Pioneers in Proteomics

John Yates, Ph.D.
Professor of Cell Biology, Head of the Proteomics Mass Spectrometry Lab
The Scripps Research Institute

In part four of the Pioneers of Proteomics series, Dr. John Yates discusses the latest trends in mass spectrometry, requirements for continued advancement in the field of proteomics, and the impact of proteomics in the clinic.

Dr. Yates is Professor of Cell Biology and Head of the Proteomics Mass Spectrometry Lab at the Scripps Research Institute. Dr. Yates received his Ph.D. in Chemistry at the University of Virginia. His graduate research involved the development and application of tandem mass spectrometry for sequence analysis of proteins. In addition to proteomics, Dr. Yates' research interests include the development of integrated methods for tandem mass spectrometry analysis of protein mixtures and bioinformatics using mass spectrometry data. He is the lead inventor of the SEQUEST software for correlating tandem mass spectrometry data to sequences in the database. He has received the American Society for Mass Spectrometry research award, the Pehr Edman Award in Protein Chemistry, the American Society for Mass Spectrometry Biemann Medal and has published over 250 scientific articles.

1. On the shotgun approach to proteomics

"…by figuring out the amino acid sequence for each of the peptides we can reconstruct the collection of proteins that were present in the original mixture."

…the best way to explain it is to look at historically how people used to analyze proteins. In the past what you would do is you would try to isolate a protein to homogeneity so you would have a purer protein and then you would try to sequence that protein in some fashion. What shotgun proteomics is rather than trying to isolate a single protein for an analysis, what we do is we take the collection of proteins and we digest it up to create a collection of peptides and then we analyze those peptides directly in the mass spectrometer, collecting tandem mass spectra. Each tandem mass spectrum represents an address that relates back to the protein from which that peptide came from. Peptides are really just short pieces of a protein. So by figuring out the amino acids sequence for each of the peptides we can reconstruct the collection of proteins that were present in the original mixture. Now the advantage to that is you can now do experiments where you pull down collections of proteins that may be related by their physical interactions and you can identify all those proteins in a single experiment whereas in the past you would have to separate all those proteins and then try to sequence them or identify them individually. So there is a huge time savings, labor savings, and so forth associated with the ability to do that. The other advantage is that when you're trying to isolate proteins, the more steps you have in that process the less likely you're wind up with sufficient

material for the analysis, so in these types of shot gun proteomics experiments you do it would be a single step purification and then you would digest and go directly to mass spectrometer so you minimize the number of manipulation steps that you're using in order to get the proteins into the mass spectrometer. So you want to do that with the minimum number of steps possible.

2. On the trends in mass spectrometry

"…people have been developing programs to do statistical analysis on the data…"

So people are trying to explore ways to pull quantitative information out of standard mass spectrometry data without having to incorporate labels and if you can do that and do that well. That would be a huge advantage because stable isotopes are expensive and they're difficult to use and they complicate the experiment and so forth. So that's an important direction. The other direction that people are going is trying to do *de novo* analysis of tandem mass spectra will be able to read a sequence directly off the spectrum and then

use that sequence and that's in various stages. That's a much harder problem because you have to consider the entire sequence space rather than just the sequence space that's within a database and it does require a much higher quality tandem mass spectrum than you would ordinarily need for a database searching approach. But for most people, you can get very far just doing database searching to identify your proteins and get answers to your questions. Now as people start to move into shotgun proteomics one of the things that they find is that coming out the back end is these huge lists of proteins so you need ways to organize those lists, filter those lists, analyze those lists of proteins and the data. So people have been developing programs to do statistical analysis on the data and to get things organized so that you can get an answer out.

3. On the evolution in mass spectrometry

"It allows us to access proteins which are of lower abundance in a cell."

…there are clear distinctions between what we could do then and what we can do now. Back then the data could be acquired relatively quickly but we would spend an inordinate amount time trying to analyze the data, interpret the tandem mass spectra so the process was very slow. It was take us on the order of a year to go from data collection to actual having a sequence. Now that happens on the order of minutes, being able to go from protein sequence to an answer because the paradigm is shifted.

The real key step that made this a useful tool, a large scale tool, is the fact that we have automated ways to analyze the data. Prior to that development, all the data analysis was done manually and that was a long and time consuming, difficult process. The other thing that's really changed between then and now is that sensitivity is about six orders of

magnitude better than it used to be so that really allows us to access proteins that are involved in key processes within the cell rather than analyzing more abundant proteins.

So we can now work with much less material than we used to be able to work with having say one by ten to the minus nine moles of material required to get a spectrum, versus having one by ten to the minus fifteen moles to get a spectrum, to get information. So that allows us to work with much less starting material. It allows us to access proteins which are of lower abundance in a cell. It just really enables one to do many more sophisticated studies.

4. On the growth of mass spectrometry

"If you get the tools in the hands of the biologists they can stay abreast of the technological changes to the extent that they want or need to."

So clearly if you want to advance biology you've got to get the tools in the hands of biologists and there are few things that limit the growth of mass spectrometry in the biology field. One is simply the ease of use of the instrumentation and manufacturers are always trying to make these things easier to use, more intuitive to use, and so if they can push that along faster, maybe with focus groups that are not experts in mass spectrometry and make sure that they can adequately use the instrumentation.

So right now the current trend has been to give money to core facilities and for core facilities to buy the instrumentation and provide the services to the local community. The problem with that model is that in a dynamic field like proteomics and mass spectrometry, it is very difficult for core facilities to stay abreast of what's going on and so they always lag behind.

If you get the tools into the hands of the biologists they can stay abreast of the technological changes to the extent that they want or need to. So if they need to do shotgun proteomics experiments they can learn how to do that. If they want to just identify bands off of a gel they can do that too. So then they won't have to rely on their local core facility as a way to get the answers to their questions.

5. On the need for standardization

"Trying to find correspondence among the databases can be real tough."

So one of the biggest problems that we have with sequence databases is that, if we do an experiment and then it takes a year or two years for the biology to get sorted out in that experiment, the databases have gone through so many revisions that the accession numbers no longer match up so it makes it difficult to report to people what the accession number might be for the proteins that were identified that isn't a current database.

Usually by the time the paper comes out those accession numbers have changed again anyway. So there needs to be a way to go to track accession numbers among various proteins. The other issue that is fairly important is that people use different sequence databases to analyze their data. Trying to find correspondence among the databases can be real tough. For us, on a practical level, this has been a real challenge is how do you keep track of all that information? So having some kind of standard and way to keep track the accession numbers of various proteins as the database is evolved would be very important for people to be able to look at data and understand how that information that shows up in a paper might relate back to what they're trying to study.

6. On collaboration

"For my lab, collaboration has been absolutely essential."

For my lab, collaboration has been absolutely essential. Most of our collaborations have been on more fundamental biological questions. But there are two things which are very important in this is that one -- we have somebody that's got the expertise to deal with the biology and the second thing is that you've got somebody who is particularly passionate about the biological question that is being asked. So, when we team up with them, we bring our expertise to bear on the problem, and we generate the data, and then they vigorously pursue that data to try to understand the biology. And that's very important for pushing this along. So, most of the collaborations that we're involved in, we would not be able to follow up the biology with the same level of expertise or the same level of passion that these other biologists bring to the plate. So, these collaborations for us have been wonderful. They have been absolutely fantastic.

It's critical that people be passionate about what they're doing**.** That way they bring the energy to the problem, and they bring their expertise to the problem.

…in a team environment, the team network environment, the value added to that is that you have other pairs of eyes looking at other data besides just your own and so you can get either new assessments on that data or you can get a critical assessment of that data, but it allows other people to look at your data and to make judgments on that data. You can get people that will have new insights to what you've done and so that's certainly a value to trying to learn new things.

7. On the challenge of studying serum

"Serum presents huge challenges for proteomics technology at this point."

I think the interest in serum and plasma is simply because it's what's been used clinically for a long time. It's fairly routine to go in a doctor's office, get a blood sample taken.

And so there's a natural tendency to say, well, we should use those for these proteomic measurements. And it may be correct, and it may not be correct, but it's certainly a convenient body fluid to obtain from a patient. Serum presents huge challenges for proteomics technology at this point. There's the ten to the eleventh, ten to the twelfth range of protein abundances in serum. And so there's a lot of interest in trying to figure out how to solve those problems so that we can use that serum -- use serum as a way to look for biomarkers.

8. On use of mass spectrometry in the clinic

"it's quite feasible that mass spec could be used for doing clinical analyses."

So, mass spec is always used in clinical analyses. So, it's used quite heavily in neonatal screening so to look for inborn errors of disease and things of that nature. So, it's quite feasible that mass spec could be used for doing clinical analyses. But the critical factor in a clinical test is how much that clinical test costs. So, a test that cost $10,000 isn't going to be very useful to that many people versus a test that costs $50.00. So, one will have to sort out on the basis of cost whether the mass spectrometry is the appropriate way to do that test versus some other method. In terms of neonatal screening, it turns out that mass spectrometry is the most cost effective and accurate manner in which to do these.

I think that probably the best strategy will wind up being that mass spectrometry is used to discover the markers and then something like a protein micro array or a protein affinity array will be the best way to test patients for those because it will be a high throughput low cost method for figuring out what disease a person might have.

Perhaps in the longer term, mass spectrometers as they become cheaper and easier to use and as people develop strategies, there might be ways to do these clinical tests in a clinical laboratory at the hospital. That would be faster and cheaper and so forth than an ELISA test. The cost -- the overriding factor in any clinical test is always going to be cost. So, can you do these costs in an effective manner and to keep the cost as low as possible? So, if mass spectrometry and proteomics can deliver cost effective diagnostics, then it will be useful; if it can't then it will just be used at the research end for discovery.

9. On the impact of proteomics in the clinic

"I think we're at the technology development proof of concept stage."

I think it's simply a time issue. I don't think that people have been trying to address clinical issues long enough. There's always a period of people trying to sort out what

are the right samples to look at, what's the right way to look at them. So, there's a technology development phase and an application phase, a proof of concept phase. So, it's just simply that not enough time has evolved.

I think we're at the technology development proof of concept stage. So, people are trying to sort out what are the right fluids to look at. Looking at what type of information do we get at when we look at that particular disease state. Can we actually find something? Can we actually find markers in that particular disease state. The other issue that I think is of concern is that when you do look at say a cancer patient versus a healthy patient, and the markers that you see, are they indicative of the cancer, or are they indicative of just some sort of distressed disease state? So, I think, what's probably going to have to happen is that people need to look at various types of diseases together. That might be where just public data sharing becomes very important in order to figure out what's a marker related to disease, what's a marker related to stress.

> …the drug industry is interested in biomarkers from a different perspective. So, they want a marker for a disease so that when they treat the disease with their compounds, they'll have something that they can report to them -- well, this marker has been decreased; therefore, we have, the drug has efficacy. So, there's actually a lot of work going on in the pharmaceutical industry trying to identify biomarkers for diseases that they want to treat. So, yes, it's having a huge impact in the pharmaceutical industry.

10. On the Clinical Proteomic Technologies Initiative

"…NCI's efforts to create reagent resources will be great for science…"

One of the marvelous things that came out of the genome project besides getting all that information about genomes is that it turned out to be an enormous resource for all kinds of different technologies -- so microarrays, proteomics, creating expression systems for various proteins. And so I think one of the lessons out of that project is the value of creating these resources. And so NCI's efforts to create reagent resources will be great for science because what that will do is it will allow people to accelerate the pace -- it will accelerate the pace of research because rather than having to make an antibody you can just go ahead and order one up and so that will increase people's ability to go from maybe a protein identification to a validation or to be able to do additional studies to try to sort out what's going on. So, the creation of resources will increase dramatically the pace of research.