

Pioneers in Proteomics

Mark Boguski, M.D., Ph.D.
Vice President and Global Head, Genome and Proteome Sciences
Novartis Institutes for BioMedical Research

In part two of our Pioneers of Proteomics series, Dr. Mark Boguski discusses the challenges of proteomics research, the changes needed to foster progress in the field and the application of proteomics in the clinical setting.

Dr. Boguski is a founder of the integrated program in genome and proteome sciences at the Novartis Institutes for BioMedical Research (NIBR), overseeing the interface of genomic services with disease areas and platforms. Known for his expertise in bioinformatics, Dr. Boguski served as director of the Allen Brain Atlas project in addition to holding affiliate faculty appointments at the Fred Hutchinson Cancer Research Center and in the Department of Medicine/Genetics at the University of Washington. He was senior vice president of R&D for Rosetta Inpharmatics and has been involved with the development of a number of enabling information resources at the National Center for Biotechnology Information.

1. On the origins of proteomics

Proteomics is a new name for an old discipline, really. Proteomics actually predates the human genome project by two or three decades. A woman named Margaret Dayhoff who almost no one has ever heard of today, actually was building the first protein databases back in the 1960's, and actually had developed some computer programs to analyze mass spectrometry before most biologists even knew what a computer was. So it really is an intellectual tradition that goes back a long time, and actually predated the Human Genome Project. Now, in the '60's there was something called the human protein index that was proposed as a big science project kind of in the wake of the Apollo Program. However, even though the concept was there and the vision was there, the technology was not. And it really took I think, the Genome Project to reinvent how science is done in terms of the sociology of research, teamwork rather than individual investigator laboratories driving most of it, and this sea change in not only in the sociology of science but in the technological advancements that occurred, actually makes the time very ripe now to execute on that initial vision of proteomics that was developed so long ago.

2. On the sociology of science

I'm thinking of a really unique example from the genome era when people came together for a common goal. When the *Drosophila* genome was sequenced as a result of a public/private partnership, in order to annotate that sequence we could have done it the traditional way - just had one small group annotate it or release the data and have the

world annotate it. But a new experiment was tried here, which I think was a very successful and very exciting, that is to have an annotation jamboree. What was done was that something like 40 or 50 or 60 bioinformatics people and computational biologists came together for two weeks in a crash program to annotate the *Drosophila* genome. That was a very fertile environment for sharing knowledge and making new relationships that carried on to collaborations even after the annotation was over. And I think more things like that ought to be envisioned in terms of annotating proteomics data, for instance.

One of the contributions of the Human Genome Project to science was creating a sea change in the way science was done. In the post genome or genome era as some people say, not only has science itself changed in terms of technological sophistication and the scale upon which we can undertake investigations of nature, but also the sociology of science has changed. Increasingly it requires multidisciplinary teams to tackle these large problems, and this is something that biologists really weren't used to before the Genome Project created the necessity for it.

3. On changes needed in education

I think one of the deficits in education of biologically trained graduate students as post-docs is a lack of appreciation for I guess it's fashionable to say, the whole system these days, systems biology. After a couple of decades of very successful but very reductionist molecular biology, I think most students if they were asked to look through a microscope at a body tissue really couldn't tell if it came from the pancreas or the salivary gland. And I think it's hard to encapsulate all that knowledge in a single individual, so I think there's two ways to address that; a broader and more liberal education in the biological sciences, if you will, and also the realization and preparation for working in multidisciplinary teams. A lot of lip service is paid towards multidisciplinary teamwork, but we don't actually prepare students really well to engage in that kind of post genome research activity.

4. On informatics and the caBIG™ initiative

Genome era thinking not only resulted in a sea change in the sociology of science and the way it's done, but also the way in which the information is disseminated and shared. In the pre-genome era the primary and exclusive route for data sharing was publication and peer-reviewed journals. And now as a result of the vast amount of data that was publicly supported by the Genome Project for instance, it completely changed the landscape of how data is shared.

I think the age of genomics has demonstrated the value of a new form of data sharing and communication in science. No longer is the fundamental data siloed in particular databases or in publications, but rather shared based on the network of common standards and universal access. I think what this does is make the data not only available to its obvious audience, biomedical researchers in government, academia, and industry, but anyone who could have a creative new idea based on an analysis of this information.

And I think caBIG is really the right concept and the right design to maximize the benefit of this investment. caBIG will make sure that the data produced by this project is absolutely available to anyone who may have a unique insight into the field and accelerate the development of proteomics.

One of the great advantages to having a public domain data source like this is that you never know whose going to look at it and have a bright idea or a creative new insight that really advances the field. In a sense it is built for a particular audience, which in this case is a combination of the public and private sector, scientists in both academia and industry in government, however you just never know when some graduate student in another field or even a high school student will look at this and have some great idea.

Isaac Newton, despite the obvious genius that he was, said at one point in his career that if I have seen further, it's only because I have stood on the shoulders of giants, the people who came before. And I think the modern version of that is that no scientist exists as an island. In order to make the real breakthroughs today we too have to stand on the shoulders on giants, but those giants are not individual scientists of the past, but rather current scientists all connected collectively in the grid of information that exists and which caBIG is designed to support.

5. On the challenge of clinical samples

Cancer is one of those diseases that routinely yields a biospecimen in the normal course of diagnosis and treatment. And you'd think that would really be an advantage for cancer proteomics because we do have access to specimens, however they're really collected for diagnosis and not for research purposes, so the standardization of collection, the permission that's granted for the use of the specimens, this has not been standardized in a way that can really advance the entire field of proteomics. So it's the quality of specimens, it's the conditions around which we can use them for research, and lastly but not least is the sheer number of specimens. When you're studying clinical proteomics it's really not an experimental study where you take a model organism or a cell culture line and perturb it in some way and have a really good control. Human specimens are

much more challenging because there are so many variables. There are so many pre-analytic factors, as the biostatisticians like to say that can affect the outcome of your analysis.

So in a very fundamental sense, clinical proteomics is not really an experimental science, it's an observational science. And that's why large number of samples, epidemiologically-informed study designs are really important to make progress in this field.

6. On clinical proteomics

With regard to clinical proteomics, it's like the title of that movie *Back to the Future*, what a lot of bench scientists don't realize is that clinical proteomics is done every day in clinical chemistry laboratories and doctors' offices. For instance the way that one diagnoses a heart attack is to look for the release of certain heart specific or cardiac specific enzymes into the serum, and this is a routine clinical measurement. For instance, one detects liver disease by looking for liver enzymes that are released into the serum. So this whole notion of measuring biomarkers for disease and serum is not only not new, it's something that is done thousands and thousands of times every day in major hospital settings. What is new however, is the fact that we're not approaching these one protein at a time, but taking the proteome as a whole as our object of study. And also using different technologies to discover and detect these proteins.

7. On accelerating clinical applications

There are two ways I think about making better progress in proteomics. One is the training of the people coming into the field, and I can talk about that in a minute. The second are factors related to the sociology of research, big science projects in biology, teamwork, communication, timelines and deliverables, all those kinds of things that are required to, I think, maximize an investment of this kind. So let me go back first to the training aspect. What are we looking for in people to be successful in proteomics and I think they're four major things. Number one is an in-depth knowledge of the disease biology. It's really important because I think several decades of molecular biology has made us very reductionistic and I think we do need to step back and take a look at the whole system again. I think the second skill that people need are knowledge of protein, biochemistry, binding and kinetics. Again a thing that was taught back when I was in graduate school but was really supplanted by decades of very successful application in molecular biology. But now I think we have to again, go back to the future. Thirdly, biology is becoming ever more quantitative every single day and I don't think you can succeed in any "omics" any field without a good working knowledge of bioinformatics

and either biostatistics or, in the case of clinical proteomics, epidemiology. And the latter because sample size and study design are critically important for what are essentially observational studies. And last is a mastery of one of the prevailing technologies in the field of proteomics. And I've purposely mentioned technology last because technology waves come and go. It gets better – some things come around that are new approaches. And this changes so rapidly that I think if you're a good scientist, a good biologist, and can frame a good hypothesis, you'll figure out how to use the technology to answer those questions.

What CPTI is doing is not only leveling the playing field from a technological perspective, it's actually raising the level of that playing field so that when a new biomarker is discovered based on this technology approach and the associated informatics approach, it has a much greater chance of rapid transition into a clinical setting. CPTI will not only standardize, but improve the quality of validation criteria that's essential to rapidly advance biomarkers into a clinical setting.