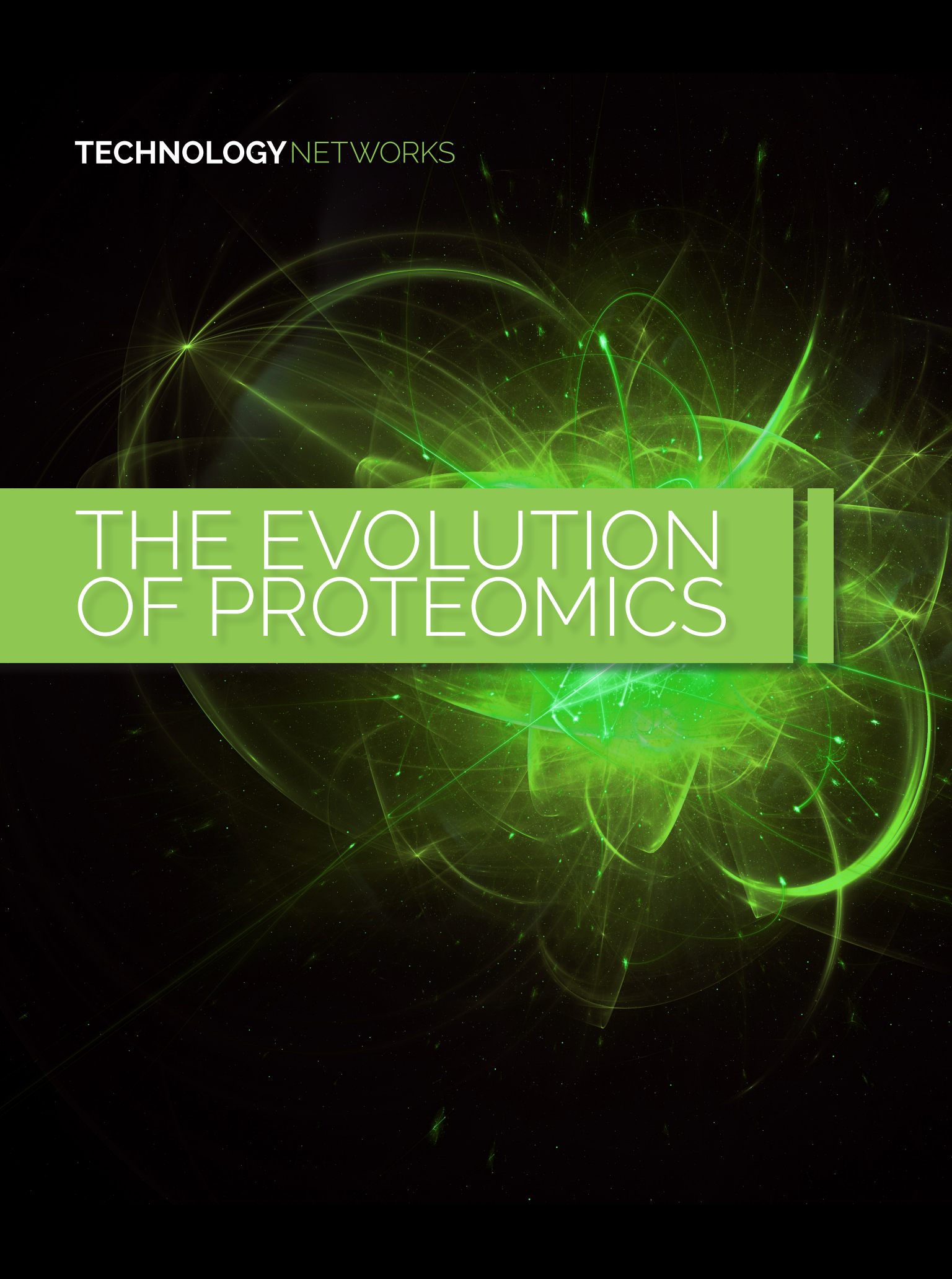


TECHNOLOGY NETWORKS

# THE EVOLUTION OF PROTEOMICS



# Contents

Professor Ruedi Aebersold	3
Dr Evangelia Petsalaki	5
Dr Mikhail Savitski	8
Professor Emanuel Petricoin	11
Dr Richard Scheltema	14
Professor Alexander Makarov	15
Dr Gary Kruppa	17
Professor John Yates	19

## Foreword

Welcome to our latest eBook, *The Evolution of Proteomics*. For a field still considered to be in its infancy, the applications of proteomics continue to grow. A wealth of research is currently available in areas such as personalized medicine, structural biology, biomarker research and proteogenomics.

To give focus to these exciting developments, *Technology Networks* conducted a series of exclusive interviews with world-renowned proteomics researchers to learn more about some of the greatest achievements in the field, the current state of play and to gain insights into the future of proteomics.

*The Evolution of Proteomics* eBook features the full collection of interviews.



# THE EVOLUTION OF PROTEOMICS

## Pioneering quantitative proteomics

**Ruedi Aebersold**



Ruedi Aebersold, former professor of systems biology at the [Institute of Molecular Systems Biology \(IMSB\)](#) at [ETH Zürich](#) and is regarded as one of the pioneers of proteomics research.

Aebersold has made significant contributions to the development of targeted proteomic techniques, including selected reaction monitoring (SRM) and data-independent acquisition. He is also one of the inventors of the Isotope-Coded Affinity Tag (ICAT) reagents used in quantitative mass spectrometry (MS).

Aebersold's research in quantitative proteomics has helped shape our understanding of how proteins function, interact and are localized in both normal and diseased states. [The Aebersold laboratory](#) utilizes high-throughput proteomic and computational methods, such as label-free shotgun proteomics, to precisely measure protein analytes in complex samples. By creating "snapshot" profiles, the research team are able to determine which cells contain abnormal levels of specific proteins, and by doing so, hope to develop novel diagnostic markers for disease. Since our interview, Aebersold has retired from active research, but he remains the head of the Tumor Profiling project at ETZ Zürich until the end of 2023.

**Molly Campbell (MC):** In your opinion, what have been some of the most exciting breakthroughs in the proteomics field since its conception?

**Ruedi Aebersold (RA):** We work on MS-based proteomics. For me the most fascinating aspect of this technique is its versatility. Essentially the same liquid chromatography mass-spectrometry (LC-MS/MS) technique and instrumentation can be used to explore the many different biologically important properties of proteins if some additional tricks are applied. These properties include, of course, the amino acid sequence and abundance of proteins, but also their half-life, state of modification, localization in cells, their participation in

complexes and the precise contact sites of interacting proteins. Recently, there has been a distinctive trend to also tackle the higher order structures and corresponding changes of proteins and protein complexes by techniques including hydrogen deuterium exchange (HDX), cross linking, correlation profiling, native MS, thermal profiling, limited proteolysis (LiP) etc. The information gained by many of these methods is frequently highly interesting and directly functionally relevant.

**MC:** Your current research in quantitative proteomics looks to compare levels of protein expression between samples. Can you tell us more about your recently published work in conducting proteomic profiling in different types of cancer for the discovery of new biomarkers?

**RA:** We have been doing quantitative comparisons between samples for 20 years, starting with the development of the ICAT technology in 1999. Out of that work we gained a lot of insights about the response of cells and tissues to different conditions. As an example, a PhD student, Ralph Schiess, discovered a set of plasma biomarkers to stratify prostate cancer with respect to diagnosis and treatment options. He then founded a company, [ProteoMediX](#), that is in the process of bringing this marker panel to the clinic. We also gained a lot of insights about specific biological processes, including their regulation by phosphorylation. As we could measure deeper into the proteome as the techniques evolved, we eventually learned that the response of cells to essentially any perturbation is very complex, typically involving hundreds of proteins.

This situation created a very challenging problem because it is not evident how a biologist would make sense of the resulting patterns and which of the many observations should be prioritized for what is commonly referred to as "biological validation". To overcome this essentially intractable problem, we decided to develop MS techniques that would allow us to quantitatively compare large numbers of samples (hundreds

to thousands) so that we could use mathematical methods like clustering, machine learning, statistical associations or regression to discover patterns indicating the biochemical changes in cells in a data driven way, rather than by prior biological knowledge. Out of these insights arose high-throughput targeting techniques and scoring software, initially SRM and tools like [mProphet](#) and a bit later [SWATH/DIA techniques](#) and tools like [openSWATH](#).

We are really excited about these techniques because they provide fascinating insights into the inner working of cells and tissues and open the door to population-based studies, for example by the use of genetic reference panels like the BXD mouse panel. By doing multi-layer measurements in such panels we try to combine genomic and proteomic data to learn how cells translate genomic variability into proteomic and eventually phenotypic variability. The same approach is also very powerful for clinical studies where the measurement of high numbers of replicates allow us to detect clinically significant signals, even in the noisy background of clinical samples.

### **MC: How important is data integration in proteomics research? How are advances in computational proteomics aiding data storage and dissemination?**

**RA:** These are really two different questions. The second is about data management and the first about relating the biological meaning of the data to other types of biological data.

Data management, including storage, dissemination and processing, pose significant financial and technical challenges because with advances in instrumentation, the data volume has also increased dramatically. It is not uncommon that a single study, e.g. population based studies as mentioned above, generate terabytes of data, a volume that is difficult to handle for many groups. Fortunately, cloud-based systems are becoming available and I would also like to highlight [Pride](#) and [the Protein Exchange Consortium](#) who have done an outstanding job of collecting and archiving data supporting published work and making the aggregate of data accessible back to the community, e.g., to support meta analyses.

The first question is even more challenging to address because in my view it is at present not clear how different data types generated from the same biological objects, e.g. cultured cells or clinical tissue specimens, are best integrated. There are rather straightforward methods such as correlation of data types (frequently, transcript vs. protein abundance) but the knowledge gained from those analyses is limited. There is an interesting discussion in the field as to whether strictly data-driven approaches like machine learning have an equal, higher or lower potential to discover properties of biological systems compared

to approaches that take into account the vast accumulated knowledge of biological processes. Personally, I came to the conclusion that for understanding the evolved biological systems we are studying, prior biological knowledge is highly useful and likely essential.

### **MC: You have worked on the development of several proteomic techniques. What technical challenges do researchers face in proteomics research?**

**RA:** For a long time, MS-based proteomic analyses were technically demanding at various levels, including sample processing, separation science, MS and the analysis of the spectra with respect to sequence, abundance and modification-states of peptides and proteins and false discovery rate (FDR) considerations. I think we are in, or are approaching, the exciting state where these challenges are reasonably well, if not completely, resolved. When we get there, we will be able to more strongly focus on creating interesting new biological or clinical research questions and experimental design, and tackle the highly fascinating question discussed above, how we best generate new biological knowledge from the available data. Personally, I am convinced that we will be most successful in this regard if we generate high quality, highly reproducible data across large numbers of replicates and it seems that proteomics is essentially at a point to achieve this.

### **MC: Your most recent paper adopted a multi-omics approach to explore heterogeneity in HeLa cells across laboratories. Why was a multi-omics approach advantageous over other techniques in this instance, and how significant were your findings for the field?**

**RA:** We undertook the study for two reasons. First, we wanted to make a fact-based contribution to the discussion about the reproducibility of research results in the life sciences, and second, we wanted to generate a presently unique multi-layered data set to explore how genomic variability affects the different layers of gene expression along the central dogma.

With respect to the first question, we found that HeLa cells used for experimentation in different labs are significantly different in their molecular makeup and that this different molecular make-up renders the cells phenotypically different. We also discovered that the cells cultured in the same lab change over time. These phenomena are the result of genomic drift. In combination with the results of some community benchmarking studies we and others have undertaken over the past few years to assess the technical reproducibility of various aspects of MS-based proteomic methods, we now conclude that proteomics has reached a state where the technical (and

reproducibility is very high. So, any potentially observed poor reproducibility of results is likely to be rooted in the complexities of biological systems. With respect to the second question we discovered that the quantitative results at each measured layer, i.e. the way and extent to which the cells respond to genomic alterations (a situation that is similar to that in cancer cells), correlate to some extent along the path of gene expression but not strongly enough to make one layer predictive of the other. We also discovered that the response to copy number variation in specific gene loci was significantly buffered at the level of protein complexes. Excess protein that is synthesized due to higher ploidy at a locus tends to be degraded if it cannot associate with its intended complex partners. This mechanism contributes to protein homeostasis at the level of the modular proteome.

**MC: Systems biology is evolving at a phenomenally fast rate. Having worked in the field for several decades, what do you envision for the future of proteomics?**

**RA:** I envision a vastly increasing significance of proteomics in systems biology, for two main reasons, both of which have been addressed above. The first is that proteomics has reached a level of maturity where large and high-quality datasets can be generated with relative ease and at a moderate cost. We have witnessed in the field of genomics that robust and accessible high throughput technologies are strongly transforming the life sciences. The second reason is that the different types of proteomic data which now can be generated contain a wealth of information that we have yet to learn how to completely understand. In short, biology and medicine are essentially about function and phenotypes, and these are strongly determined by the composition and modular organization of the proteome, a state that we describe with the term the proteotype.

*Ruedi Aebersold was speaking to Molly Campbell, Science Writer for Technology Networks.*

# THE EVOLUTION OF PROTEOMICS

## An omics approach to studying cell signaling

**Dr Evangelia Petsalaki**



**Evangelia Petsalaki, PhD is a Group Leader at the European Bioinformatics Group, where her research team study human cell signaling in health and disease conditions.**

The [Petsalaki group](#) uses interdisciplinary approaches, including data-driven network inference, modeling of cell processes and data integration to understand how different environmental or genetic conditions affect cell signaling responses leading to diverse cell phenotypes. Their aim is to make both predictive and conditional models so they can anticipate what might happen in a biological network under different conditions.

The research group also collaborate with experimental teams that specialize in MS, imaging and cell biology to enhance their data sets and validate their models. Such models are being designed to help researchers answer specific biological questions, such as how stem cells “decide” what type of cell they will become, and what is the effect of cell signaling on cell shape and migration (i.e. where it “goes” in a tissue or organ).

**Molly Campbell (MC): What do you regard as being the most exciting breakthrough in proteomics research since the field’s conception?**

**Evangelia Petsalaki (EP):** Proteomics in the last 10 years has been galloping on all fronts. I don’t think that there is a single most important breakthrough. Rather, the entire field has managed to develop technologies and methods that have allowed unprecedented views into the proteome of the cells, from a very large array of conditions and sample types, from cells, to patient samples, and everything in between. If I had to choose one, the [SWATH technology](#) developed in the [Aebersold group](#) at the Institute of Molecular Systems Biology at ETH Zürich, really provides in-depth quantitation of entire proteomes.

However, the technology I am most excited about is not quite ready to be called a breakthrough yet, but I expect it to be revolutionary in the future. I am talking about the work from Swaminathan et al, published in 2018 in *Nature Biotechnology* from the [Marcotte group](#) at the University of Texas at Austin. Using Edman degradation, they were able to identify proteins from protein mixtures. They still have a lot of issues to overcome before this technology works at scale, on protein lysates and is affordable. But when this is achieved, we are looking at a revolution in proteomics, where accurate, comprehensive proteomes and respective phosphoproteomes (and other post translational modifications) can be measured effectively, similar to the way that genomes, transcriptomes and epigenomes are measured now.

**MC: Your research group studies human cell signaling with the aim to understand what controls different cell responses. Why it is useful to study this area from an omics approach, particularly with a focus on phosphoproteomics and proteomics?**

**EP:** Cell signaling represents the set of processes that define how a cell will respond to perturbations in its environment or messages from other cells. These processes are critical in the cell and their deregulation leads to many diseases, including cancer, which is in fact largely a signaling disease. Because of their importance they have been studied for many years. Most of our knowledge comes from very detailed studies done a long time ago, where signaling pathways were discovered and annotated.

Since the “high throughput” era began we have made some additions to these pathways, but we are still heavily relying on these initial annotations. While they have provided amazing contributions to the field and our knowledge, there are two issues with them: The first one is that they represent the “average” pathway, however, cells respond differently to different conditions even if they activate similar “pathways”. Therefore,



assuming that pathways always have a specific structure regardless of the cell type or condition is an oversimplification. “Omics” approaches can help us fit these pathways to the observed data and adjust them, and even better, to use data-driven approaches to extract them directly from the data.

The second problem is that, as these pathways were discovered with very small and detailed studies, they cover a very small space of the actual signaling networks in cells. Omics data opens the door to exploring the rest of this space. As signals in the cell are transmitted largely through a relay of phosphorylation of proteins, proteomics and phosphoproteomics in particular, represent the actual signaling state of the cell at the time of measurement. They are therefore the ideal type of omics data to study this type of processes.

### **MC: As a computational lab, what approaches do you use to interpret phosphoproteomics data?**

**EP:** First of all, we aim to use data-driven approaches. This means using statistical approaches to extract patterns from the data without restricting it to what is already known about the system. The reason for this is that the majority of knowledge in cell signaling is accumulated around a handful of very well studied kinases and pathways. If you think of the cell as Europe, we only have a limited map for Portugal and a bit for Spain and the rest of Europe is uncharted. Currently, most studies try to venture out just a bit out of the map but still very close to the charted territory.

Since we collect the data from the entire cell (i.e. snapshots of the entirety of Europe), if we only restrict our study around previously known information, then we are ignoring an entire world of potential new discoveries. The other focus of the group is on integrating the phosphoproteomics data with other omics datasets that can provide information on other layers of cell regulation. To go back to the map analogy, imagine it being like getting different types of pictures of Europe, including the roads, the mountains etc. Integrating different types of information can give us a more complete picture of how cells work.

### **MC: What challenges do you encounter when handling proteomic data? How can these challenges be overcome?**

**EP:** The major two challenges are that the data is very sparse, and that we have trouble measuring low abundance proteins. So, every time we take a measurement, we sample different parts of the proteome or phosphoproteome and we are usually missing low abundance players that are often the most important ones, such as transcription factors.

In my group, one approach to mitigate this issue is to map the identified peptides on protein interaction networks and diffuse the signal on this network. This reduces the noise from spuriously identified proteins and enhances the functional signal. It also allows us to observe regions of the network that are highlighted by the different datasets and compare and study these, instead of trying to compare the sparse datasets between each other.

However, with the advances in MS technologies developed by many companies and groups around the world, including the Mann group at the University of Copenhagen, Aebersold and other emerging technologies that promise to allow “sequencing” proteomes, analogous to genomes, developed by the Marcotte group and colleagues, I expect that these will not be issues for very long.

### **MC: You recently published a paper titled “Allosteric Modulation of Binding Specificity by Alternative Packing of Protein Cores”. Your research group suggested that your findings could be used to engineer proteins with novel functions. Please can you expand on this?**

**EP:** This is a project completed during my postdoc time in Toronto and the lead on this is [Dev Sidhu](#), at the Donnelly Center of the University of Toronto. He is a wizard in protein engineering and has done very important work in the field. In this paper, we showed that modifying amino acid residues from the core of the protein provided conformational flexibility to the protein, resulting in changes in its ability to recognize specific ligands and even the binding site for these ligands.

This has direct implications for its function and its effect in the cell's functions. I am not a protein engineering expert but as far as I know, typically, modifications on the surface of the protein are used to modulate its ability to bind different ligands. Our finding shows that modifications in the core, can provide structural flexibility and therefore more options as a starting point for engineering specific binding properties. By understanding the effects that changes in the protein core have on the protein surface and its binding properties, we can engineer proteins to have additional or modified functions.

### **MC: What advancements would you like to see in the next 10 years in computational omics research?**

**EP:** I think that despite all the advances with data generation, analysis and integration methods, an approach or set of approaches to truly integrate these data and generate testable hypothesis to push the boundaries of our knowledge forward is still elusive.

I am excited about efforts to create whole cell models that are happening in different groups around the world, such as the [Covert](#) and [Karr](#) groups, in Stanford University and the Mount Sinai School of Medicine respectively, and others in Japan, and we are also joining that effort now.

I think that combining true data integration efforts with executable models of cell function will provide breakthroughs in our understanding of how cells work, what is wrong in disease, why different human cells (either same human different cell types, or same cell type and different humans) respond differently to drugs, and many other important questions in biology and medicine.

*Evangelia Petsalaki was speaking to Molly Campbell, Science Writer for Technology Networks.*



# THE EVOLUTION OF PROTEOMICS

## Deepening our knowledge of protein interactions

**Dr Mikhail Savitski**



Mikhail Savitski, PhD, is the Team Leader and Head of the **Proteomics Core Facility** at the **European Molecular Biology Laboratory, Heidelberg**.

The **Savitski laboratory** uses and develops stability proteomics for understanding the phenomenon of aggregation and disaggregation, cell phenotyping, and detection of protein interactions with drugs, metabolites, DNA and RNA.

Savitski has made several impressive contributions to the proteomics field, including the development of thermal proteome profiling (TPP); a technology that enables the identification of drug targets in situ on a proteome wide scale that has had a major impact in the world of drug discovery.

In a **paper** published in 2019, the Savitski lab used TPP to show the high affinity interactions of ATP as a substrate and as an allosteric modulator with widespread influence on protein complexes.

**Molly Campbell (MC): In your opinion, what have been some of the most exciting breakthroughs in the proteomics research field thus far?**

**Mikhail Savitski (MS):** For me this has certainly been the development of the Orbitrap by **Alexander Makarov**. This has had a tremendous impact on the field of proteomics. Development of multiplexed quantitative MS has also in my opinion had a tremendous effect on the kind of biological questions that can now be tackled by MS.

**MC: Your research looks at protein-drug, protein-metabolite and protein-protein interactions in the context of biological processes. Why is it important to study these interactions? What can they tell us?**

**MS:** Protein-drug interactions are essential to decipher in order to make progress in all types of disease treatments.

Whilst protein-protein and protein-metabolite interactions are aspects of more basic biology, deepening our knowledge of these interactions in healthy and diseased states will lead to new strategies for treatment of disease in the future and will, in general, enhance our understanding of how a cell works.

**MC: You developed the proteomics method TPP. Please can you tell us about the creation of this method, how it works, and its applications in the field of drug discovery?**

**MS:** TPP was developed when I still worked at Cellzome. Our focus there was on deciphering the mode of action of small molecule inhibitors. In 2013 a paper came out from the Karolinska Institute which extended the use of the thermal shift assay for detecting drug interaction with a recombinant protein to the cellular milieu. We realized that we could combine the principle of this cellular thermal shift with multiplexed MS and by doing so develop the first unbiased technology for detecting protein-drug interactions inside living cells across the proteome.

We were incredibly excited when the setup worked and, already, the very first experiments have provided novel insights into the off-target activities of drugs. Meanwhile, we have significantly improved this technology and also applied it to bacteria as well as for studying fundamental biological process such as the eukaryotic cell cycle. TPP is continuously becoming more and more widely used by other groups and pharmaceutical companies for finding targets and off-targets of drugs.

**MC: What challenges do researchers face in the development of novel MS workflows?**

**MS:** I think the main challenge is having the right ideas! The field is getting more and more interdisciplinary. My lab consists of biologists, chemists, biochemists, and bioinformaticians. Often people have no background in proteomics before they join us, but they are excellent and very often come up with exciting new ideas. So in short, my answer would be to surround yourself

with talented, diverse people and create the best working environment possible, then the great ideas will come!

**MC: Can you give an example of a completely unexpected finding that you have stumbled upon recently in your research?**

**MS:** Our most recent work focused on interactions of ATP—the most common metabolite in the cell—with the proteome. Remarkably we discovered that in addition to the known canonical roles of ATP, it also turned out to modulate the solubility state of a quarter of the solubility transitioning proteome—proteins that can transition from soluble to insoluble under physiological conditions—at physiological concentrations of the metabolite. A previous study had shown that ATP could prevent the aggregation of certain recombinant proteins, and we were incredibly excited to see these broad effects on the proteome-wide level.

**MC: Technologies and methods used in proteomics have evolved rapidly over recent years. What further advancements do you anticipate occurring in the next 10 years?**

**MS:** I think we will see a tremendous progress in functional proteomics over the next decade. By that I mean that the field will move beyond merely measuring the changes of expression of proteins but rather use quantitative proteomics in combination with innovative biochemical assays and computational methods to assess the functional state of proteins in different ways. This is already happening in the field, but I expect we will see much more of this in the future.

*Mikhail Savitski was speaking to Molly Campbell, Science Writer for Technology Networks.*

# THE EVOLUTION OF PROTEOMICS

## Driving applied proteomics

### Professor Emanuel Petricoin



Emanuel Petricoin is a Professor and Co-director of the [Center for Applied Proteomics and Molecular Medicine](#) at [George Mason University](#).

He has dedicated his career to driving the clinical proteomics field forward and advancing personalized medicine. Petricoin's research focus is on the development of cutting-edge microproteomic technologies, identifying and discovering biomarkers for early stage disease detection, developing novel bioinformatic approaches for protein-protein interaction analysis and creating nanotechnology tools for increased analytical detection, drug delivery and monitoring.

He is a founding member of the [Human Proteomic Organization](#) (HUPO), has authored over 40 book chapters and is on the editorial board of several publications including *Proteomics*, *Proteomics-Protocols*, *Molecular Carcinogenesis* and *the Journal of Personalized Medicine*. Petricoin is a co-founder of several life science companies and is a co-inventor on 40 filed and published patents.

**Molly Campbell (MC): In your opinion, what has been the most exciting breakthrough in proteomics research thus far?**

**Emanuel Petricoin (EP):** Beyond the advances in the technologies of MS, such as the Nobel prize winning work of [Koichi Tanaka](#), [John Fenn](#) and others for matrix assisted laser desorption/ionization (MALDI) and electrospray, I think that the automation and computational analytics software packages and overall workflow build-ons that have occurred in proteomics have been extremely exciting and have really pushed the field; especially in clinical proteomics where people want to see proteomics have a true clinical impact.

The underpinning technologies in proteomics have had to adapt for clinical application, going from low throughput and clunky

research technology to high-speed and high-volume technologies that produce experiments with high reproducibility and with low sample processing costs. This isn't done yet – it's still evolving.

Secondly, I think the explosive interest in multiple reaction monitoring and selective reaction monitoring- and panel-based assays using triple quad technology and other new tribrid high-resolution MS equipment is exciting. This is directly in the field of clinical proteomics where measuring discreet panels or discreet signatures is going to be useful as a clinical decision support tool, an early detection tool and as a monitoring tool for patient treatment response etc. The development of reference standards and standard operating procedures (SOPs) has also been tremendous for the field.

I think the third area that is the most exciting is (carrying on with the theme of looking beyond protein discovery to protein panels) is the development of robust protein array technologies and new types of multiplexed, histomorphological based proteomic analysis. In proteomics we're now intersecting with the geospatial era of not just how much of a protein there is, but where exactly in the tissue and cell it is.

We invented the [reverse phase protein array technology](#) in our laboratory over a decade ago, and that technology has exploded in world-wide usage, clinical implementation and pharmaceutical company interests. Out of all the proteomic technologies that I have been involved with, the reverse phase protein array technology has been accelerating the most rapidly and truly has an impact on patients, treatment selection and outcome. We're going beyond discovery and into robust clinical measurements in regulatory environments.

**MC: Your research explores personalized medicine using cutting-edge microproteomic technologies. Please can you tell us more about the development of these technologies and what benefits they offer in proteomics research?**

**EP:** Absolutely. One of the quandaries we face in the proteomics field is that there is no polymerase chain reaction (PCR) -like technology that the proteomics field can use to routinely amplify low abundance proteins. In the field of genomics, we can speculate that the PCR technology catalyzed, electrified or perhaps even inaugurated the genomics revolution. The inability to amplify low abundance molecules has meant that the proteomics space has lagged behind genomics. In proteomics, whatever you have in your sample is all you're going to get.

The reason I raise this point is that in the field of clinical proteomics and precision medicine, we're left with the daunting challenge of having both extremely small amounts of material in our sample to begin with and the desire to develop multiplexed assays. In this field, we're wanting to measure many different protein analytes that are becoming extremely interesting to physicians and pharmaceutical companies because they're the targets for so many drug compounds – take kinase inhibitors in oncology for example. It's problematic, therefore, that these analytes are extremely low in abundance and you have only a few hundred cells to begin with.

The proteomics field in the past was simply more research driven, and so had the luxury of beginning with experiments where growing trillions of cells in an incubator as the input for MS experiments or other proteomic techniques are routine. However, that luxury does not exist in the space of precision oncology, and clinical proteomics. In these areas, you're left with very small amounts of cells as your input because the input is typically surgical biopsies. The amount of material that a pathologist needs for diagnosis has dropped dramatically compared to say 10 years ago, and so there is a growing pressure in proteomics to match this standard and take even smaller biopsies for proteomic analysis.

The precision oncology space is exploding with therapeutics that specifically target proteins and not genes. So, how do we measure these drug targets effectively in a patient sample, under the auspices of being able to use this information for treatment guidance, stratification and to create predictive markers, when we only have maybe a few hundred to a thousand cells in the biopsy sample? We have had to develop micro-proteomic technologies to meet the demands of the clinical space, because the clinical space is not going to adapt for us and nor should they from a patient's perspective.

That was the underpinning motivation for us to develop the reverse phase protein array. We wanted to develop a tool that measures highly important proteins and phosphoproteins that are of extremely small abundance in tiny biopsy samples. This technology allowed us to enter a clinical space that otherwise was shut off to investigators and dominated by genomics, an area where you can measure DNA, RNA and microRNA in

very small amounts of material. The reverse phase protein array technology allows us to quantitatively measure hundreds of low abundance proteins and phosphoproteins from extremely small amounts of clinical material in a robust way.

We have taken this technology from an invention and graduated it all the way to clinical implementation as a **CAP-accredited** assay that can be used in a clinical trial setting to make patient treatment decisions. It's only because of its ability to measure such small amounts of material that has really allowed this technology to thrive.

### **MC: Why is it important to consider proteins as potential biomarkers in early disease detection?**

**EP:** A lot of people think that genomic based detection of disease is more desirable because of the ability to specifically measure a genomic alteration, DNA or RNA fragment, transcript, etc. from a pathogen or from the disease process itself. One of the reasons why people like the genomic detection methods is because there is some consideration that the genomic DNA or RNA is more stable in the body and doesn't degrade as much as a protein biomarker.

Of course, as we said, we have really sophisticated ways of amplifying these signals by PCR and other methods, so there are a lot of genomic based diagnostic testing and early detection tools currently being implemented. Again, it all ties back to abundance of the target analyte. Early detection means you're trying to detect the disease before it can be detected by any other means, early on. There is no reason to detect disease late, we want to detect it early to make a clinical difference, but if you detect a disease early it means that the amount of analyte that's coming from the diseased cell is going to be very low. So, you're going to have a technological wall that you meet when it comes to analysis of low abundance molecules.

However, proteomics has, in its advantage, the natural cellular amplification that can occur simply by what proteins normally do during the transcription to translation to expression process. Per cell, there is very often many more copies of a protein than there is the DNA that encoded it, making proteomics an intoxicating area to look for disease biomarkers.

A second reason is that there are already a number of FDA approved/cleared assays that measure protein biomarkers for early detection and are in use for a variety of diseases – this is a space already quite robust. We have molecules such as PSA for prostate cancer for example, troponin for heart disease detection, haemoglobin a1c for diabetes monitoring and detection and insulin monitoring. Insulin itself is a protein. When you start to think about it philosophically, the proteomic space has already kind of “owned” the diagnostic space for quite a while, however



these proteins are usually measured one at a time and aren't thought about as a proteomic multiplexed tool.

I would say the biggest issue in the early detection space is always specificity performance of your biomarker and not as much on the sensitivity (although you would like to be very good at both) because most diseases occur at low relative frequency thankfully, compared to other benign/inflammatory conditions that present coincident to the pathogenic process and occur at a much higher relative frequency. Considering this, you have to have a biomarker that is very specific to the disease to reduce false positives.

### **MC: Can you tell us more about your work in developing high-throughput proteomic sensing technologies and microfabricated biosensors?**

**EP:** One of the things we're trying to do is measure proteins in a way that has clinical relevancy. In our lab, for example we are working on identifying new protein biomarkers in saliva for traumatic brain injury.

We have developed some new technologies that could potentially go right into the mouthguard of say, an athlete, or even potentially a war fighter in the military, where the mouthguard basically has fabricated nanoparticles that can change color when a specific protein bind to them. That way you could detect a concussion for example by looking at the color change in the mouth guard. This technology is not ready yet, but this is where we're trying to go. In clinical proteomics it's not just about discovering the protein biomarker but also incorporating its measurement into devices.

One of the fields evolving in parallel with proteomics is the sample preparation field. There are technologies in sample preparation that are pulling along the proteomic space, and likewise, proteomics advancements are pulling along the sample prep field – they're inexorably linked. In any scientific field there is always a weak link that effectively "holds the field back". In proteomics one of the weakest links in the past was the sample preparation side. In MS for example, there have been a number of developments in the physics of the instruments themselves along with approaches to the upfront fractionation technology that often dilutes the starting protein concentration of a given sample. Fractionation is not concentration.

If you inject pure serum into a mass spectrometer, you pretty much just sequence albumin, whereas if you fractionate that serum sample beforehand you can see a whole universe of proteins that would not have been detectable before. These developments are all "sample prep" if you think about it philosophically. However, the problem is that none of these approaches are concentrating, they are simply fractionating. We

need concentrating and fractionating at the same time. Lots of new technologies are trying to do that, for example there are new types of "paper origami-like" sample prep technologies emerging in the field. You can take a saliva sample or blood sample for example and fractionate it across nanofabricated paper wicking devices that can then be put straight into a mass spectrometer. These technologies are low cost too.

Our laboratory has developed new types of nanoparticles that are like biomarker vacuums. They're caged molecules that you can nanofabricate to capture all sorts of types of proteins, and then these nano-cages open up and spill their contents into the downstream detection platforms such as enzyme-linked immunosorbent assay (ELISA), MS or onto a protein array. Simply, developing a new sample prep technique can revolutionize the proteomic space using existing proteomic technology. Some examples are [Biorad's ProteoMiner™](#), [Ceres Nanoscience NanoTrap™](#) [Disclaimer: Petricoin is on the board of directors at [Ceres Nanoscience](#)], and the various magnetic bead technologies that can be used as a chromatographic reagent and/or coupled to antibodies etc for targeted capture.

### **MC: As one of the co-founders of the HUPO, can you tell us more about the [Human Proteome Project \(HPP\)](#), including its aims and some of the project's achievements to date?**

**EP:** HUPO in general always sought to be an organization that helped to provide structure, non-prescriptively, to a field that is inherently extraordinarily more complex and disparate than the genomics field. When we were first founding HUPO with folks like [Gil Omenn](#) and [Sam Hanash](#) and many other early proteomics pioneers, we wanted to figure out constructive ways to help move along the entire field. In proteomics you have so many different technologies and methodologies: protein array technologies, MS technologies, there's sample prep technologies, there's cell-based technologies and non-cell based technologies and sub-classes of proteomics including the glycol-proteomics field, the phosphor proteomics field, the lipidomic and the lipid phosphor proteomic field. All of these specialties and sub-specialties have different cohorts of scientists that in themselves are in their own little sub-groups. HUPO wanted to have an overall organizational structure that represented the efforts across the globe in different areas, and also, we wanted to try to develop what I call "campfire" type projects that people could congregate around and participate in together to advance the field.

Omenn and Hanash along with others helped us start the [Human Plasma Proteome Project](#) that HUPO helped to sponsor and initiate. That was a huge success of being able to say, hey look, let's distribute a single common sample and no

matter what technology you use, no matter what MS workflow you adopt, you can analyze this sample and deposit the data back into a central database that can be shared. This allows for a common portal to basically display the data for the field and for people to do the comparative analytics and say this worked better than this, this is why.

Beyond just convening an annual meeting, beyond just having sponsored conferences, I think HUPO has tried to develop an overall philosophy of ensuring that there are specific types of projects that can be worked on and confederated; the development of reference standards for example, or the development of SOPs and sharing SOPs for the field. These are all things that HUPO really started.

I think if you look at some of the founding principles of what HUPO wanted to achieve, they are replicated in organizations such as the NIH's [clinical proteomics consortium](#). HUPO stands as a showcase for other countries and governmental bodies. When they want to fund life science research at the national level, they look to HUPO because it was the first organization there, and I think that's been a great attribute.

**MC: Thanks to technological advances, the proteomics field has evolved rapidly over recent years. What do you believe the field will look like in 10 years' time? What obstacles currently stand in the way of proteomics advancements?**

**EP:** That's a great question. I guess I'm expecting, or envisioning, that the field is going to be less about the detection methods and more about the stitching of those detection methods into practical applications that we see in our everyday life. What I mean by this is the development of proteomic detection methods in wearable devices, proteomic detection methods that are sensing the environment, the water, the air, or nanosensors implanted inside the body.

For me, it would be extraordinarily depressing if in ten years or fifteen years' time I'm going to an ASMS meeting or HUPO meeting and the focus remains on the classic proteomic techniques themselves. If the proteomics field is still simply talking about the next new MS, or some interesting software tool that allows you to measure this or that better, then the field is going to have stagnated drastically.

The field must get out of just displaying new types of MS equipment. The equipment needs to be in the background and what you are doing with it needs to be in the foreground, as is what happened in the genomics space. If it's just about the machinery then proteomics will always be a "poor step-child" to genomics. At conferences we want to see the application of

proteomics, for example "we can take this machine and now we can do this with it and we can find these biomarkers".

Another way that proteomics is limited currently is a lack of financial investment. The genomics field has sucked a lot of money into their space, perhaps rightfully so, but we need capital infusion into the applied proteomics and clinical proteomics areas.

Furthermore, the field itself hasn't yet identified or grabbed onto a specific "moon-shot" project. For example, there will be no equivalent to the human genome project (HGP), the proteomics field just doesn't have that. The "human proteome" is a constantly fluctuating information archive. Every cell type has its own unique proteome – it depends on what the function of the cell is, and it depends on what point in time you're measuring the protein content of the cell. Projects such as the HGP attracted a lot of PR and money investments for genomics, and so it is a shame that proteomics will not have an equivalent "moon-shot" project.

*Emanuel Petricoin was speaking to Molly Campbell, Science Writer for Technology Networks.*

# THE EVOLUTION OF PROTEOMICS

## Tackling complex biological questions with XL-MS

**Dr Richard Scheltema**



**Richard Scheltema** is an Assistant Professor at Utrecht University where he heads up the Scheltema laboratory within the wider context of the group of **Albert Heck**.

The research group focuses on MS based structural proteomics, for which they develop advanced LC-MS/MS platforms and **analysis software**, working closely with other researchers such as Alexander Makarov who features later in *The Evolution of Proteomics*.

Scheltema is the core developer of **XlinkX** for Proteome Discoverer, and with his team adopts this technology to gain an

in-depth quantitative view on proteins, in addition to gathering special information to answer interesting biological questions.

**Molly Campbell (MC):** What has been the greatest breakthrough in the proteomics research field in recent years?

**Richard Scheltema (RS):** To me, the applications that MS based proteomics have seen over the years are very exciting. An instrument that is basically only measuring masses and rough abundances is capable of:

- extracting the presence of proteins

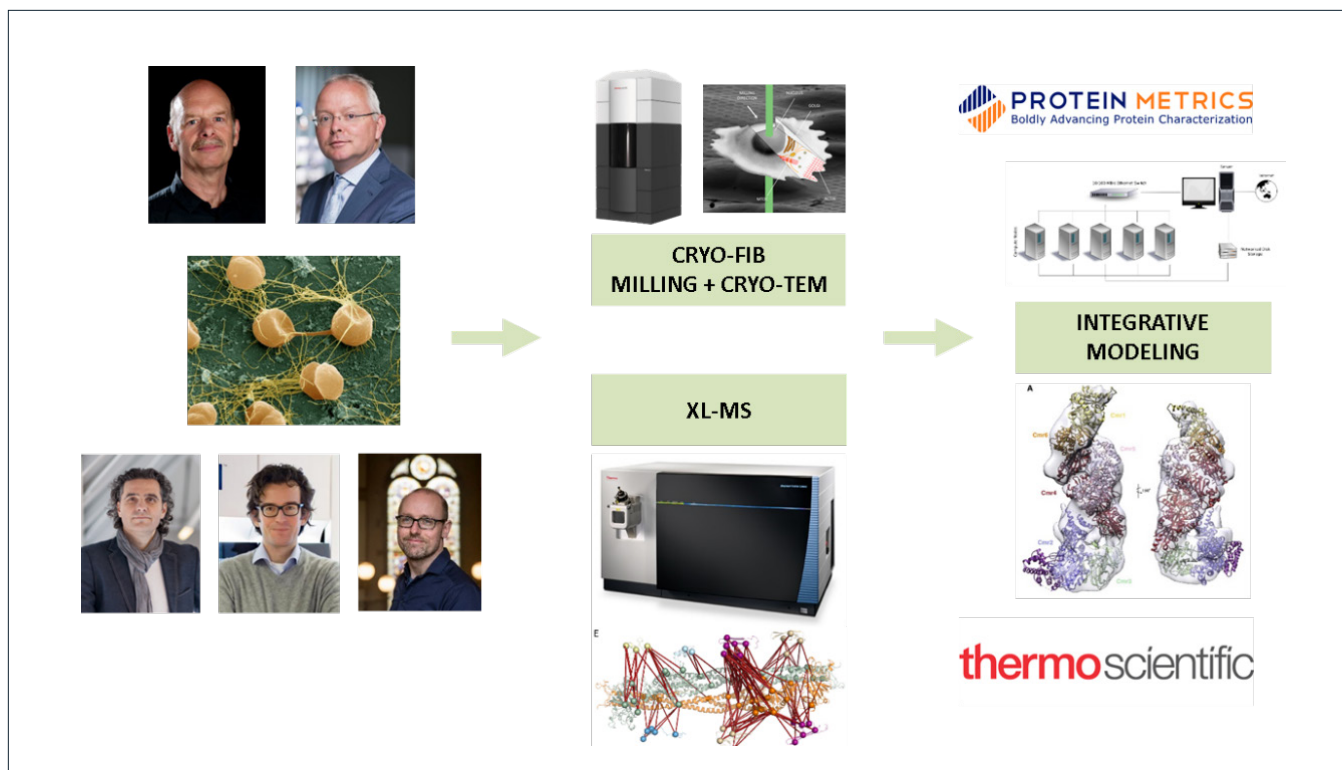


Figure 1: General XL-MS workflow. Original figure. Credit: Richard Scheltema.

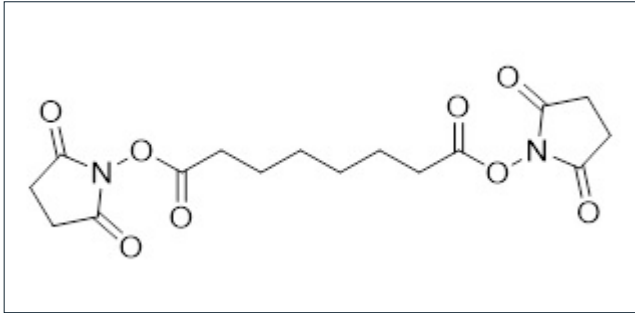


Figure 2: The DSS crosslinking reagent. Original figure. Credit: Richard Scheltema.

- extracting which post-translational modifications are present on which proteins and in which stoichiometry
  - estimating copy-numbers
  - uncovering which proteins are interacting
  - deriving structural information
  - figuring out the cellular localization of proteins
- and so much more – it is mind blowing.

However, even though the ideas behind all these applications that make them work on a functional level are incredibly clever, they all stand or fall with the quality of the used MS platforms. With this in mind, I think the only answer really can be the continued development of the MS platforms in terms of speed and sensitivity over the last decades. The advances we have seen in the platforms have enabled researchers to explore the clever ideas leading to the eventual applications of MS. The Orbitrap family of mass spectrometers illustrates the astounding advancements in MS.

**MC: Your research utilizes cross-linking mass spectrometry (XL-MS). Can you tell us more about this approach and why, for your work, it is superior to other available techniques?**

**RS:** In XL-MS we are interested in investigating protein structure and protein-protein interactions. This is achieved at

the “crosslinking” stage by solubilizing the protein(-complex) in its native state and incubating it with crosslinking reagents – small chemicals with two amine reactive ends that form a covalent bond between two amino acids. After the crosslinking stage, the protein(-complex) is processed which finally sees the protein(s) cut into peptides by a protease (Figure 1, panel I).

This results in a sample containing normal peptides and copies of two peptides connected by the crosslinking reagent. This mixture is subsequently measured by MS for identification – where in most cases the amino acids involved in the crosslink can be assigned providing us with a distance constraint defined by the length of the spacer arm and the two side chains. These distance constraints provide valuable information on how the protein is folded (two peptides originating from the same protein) or on which proteins are interacting and where the interface of this interaction is located (two peptides originating from different proteins).

The reactive ends are separated by a spacer arm and typically the sidechains of the amino acids are targeted, resulting in a structural resolution between 15 and 50 Å (Figure 2). This is by no means near the resolution that one can achieve with techniques such as crystallography, electron microscopy (EM), or nuclear magnetic resonance (NMR) spectroscopy. Therefore, we do not aim to compete with these techniques for resolving structural information – I rather see what we do as highly complementary. We are, for example not limited in size of the proteins and protein-complexes under investigation, can peak inside the protein structure, are not bothered by flexibility in the protein structure and can deal with highly complex mixtures. Even though the information is of relatively low resolution, when combined with high resolution crystal structures of individual subunits and/or high-resolution EM maps, a very detailed structural picture can be built up.

Our ability to tackle complex lysates will most likely result in XL-MS gaining a lot of traction in the foreseeable future.

Transmission electron cryomicroscopy (cryo-TEM), where

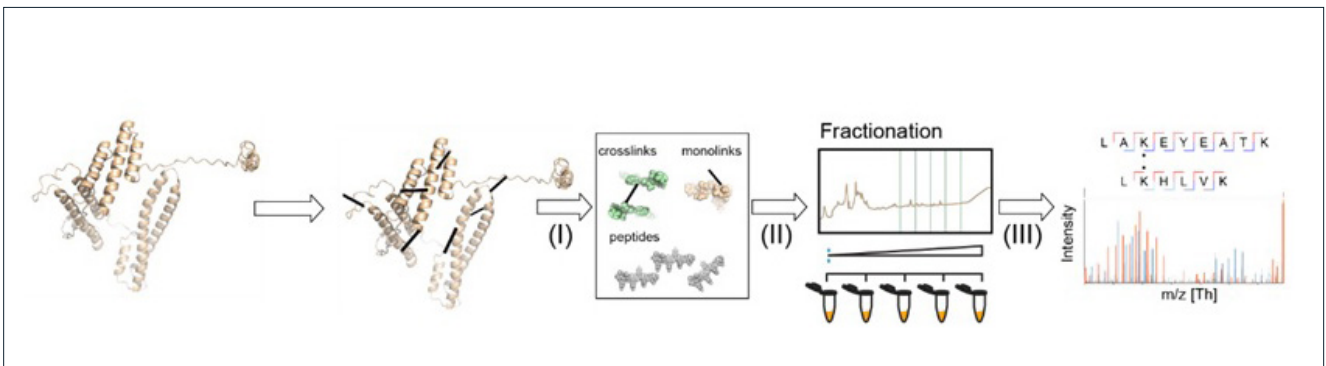


Figure 3: XL-MS combined with Cryo-TEM. Original figure. Credit: Richard Scheltema.



whole cells are imaged by EM, is seeing a great surge in use. From the protein structural outlines recorded with this technique, it is difficult to identify which proteins are involved with each other and how they are arranged within the full structure – questions we are providing answers to.

With this in mind, we applied for a grant at the Dutch Science Council with the goal of developing approaches to [marry Cryo-TEM with XL-MS](#). We hope to achieve this through the use of automated protein docking solutions where we want to infuse a lot of knowledge gained from the MS measurements. We are very confident that we will be able to succeed and we aim to investigate organisms with a large potential for discovering new protein complexes. The project is led by an excellent team of principal investigators, such as [John van der Oost](#), [Albert Heck](#), [Alexandre Bonvin](#), [Friedrich Förster](#), and myself.

### **MC: What key features are attractive in a proteomic data handling software? What processes are involved in the development of a novel software?**

**RS:** Proteomics data processing software is one of the focusses of my laboratory in the larger context of the Heck group. The samples recorded in XL-MS studies have an extra level of complexity as we record data with two still connected peptides (in contrast to normal proteomics experiments where a single peptide is measured). When we started developing our [software](#), solutions were already available; however, we wanted to step into this field to enable experiments with highly complex lysates – an area at that point then not yet covered by existing solutions. In addition, we found the ability to change the data analysis software also highly beneficial for the flexibility it offers to perform “out-of-the-box” experiments. During development, besides ensuring correctness of the extracted peptide identifications, we placed emphasis on user friendliness. We defined this as (1) ability to run on any desktop pc (which requires a large amount of optimization of the algorithms), (2) easy presentation of the results, requiring graph visualizations and browsable tables and (3) support in case of questions and/or problems.

We achieved these goals partly by integrating the developed data analysis in the already established environment Proteome Discoverer. Here, we had the advantage of user-friendly table representations and visualization tools in place, an existing support structure with a lot of people already attuned to how to use the Proteome Discoverer, and a helpdesk available in case of problems.

An additional point of concern for us was that a large amount of bioinformatic solutions appear to be dead on arrival, meaning that once they are published the support and push to

continue their development falls away. Eventually, this means that the software becomes unusable. We were very keen to prevent this with our software solution as we envisioned it being utilized world-wide – which has been the case.

### **MC: In 2018 you published a study in which your team looked at histone protein interaction landscapes using XL-MS in intact cell nuclei. Can you tell us about your findings and what they contribute to the field? How can further research expand the data?**

**RS:** Crosslinking of the intact nucleus provided a snapshot of the histone interaction network and set the basis for investigations into how stimuli influence chromatin organization and influence/regulate the histone interactome. However, further experiments – and importantly, increased depth of analysis over what we achieved – are needed. We would like to apply our nuclear XL-MS workflow in cells treated with histone deacetylase inhibitors (HDAC inhibitors) that promote unpacking of chromatin. This relaxed state of chromatin promotes transcription and will allow us to investigate the dynamics and organization of endogenous transcriptional complexes.

What we have already in part uncovered is how repressive and activating histone marks (methylation and acetylation of histone tails) drives interactions to other proteins. Additionally, we uncovered for known interactors the interaction interfaces to the histones, which we could model on the existing structure of histone.

### **MC: What are some of the key challenges you face in structural proteomics?**

**RS:** There are two major concerns for XL-MS experiments. The first has to do with the low abundance of the crosslinked peptide products. From available data, we estimate that over 99% of the material injected into the mass spectrometer consists of normal peptides (i.e. peptides not modified by the crosslinking reagent), carrying no structural information.

This makes it difficult to detect the crosslinked peptide, carrying the structural information we are after. Over the years this has forced researchers to take two approaches where either the sample is heavily pre-fractionated based on size (two crosslinked peptides are bigger than normal peptides) or charge (two tryptic peptides have twice the charge potential of a single normal peptide). This turns out to be a costly business, as for large-scale projects we have to run 20x three hr measurements to get all the fractions measured. Because we also want to do replicates, this multiplies again by a factor of three if we are looking

at a single system and even higher factors if we include a stimulus to see how the system changes.

Another approach researchers have been taking is to integrate an enrichment handle directly on the crosslinking reagent – creating a so-called tri-functional crosslinking reagent. Previously, biotin has been used as the handle. The produced reagents, however, never found traction as it remains difficult to efficiently detach biotin from the beads used for capture. The biotin handle additionally makes the reagent very bulky, potentially leading to steric hindrance and low access into the protein structure under investigation.

Our solution to this problem – and we really think this is now resolved – was to take a cue from the developments in phosphorylated peptide enrichment with immobilized metal affinity chromatography (IMAC) technologies. The advantage of IMAC is that it has seen large scale automation, phospho-groups can be easily detached from the beads, it has fantastic specificity (meaning non-phosphorylated peptides can be almost completely separated from the phosphorylated ones) and importantly the phospho-group is very small in comparison to biotin. To this end, we developed a tri-functional crosslinking reagent, **PhoX**, that incorporates a phosphor-enrichment handle. From our experiments we found that with this reagent we indeed get incredible performance in creating pure samples and thereby making the detection of crosslinked peptides very easy.

The second major concern for these experiments is that we are able to generate a lot of structural data which currently has no protein data bank structure associated to it. Especially from complex lysates, we observe many complexes of which the individual subunits never were (successfully) crystallized or recorded with EM. This essentially prevents us from building up a structural picture of the complexes based on crosslinks between the individual subunits of the complex. To deal with this situation we and others revert to structural modeling where we use the detected intra-links for each of the individual subunits. The developed tools available are really fantastic as they are able to provide a structure for a given amino acid sequence that is, in a lot of cases, very reasonable. The problem is that we tend to have to search for the best structure in a pool of many generated possibilities. For this, the detected crosslinks are very helpful in validating and filtering, but a large amount of manual work is also required to extract the biologically most relevant model from all the possibilities. This leads to very protracted timeframes for projects trying to uncover structural details. For this, so far, no easy fixes have been proposed and the hunt remains ongoing.

**MC: The proteomics research field is constantly**

**evolving and changing. What key breakthroughs would you like to see occur in the next few years?**

**RS:** The obvious advancement would of course be further improvements in the MS platforms we utilize, but also the liquid chromatography platforms we use to separate the peptide mixtures over a single measurement would be very much appreciated – both in the terms of performance and stability, but particularly stability remains a point of some concern. Also, the development of novel fragmentation techniques and improvements to existing approaches would be very nice to see.

Finally, I would be very excited to see advances in protein purification, sample preparation, data analysis software — basically every aspect of proteomics. All information generated is useful for our structural studies, so, everything is welcome!

*Richard Scheltema was speaking to Molly Campbell, Science Writer for Technology Networks.*

# THE EVOLUTION OF PROTEOMICS

## The Orbitrap inventor

**Professor Alexander Makarov**



**Professor Alexander Makarov, PhD, is the Director of Research in Life Sciences MS for Thermo Scientific and the world-renowned inventor of the Orbitrap mass analyzer technology.**

The Orbitrap-based instruments have been integral to the advances seen within the proteomics field over recent years in addition to enabling significant progress in areas such as metabolomics, environmental analysis, molecular screening and toxicology.

Makarov continues to develop and optimize the Orbitrap technologies, striving for the democratization of MS instruments and their increased availability in clinical environments to enhance disease diagnostics.

In addition to being a widely published author and holding over 50 patents, Makarov is the recipient of several prestigious awards for his contribution to the proteomics field, including the Award for Distinguished Contribution in Mass Spectrometry of American Society for Mass Spectrometry (2008).

**Molly Campbell: In your opinion, what have been some of the biggest breakthroughs in MS-based proteomics?**

**Alexander Makarov (AM):** Over the last 20 years MS-based proteomics has been on a roller coaster as it evolved from the “biomarker gold rush” to a mainstream laboratory technique for the elucidation of disease pathways and mechanisms.

2019 marks 20-years since I presented the Orbitrap technology at the annual conference of [American Society for Mass Spectrometry \(ASMS\)](#), and it has been a period marked by significant advances in the applications of the technology across a broad range of scientific fields including proteomics, metabolomics, environmental and food safety, all of which play an essential role in making our world healthier, cleaner and safer. It has been a privilege to see how much Orbitrap technology has contributed to the rapid evolution of science!

This contribution was well matched by tremendous progress across all other major stages of proteomic experiments i.e. sample preparation, liquid separation, analytical methodology and data processing. As examples, I would like to mention automated extraction and digestion stations for sample preparation, ultra-high pressure nano- and capillary-flow liquid chromatography for liquid separation, data-independent acquisition and tandem mass tag workflows for analytical methodology, and software suites such as the Thermo Scientific [Proteome Discoverer Software](#) for data processing, as well as third party options.

**MC: As the inventor of the Orbitrap mass analyzer, which research study utilizing the equipment in the field of proteomics has excited you the most?**

**AM:** In the field of proteomics, I am most excited by research that explores the intricate mechanisms of different diseases, as it is truly detective work to correctly identify the real band of culprits behind human suffering. I am also excited by the applications of proteomics for the analysis of historical objects such as pictures, manuscripts and bones – this research requires equally elaborate and inquisitive intellectual effort and deep knowledge of multiple disciplines.

**MC: Increasingly, research laboratories are striving towards single-cell proteomics. How can the family of Orbitrap technologies facilitate this movement?**

**AM:** Orbitrap mass spectrometers are already actively used in pilot studies to advance single-cell proteomics, and I expect that both data-independent acquisition and tandem mass tag workflows will find increasing use in this important and rapidly emerging field of science. There is also significant performance and throughput improvements to come in the future from novel, yet to be implemented, instrument enhancements.

Naturally, progress in MS needs to be matched by advances in adjacent areas of technology such as microfluidics, sorting flow cytometry and liquid separations – most likely with capillary electrophoresis gaining importance in the latter.

**MC: You have worked on the development of several proteomic techniques. What technical challenges does your team continue to face in the development of novel MS technologies?**

**AM:** As MS developed from a cottage industry of the 1980s and 1990s into a modern industry like aviation in 2000s-2010s, each new development required larger and larger research and development teams to match the increasing complexity of instruments and the skyrocketing importance of software at all levels, from firmware to application. All this extends the cycle time of each innovation and also forces us to concentrate on solutions that address the most pressing needs of the scientific community.

In parallel, the increasing democratization of MS brings with it new requirements for instruments, such as far greater robustness and ease-of-use, which need to be balanced against some aspects of performance.

**MC: In a recent talk at the 10th MaxQuant Summer school, you discussed the “death ladder” of MS, in which several orders of magnitude of sensitivity are lost throughout the MS process. What are the reasons for this loss, and can you describe how the Orbitrap technology overcomes this?**

**AM:** The life of an ion in MS is tough and full of discrimination! It all starts with competition for charge during ionization, which continues during space charge-dominated ion transportation from atmosphere to the vacuum, mass selection, fragmentation and then mass analysis. Orbitrap analyzers allow us to reduce losses during the last stage by at least an order of magnitude when compared to other high-resolution techniques. However, each of the remaining stages (except for fragmentation and only when done with gas collisions) still result in orders-of-magnitude lost.

The current frontier of instrument development is focused on reducing losses at the stage of mass selection, with a number of solutions ranging from data-independent acquisition with widened mass filtering window, to linked trapping/ion mobility separation/quadrupole selection/mass analysis operation. However, even modest improvements during preceding stages would give us still another order-of-magnitude improvement when compared to the current best-case scenario.

**MC: Systems biology relies heavily on MS. As this field of research continues to grow, how can MS technology be further developed to meet the needs of researchers?**

**AM:** I foresee MS to continue its rapid development along two somewhat divergent but also closely linked directions. First, the performance envelope of high-end mass spectrometers will continue to improve, enabling more versatile and deeper analysis of new types of analytes. Second, this progress will be accompanied by the already mentioned democratization of MS, with instruments hopefully becoming as ubiquitous and affordable as liquid chromatographs or even microscopes. Interestingly, the amount of innovation and effort needed to fuel the second trend often appears to be even higher than that required for the first trend!

*Alexander Makarov was speaking to Molly Campbell, Science Writer for Technology Networks.*



# THE EVOLUTION OF PROTEOMICS

## Advancing MS-based plasma proteomics

**Dr Gary Kruppa**



**Gary Kruppa, PhD, Vice President of Proteomics, Bruker Daltonics Inc., has over 30 years of experience in the field of MS, having served as a Vice President at Bruker Daltonics for over 20 years.**

Kruppa received his PhD in chemical physics from the [California Institute of Technology](#), and his BS from the [University of Delaware](#). Kruppa oversees market and applications development management for Bruker's innovative solutions for research in proteomics.

In this interview, Kruppa discusses the recent technological advances that are driving MS-based plasma proteomics for biomarker discovery and beyond.

**Molly Campbell (MC): Can you provide some background into the need for new plasma biomarkers? How can new methods improve biomarker discovery?**

**Gary Kruppa (GK):** Many existing diagnostic methods that are currently based on proteins measured by antibody binding could be translated to MS-based assays, improving specificity. Leigh Anderson's review paper on [The Human Plasma Proteome](#) as early as 2002 provides a good reference to both the advantages of plasma proteomics, as well as its inherent challenges.

There is an unmet need for new biomarkers for many diseases, but the protein content of plasma is very complex, making discovery and validation a challenge. A number of cancers currently have no known plasma biomarkers, the early detection of which would revolutionize patient care. New techniques for biomarker discovery, as well as instruments with high sensitivity and robustness, are required to meet this need. The abundance of biomarkers for early stage cancer in plasma is likely to be quite low, so very good sensitivity and very high coverage of proteins in plasma is required to detect such biomarkers.

**MC: In addition to medicine, what other fields of science can benefit from proteomics insights?**

**GK:** The development of new proteomics technologies is key for pharmaceutical drug development. Many drug targets are proteins, so gaining a deeper understanding of the cell mechanisms that influence normal concentration range, turnover rates, accessibility to protein pockets for suitable targets, protein-lipid complexes on the cell surface, and multiple target or off-target hits are important factors that MS-based proteomics helps unravel. Once researchers have a suitable target and a drug with which to target it, they want to assess the proteomic profile to ensure they are knocking down the expression of the target protein and affecting only that pathway. The more sensitive the methods, the more you can study small on and off-target effects on the proteome.

Much early stage drug development is done in cell culture, so sensitivity and dynamic range are not as much of a problem as they are in plasma, but the whole range of pharmaceutical sciences is very interested in proteomics for numerous applications. In the field of biology, metaproteomics is a major area for study, which includes the study of how organisms interact, which organisms are present, and how their proteomes are affected by these interactions.

A key area of metaproteomics includes the study of gut microbes in humans, from both a fundamental science standpoint and studying the effect of the human microbiome on health. Host-pathogen protein-protein interactive networks (PPI) is another key area of focus, where drug development could be tailored to target virus borne diseases, such as Zika, Dengue or Ebola, which hijack the cellular protein machinery of the host and often lead to human-wide illness and even death.

### **MC: Can you explain the use of 4D matching, and what this means for biomarker discovery?**

**GK:** 4D matching has a huge impact on biomarker discovery, pharmaceutical drug development, and fields like metaproteomics and PPI. In standard bottom up proteomic studies, the proteome is digested and peptides are detected as they elute from a liquid chromatography (LC) system. The retention time on the LC column is one dimension of analysis, the mass measured is the second dimension, and the intensity of the peaks is the third dimension. The resultant 3D peaks are integrated, which reflects the intensity of the peptide, which is identified by MS/MS and tells you what protein it came from, and by inference how much of that protein was in the sample.

The fourth dimension (4D) refers to the addition of ion mobility, which has been around for a number of years but has not been used routinely in proteomics. The invention of trapped ion mobility spectrometry (TIMS) by Bruker has made the routine use of ion mobility in proteomics possible. Additionally, the parallel accumulation-serial fragmentation (PASEF) scan technique in the TIMS cell, increases sensitivity and speed. As the ions are trapped and then elute as a function of their mobility, this additional dimension of information can be used to improve identification. As multiple peptides co-elute off the nano-LC column, their unique collision cross sections (CCS) allows for further gas phase separation in the TIMS cell, allowing for more peptides to be identified. This gas phase separation as a result of TIMS is the fourth dimension in addition to retention time, mass-to-charge, and intensity, and termed as 4D matching.

The benefits of 4D matching due to the PASEF scan allows researchers to identify lower abundance proteins, such as tissue leakage proteins or signal proteins, with higher confidence and with the required high sensitivity. For example, for data dependent analysis (DDA/PASEF), the CCS values will be used as an additional identification criterion in the search engine to provide confidence in the peptide identification. In the case of data-independent analysis (DIA/PASEF), or an intermediate method called “match between runs”, you can use the CCS value as a unique peptide signature to help align features and increase confidence in assignments.

### **MC: What stages are involved in the development of a novel technology for use in proteomics research?**

**GK:** Firstly, you have to identify the unmet need of a particular application, which in proteomics is usually the depth of coverage of the proteome, speed, and sensitivity.

There are an enormous number of proteins in plasma, over an incredible dynamic range, so this is the major unmet need that has been recognized for years.

While you want to drive the higher sensitivity and dynamic range by improving the specificity, sensitivity and speed of the mass spectrometers, you have to keep in mind that in the clinic, samples must be analyzed quickly. This is another unmet need – to be able to generate a proteome from a person in an hour at a reasonable cost, to ensure applicability in the clinic. Researchers may be willing to spend hours to days on a single sample to find a biomarker, but validation cannot take this long because you need a minimum of a thousand samples, and for that to work you need a method that can be done in an hour or less. For routine application in the clinic with thousands of patients per day, it has to be even shorter.

Robustness is another need to be met. To routinely measure a patient’s proteome and compare them at regular intervals, results must be reproducible in order to observe changes in the patient, rather than in instrument performance. Thus, you need a mass spectrometer that is very robust. In addition to its 4D matching capability which adds specificity, and the speed which enables you to hit a lot more targets, the robustness of Bruker’s timsTOF Pro is a big advantage.

Once the unmet needs are identified we must then develop solutions to solve them. In many cases, partnerships and collaborations are also crucial. Bruker works closely with both commercial and academic software partners, e.g. [Bioinformatics Solutions Inc](#) (the producer of [PEAKS software for proteomics](#)), [MaxQuant](#), [Skyline](#), and [Protein Metrics Inc.](#) (the producer of [Byonic™](#)) to help dig deeper into results.

We also partner with chromatography experts to maximize robustness and speed. Samples are injected into an LC system to separate the peptides, and generally for sensitivity purposes this is done with nanoflow chromatography. However, alternative methods may be more suitable for clinical applications because nanoflow can pose some practical challenges.

With the timsTOF Pro, the nanoflow chromatography is one of the bottlenecks, so we have partnered with Evosep which has developed a very robust, high-throughput LC system, the Evosep One, with moderately low flow rates that are ideal for clinical research applications. By partnering with these different companies and academic institutions, we can help speed up the development and bring best practices from different sources to meet these needs.

**MC: In your opinion, what have been some of the most exciting technological developments in the proteomics research field thus far? What major advances do you see in the future of the proteomics research field?**

**GK:** Even after nearly 25 years of proteomics, the field remains largely fragmented, especially in contrast with genomics and the landmark human genome sequencing work done by [Venter](#) and [Lander](#) in the mid-2000s.

In proteomics, the challenge remains in achieving relevant proteomics depth in the shortest possible time, which is no easy task given that the human genome contains about 20,000 protein coding genes. To achieve the required proteomics depth, given the duty cycle of modern mass spectrometers, two mutually exclusive approaches are employed, each with its pros and cons; DIA and DDA with “match between runs” philosophy.

Furthermore, LC-MS is a hyphenated technique, and the LC methods segment into nanoflow and microflow, each with its strength and weakness. So unlike a DNA sequencer that benefited from PCR amplification and removed the focus from the analytical technique itself, proteomics very much remains dependent on the advances made in both LC-MS and in bioinformatics, trying to decipher the complex information acquired per second.

The [Orbitrap mass spectrometer](#) made high resolution, accurate mass data, routine which in turn triggered bioinformatics tools that used the accurate mass information to dramatically improve the confidence in analysis. This

key development resulted in the wide-spread use of MS-based proteomics, leading to a rapid advance in the field. The timsTOF Pro with PASEF technology represents a new step-change for MS-based proteomics, as it adds an additional key qualifier – the peptide CCS value. Even today, chromatographic retention time plays a key role in bioinformatics, so adding peptide CCS values – a critical gas phase separation signature of the peptide – could exponentially improve the confidence of analysis by reducing false discovery rates, discover multiple site PTMs for the same peptide sequence, and go deeper into the proteome by triggering MS/MS in windows around certain ion mobilities. This unique CCS signature could be advantageously used by adding a level of intelligence, for example in the immunopeptidomics (and other non-tryptic peptides, targeted proteomics) using PASEF triggered parallel reaction monitoring (PRM), or for connecting the various PPI pathways to study host-pathogen infections to discover new pharmaceutical drugs.

The field has been developing rapidly for the past 25 years, thanks to the continuous evolution of MS-based proteomics. The dramatic improvement in the robustness and speed performance the timsTOF Pro makes large cohort studies possible, analogous to the impact made by the genome-wide association study (GWAS). We believe the TIMS/PASEF approach with the critical CCS-peptide signature information, together with advances made in machine learning capable bioinformatics tools, make it an important consideration in enabling proteomics to become more clinically relevant.

*Gary Kruppa was speaking to Molly Campbell, Science Writer for Technology Networks.*



# THE EVOLUTION OF PROTEOMICS

## Developing research strategies to propel proteomics into a new era

**Professor John Yates**



The final instalment of *The Evolution of Proteomics* series features an interview with **Professor John Yates** from the Department of Molecular Medicine at **Scripps Research**. The **Yates laboratory** is focused on developing strategies and tools in proteomics to answer basic biological questions.

The work of Yates and his lab has been instrumental in driving the evolution of proteomics, with key achievements including the development of shotgun proteomics, the creation of the **SEQUEST** algorithm allowing tandem MS to be correlated with protein sequences, and of course the development of Multidimensional Protein Identification Technology (**MudPIT**) that resulted in a shift from traditional 2D gel-based MS techniques to liquid chromatography approaches in proteomics.

**Molly Campbell (MC):** In a 2018 talk you mention the idea that proteomics was a “great unintended consequence of genomics”. In your opinion, what have been some of the most exciting breakthroughs in proteomics?

**John Yates (JY):** The biggest breakthrough is that proteomics exists at all. Back when genomes were first being sequenced, protein biochemistry analysis focused on one protein at a time – it was laborious, you could spend an entire year trying to sequence just one protein. It was also incredibly inefficient, relative to what we can do today. Now, in just a few hours, you can sequence an entire protein complex and identify what each component of the protein complex is doing. The advances have been stunning.

The reason that proteomics is great unintended consequence of genomics is that nobody was talking about the impact of genome sequencing on protein biochemistry, it really was

something that came out of nowhere and had a huge impact. When you read a report by the National Academy of Sciences in the US, and why they should sequence the human genome, most of the discussion centers around “oh, bioinformatics will figure out what everything does, and we’ll learn about how cells work” and so forth, and really no discussion about the impact it might have on protein biochemistry.

**MC: The Yates Laboratory at the Scripps Research Institute develops and applies MS-based proteomics techniques to study conditions such as Alzheimer’s disease, schizophrenia and depression. How can a proteomics approach enhance our understanding of the pathophysiology of these conditions?**

**JY:** These are very complicated diseases. There have been a number of genome wide association (GWA) studies trying to figure out what the genetic components of these diseases are that have been unfortunately somewhat unsuccessful. As a result of such studies, the concept of “missing heritability” came about – but maybe it’s not missing heritability? Maybe it’s not genetics, maybe it’s the environment, together with the genes that is affecting protein networks in ways that we don’t quite understand yet.

Alzheimer’s disease in particular seems to be a disease of a breakdown in the proteostasis system, the system that maintains protein folding and degrades proteins. When they misfold, you get an accumulation of misfolded proteins in the brain that becomes toxic to cells and so forth. There are a number of diseases now which are clearly failures in the proteostasis network, where protein misfolding can result in a loss of function or a gain of function. So, we really need to study these diseases at the protein level, as you will only get so far with genetics and genomics. In order to do more at the protein level, we still need to advance our technology so that we’re competitive with genomics technologies.



**(MC): You have pioneered the development of several methods and software systems that have shaped proteomics research. What technical challenges do you face in further refining proteomics techniques so that they are increasingly sensitive and specific?**

**JY:** Some of the trends that are occurring in the field include people trying to come up with ways to be more efficient and more high-throughput. One of the complaints from funding agencies is that you can sequence literally thousands of genomes very quickly but you can't do the same in proteomics. There's a push to try to increase the throughput of proteomics so that we are more compatible with genomics. One of the real exciting things in my opinion is the move of proteomics to single cell. People are finally making progress on cells that are biologically relevant, not just those that are packed with a few proteins such as red blood cells. That's going to be a great area.

I just went to a think tank, sponsored by one of the NIH Institutes, that was discussing single cell proteomics. I think there's enough excitement there that funding agencies can start putting some money into it to advance it.

One of the things that we are dependent upon in the MS field is for instrument manufacturers to keep advancing the technology. Some of the very fundamental basic research in MS takes place in academia, but really in order to make that technology useful it must be commercialized and advanced with the quality control and standards that commercialization brings to the instruments. It's always exciting when you go to [ASMS](#) to see what instruments or technologies are going to be introduced by the manufacturers.

**MC: Please can you tell us about your recently published work in cystic fibrosis, and how this research may help to identify novel drug targets?**

**JY:** One of the papers we published looked at the interactome of the protein that is involved with cystic fibrosis, called the cystic fibrosis ion transport regulator (CFTR).

We looked at the interactome between the wild type version of the protein and the most common disease form of the protein, which is the Delta F508, and there was a disease specific interactome. As we began to study the interactomes we found about 40 proteins where if we knocked down their expression, we could influence the maturation of the disease form of the protein in some fashion.

We tested a handful of these, about eight, to make sure that they actually restore channel function. Out of the eight that

we tested, seven did. The ones that are enzymes would be fairly easy to target by drugs as you can inhibit their activity.

We actually did an experiment where we took one of the one of the proteins that we were studying and found an inhibitor for it published in the literature. We made the inhibitor and tested it and what we found was that we could rescue the mutant form of the protein. We've identified a number of proteins, which are potential targets, where if you inhibit their activity, you can rescue the protein.

A number of the proteins that interact with CFTR are kinases and phosphatases, and so we started looking at the modifications of CFTR and we found some modifications that looked like they may be important to the decision making process of whether the protein is mature or not and should be sent to the cell surface. We established that there is in fact a post-translational modification code that determines whether a protein is mature. I'm not sure how that would turn into the creation of drug targets, but it is certainly interesting biochemistry.

**MC: Your research encompasses the areas of bioinformatics and software development, methods development and biological applications. Do you face any difficulties in integrating these elements, and if so, how do you overcome those difficulties?**

**JY:** It's not really that difficult to integrate them, the challenge has always been trying to prioritize which elements need to be done first (especially in a lab where a lot of people are doing different things)!

We've got a fairly robust and well-established pipeline of software tools that are used for a wide variety of things that are used by anybody that's doing any kind of proteomics research at Scripps.

Where a lot of people spend time is addressing the question of "what biology have I discovered in my experiment?" and trying to come up with the tools that help people become more efficient at answering that question. When we have group meetings and discussions about bioinformatics, they are always the most contentious, heated and lively discussions and they are a very important topic.

**MC: As an expert in quantitative proteomics with many years' experience in the field, what do you envision for the future of proteomics?**

**JY:** These are always tough questions. I think proteomics is going to advance in a few areas. It is going to be more sensitive

as we push down towards single-cell analysis. It's going to become more high-throughput so that we can analyze more patient samples and so forth, enabling it to be on par with RNAseq type strategies. The scale of proteomics is going to advance to the point where we can obtain an entire proteome in a single experiment. We're close to this now, and we may actually be close enough. Some of the experiments that we're seeing in single-cell experiments see 1200 to 1500 proteins, and if you look at the RNAseq experiments, they're only seeing around 3000 or so genes – so we aren't far off. Another main goal in proteomics is to bring down the cost of mass spectrometers.

*John Yates was speaking to Molly Campbell, Science Writer for Technology Networks.*