectrometry approaches, the most prevalent approaches used in broad-scale protein analysis today, all operate by detecting only fragments of proteins known as peptides, that are generated by first breaking proteins down into smaller pieces. The process of analyzing peptides versus intact, full-length proteins complicates data analysis and protein quantification. Furthermore, information about the context of post-translational modifications that give rise to proteoforms is lost, further limiting studies from effectively characterizing the entire proteome.

Affinity-based proteomics

Affinity-based approaches can offer measurements of the proteome with improved sensitivity and specificity over mass spectrometry techniques through the binding of antibodies or other affinity reagents, such as aptamers, directly to a protein target of interest. These technologies can be used for bulk protein measurements, but by design they all require an affinity-binding reagent that must be highly specific to an already known protein target.

Performance of affinity-based technologies is directly linked to the availability of high-quality, highly specific, and sensitive affinity reagents, which typically limits the breadth of a study to fewer than hundreds of proteins in a sample. Simultaneous targeting of multiple proteins requires a specific reagent for each individual target, which ultimately prevents these technologies from scaling to the entire proteome. Furthermore, cross-reactivity between reagents can hinder quantitative accuracy.

Like peptide-based measurements such as mass spectrometry, bulk affinity-based measurements also limit proteoform deconvolution. This further hinders the ability of these technologies to detect and quantify the complexity of the proteome with respect to precise isoform abundance and patterns of post-translational modifications.

Peptide sequencing

Detecting proteins through peptide sequencing and assembly is not new. The initial techniques based upon Edman degradation date back to the 1950s and typically required a substantial amount of pure peptides. A number of new approaches have recently emerged that seek to read amino acid sequences, or portions thereof, with higher sensitivity and parallelism than historical methods. By increasing the number of peptides that can be sequenced in parallel and within more complex mixtures, these methods aim to address ongoing obstacles to sensitivity, breadth, and quantification in proteomic analysis.

However, the requirement that many peptides per protein must be sequenced (potentially hundreds) subjects the approach to a limit on the ultimate ability to scale at high throughput. Additionally, the error-rates, practical length limitations, and biases that are encountered in the field of genomic sequencing may also be applicable to peptide sequencing. Similar to mass spectrometry, peptide-level resolution generates only partial reads of each molecule. This same limitation may again constrict our view of important biological context and proteoforms visible only by detecting intact, full-length proteins present in a sample.

The future of proteomics

No existing technology using these three main approaches has been demonstrated to be capable of measuring the entire proteome. To deliver on its full potential, the field of proteomics will require a completely novel technology, similar to what was observed in genomics when routine DNA sequencing was achieved through completely novel approaches that delivered a massive improvement in data quality and scale. This in turn created access to easy, affordable, and widely available technologies for any lab around the world to read DNA.

Compared to DNA, the extreme diversity and complexity of proteins necessitate an extraordinary approach for enabling visibility into the proteome. The ideal system should be capable of identifying and decoding unique protein patterns at scale to fully comprehend the

processes and mechanisms that regulate all aspects of our physiology and meet the growing need for identifying and developing effective diagnostic and therapeutic targets.

With the easiest drugs for the least complicated diseases having already been created, the discovery of new drug targets is only becoming more critical to addressing rare and complex diseases and the future of our health. A dramatic acceleration in the understanding of the proteome comparable to that of the genome is urgently needed to discover key biomarkers, accelerate drug development, and tackle complex and rare diseases.

Technology development has focused on enhancing the existing technologies described above. But even after decades of incremental improvements to mass spectrometry, affinity-based methods, and peptide sequencing, they have not yet delivered on the promise of comprehensive proteome detection and quantification. Emerging technologies are addressing sensitivity and throughput in mass spectrometry, breadth in affinity-based approaches, and scale in peptide sequencing. However, the fundamental limitations described above may continue to restrict their ability to easily, effectively, and deeply measure the entirety of the proteome, which includes the diversity of protein modifications and proteoforms that are universally undetectable using such methods.

The Nautilus Platform

A map drawn with only a few lines is nearly incomprehensible. However, as more and more lines are drawn through exploration, the map becomes increasingly evident and detailed. We believe that the proteome can be similarly deciphered by assembling many individual data points to provide a comprehensive analysis of all the proteins in a sample.

The Nautilus Proteomic Analysis System is designed to be the first large-scale, single-molecule platform with the goal of achieving quantification of more than 95% of the proteome. Aiming to not only solve the singular technical challenges of previous methods but also ultimately enable an end-to-end process for measuring billions of individual protein molecules at a time, our technology is designed to preserve and measure full-length proteins and their unique proteoforms. The approach is centered around four key technologies:

1. A nanofabricated single-molecule protein flow cell that can capture billions of individual intact protein molecules for analysis across an extremely large dynamic range
2. An integrated multi-cycle optical and fluidics instrument to repetitively probe and interrogate individual protein molecules massively in parallel
3. A novel class of multi-affinity protein binding reagents that when used sequentially, are capable of decoding the identity of proteins at the single-molecule level
4. A machine learning software that improves the accuracy of protein decoding with growing sources of data, and is designed to decode greater than 95% of the human proteome using approximately three hundred cycles of multi-affinity probe data per system run

This technology platform has the potential to deliver unparalleled sensitivity and scale to proteomics, effectively establishing a new gold standard of accelerated speed, simplicity, accuracy, and ubiquity in the field.

Our vision

By changing the scale of what is possible in proteomics, we aim to unlock the proteome and deliver on its untapped potential to revolutionize how biological research is conducted, drugs are identified and developed, and human disease is treated. Beyond novel therapeutics and diagnostics are additional opportunities for bettering the human condition, from food and environmental sciences to basic science research.

In fully democratizing access to the proteome and all of its complexity, we can move toward understanding biology at the deepest level and enable new discoveries that can potentially positively impact the health of millions of people around the world.

References

1. Santos, Rita et al. “A comprehensive map of molecular drug targets.” Nature Reviews Drug Discovery 16(1): 19-34, 2017.
2. EvaluatePharma, *World Preview 2020, Outlook to 2026*, 2020.
3. Deloitte Center for Health Solutions, *Ten years on Measuring the return from pharmaceutical innovation 2019*, 2020.
4. Poste, George. “Bring on the biomarkers.” Nature 469: 156–157, 2011.
5. Tian, Qiang et al. “Integrated Genomic and Proteomic Analyses of Gene Expression in Mammalian Cells” Molecular & Cellular Proteomics 3(10): 960-969, 2004.
6. Adhikari, Subash et al. “A high-stringency blueprint of the human proteome.” Nature Communications 11: 5301, 2020.
7. Ponomarenko, Elena A. et al. “The Size of the Human Proteome: The Width and Depth.” International Journal of Analytical Chemistry 2016: 7436849, 2016.

info@nautilus.bio
+1 (206) 333-2001

@NautilusBio @Nautilus Biotechnology