Prof Jasminka Godovac Zimmermann
Proteomics@Moleular Cell Dynamics, UCL SLMS
j.godovac-zimmermann@ucl.ac.uk

High-throughput, big data and complexity in clinical proteomics: an interview with Jasminka Godovac-Zimmermann

Professor Jasminka Godovac-Zimmermann is Head of the Proteomics and Molecular Cell Dynamics Group at University College London, UK. Professor Godovac-Zimmermann trained at the Max Planck Institute of Biochemistry, Germany, and specialized in protein chemistry. Her research focuses on proteomics in cancer and systems biology. Here she talks about the clinical impact of her work and her hopes and predictions for how proteomics and diagnostics could work together in future.

What is the most exciting aspect of your current research?

The advent of high-throughput molecular analysis means that we can now look globally at the exquisitely intertwined genomic, transcriptomic, proteomic and metabolic networks that govern the complex adaptive systems called cells. Most efforts so far concentrate on defining the identities and amounts of the players, be it genes, proteins, RNA or metabolites. A different, crucial aspect that has been somewhat neglected is that cells are inhomogeneous with specialized organelles in different spatial locations that carry out many specific functions. This requires sophisticated cross-organelle communication between different subcellular spaces. We now have extensive data about the dynamic subcellular redistribution of multiple proteins involved in processes like DNA replication or response to oxidative stress. Focussing on the requirement for efficient subcellular spatial communication is giving new pictures of cellular functional organization and its connection to disease. For example, we have initial results suggesting that perturbation of cellular spatial control is a major contributor to dysfunction of breast cancer cells.

How might your recent work impact on clinical disease treatment/management?

New concepts about cellular function will lead to new approaches to development of treatments. If the major response of breast cancer cells to estrogen exposure is not changes in gene expression, but rather massive changes in subcellular spatial organization of proteins, we should be thinking about the dominant cellular response when trying to design pharmaceuticals for treatment.

What changes would you like to see made to address challenges associated with big data in your field?

At present there seems to be a schism between big data and other forms of research. For example, there are enormous numbers of large scale measurements of gene expression. In terms of analysing the fundamental organization of cells we think these kinds of measurements

have reached a plateau in their usefulness and that even more measurements will not help much. This can, for example, be seen in attempts to correlate genetic variability with complex diseases. We already have lots of examples of genes that are associated with a disease, but are neither necessary nor sufficient for the disease. This seems to be related to population differences, e.g. that a gene/protein is only "defective" in the context of the genome/proteome of an individual. This is leading to the concept that functional networks involving concerted action of many genes, proteins, metabolites, etc. need to be identified by combining many kinds of input information with clinical information. This gets us back again to incomplete understanding of basic aspects of cellular function – we can detect hundreds, even thousands of genes that correlate with some disease, but we very often don't have much idea of why this is so.

In some ways the most useful and reliable information about individual genes/proteins is still coming from more conventional, low-throughput approaches that incrementally analyse many functional aspects of the same gene/protein. For example, *de novo* prediction of gene/protein function from massive amounts of high throughput data has not been very successful and conventional low-throughput experiments regularly identify new, "unexpected" functions of proteins. We think that one reason for this relates to spatial organization of cells. Many proteins have multiple functions in different subcellular locations and a defining characteristic is the dynamic redistribution of proteins between different functions/locations in response to cellular environment. This often occurs with no change at all in the expression/abundance of the protein.

A schism arises because the high-throughput data is often filed away in databases and the computational, predictive efforts based on such data are often in terms of "global" parameters and are not analysed and presented in forms that "wet biologists" find useful or even consult in their research. Remember that in many high throughput data collections 90+% of proteins/genes show no significant change. It may be that at present the most useful result of the high-throughput studies would be to identify which proteins in which contexts the wet biologists should investigate in detail. Our recent experience is that while initial high coverage of large numbers of proteins is necessary for identifying "potentially interesting" proteins, then priority should be given to methods that provide deep coverage of abundance, transcriptional/ translational isoforms, subcellular location, PTMs, binding partners, etc. for about 500 proteins. This seems to be a sufficient number, even for strong cellular perturbations such as cell cycle arrest occasioned by blocking DNA replication or by oxidative stress. This tends to give lots of fragmented functional networks that are ripe for further wet biology characterization by the many specialists in particular functional subsystems, although it is hard to get journals to let one present them all in a single publication. A lot more thought and effort should go into how to integrate and coordinate the typically incomplete, fragmented large scale results with smaller scale studies. Maybe adapting recent efforts in systems biology markup languages to include things like dynamic spatial location would provide a useful interface between the two communities. This would facilitate contextual recording of "predicted" fragment networks in forms that wet biologists could use, allow recording of small scale results in an efficient form, record "confirmed" fragments that would be useful components in trying to construct more informative global networks, could be adopted by journals as a publication standard to facilitate presentation and interchange, and could be collected in open access libraries.

How do you envision proteomics taking a more holistic approach in future?

Proteomics is already both holistic and fine grained. For example, we can already monitor abundance for most of the roughly 10,000 proteins that are used in any given cell type. Because transcription, translation and degradation all contribute to abundance of proteins in cells, this gives more complete information on the state of cells than genomics or transcriptomics measurements. We can already monitor at a global proteome-scale crucial processes such as PTMs involved in signalling systems that are invisible to genomics methods. We are making progress towards proteome-scale measurement of the dynamic spatial distribution of proteins. I think we are currently on a path similar to what has happened with genomics and evolution. Massive amounts of new data are replacing the original concept of the genome as a read-only

memory with concepts of the genome as a read-write memory in which proteomic, and also metabolic, inputs re-shape the genome. Epigenetics is the short term reshaping of the genome and evolution the longer term reshaping. In some ways we might even regard the genome as a kind of flash memory that a complex adaptive system interacting with its environment finds useful in maintaining itself, but is re-written when maintenance makes that useful.

Sometimes the devil is in the details and proteomics is very good at focussing on detailed changes for a few target proteins. Seeing all the details during global measurements is still beyond our current technology. We will get better at including more of the details in global measurements, but I suspect that in many cases it will be more productive to flesh out the global picture with targeted in-depth studies that focus on subsystems. From the standpoint of practical applications in therapy and diagnostics, I think it remains to be seen whether global features or local, detailed features are the most useful. Probably the most efficacious level for both therapy and diagnostics is an intermediate regime of functional subsystems that smooth out differences between billions of individuals. At the moment we have a kind of paradox that we can increasingly measure individual genomic differences in excruciating detail, but we often don't know enough about the intermediate regime of cellular function to make those differences useful in the development or application of therapy and diagnostics. So far proteomics seems to be the technology that best monitors concurrently many different aspects of the intermediate regime and it therefore has a crucial role to play in developing both therapy and diagnostics.

What is your proudest achievement in your career so far?

Difficult question. I think that I am most satisfied to have made a useful contribution towards new concepts during each epoch of my career. I had the good fortune to do my PhD and continue to work at Max Planck Institute in Munich in contact with scientists of the calibre of Perutz, Goodman, Mayr, Edman and Braunitzer. Protein structural biology was still in its infancy and I was given the job of determining the sequence of avian hemoglobins so we could think about why some birds can fly so high or don't suffer from rapid changes in altitude (oxygen pressure). That was an early beginning of fields like computational structural biology that eventually led to further Nobel prizes. But I left Munich and in Canberra I went in a different direction to look at transport of insoluble molecules like retinoic acid. We defined a new family of related protein structures called lipocalins[1] that made it into the textbooks as important physiological transporters of lipidic molecules. During the next stages I established new methods for extracting G-protein coupled receptors from membranes and determined structures and post-translation modifications (PTMs) for several[2]. This got me interested in cellular signalling and we published some of the first papers on intertwined dynamic changes in cellular phosphorylation networks and on multiple, different PTM patterns for the same protein that represent different functional states. One aspect was competitive modification of phosphorylation or acetylation in RhoGdi Protein PTMs in signalling systems has since become an enormous field that has feverish activity and in which we still participate. I moved on to spatial aspects of cellular function and we published work on nucleo-cytplasmic trafficking of proteins, including things like glycolytic enzymes in the nucleus, well before this became wildly popular in areas like the molecular-level connections of hypoxia to cancer[3]. This brought me to our present attempts to better define the role of dynamic spatial distribution of proteins in cellular function.

What do you predict for the interplay between proteomics and molecular diagnostics in future?

I think there are three crucial components to molecular diagnostics: sensitivity, specificity and what I will call uniqueness. We have only been peripherally involved in molecular diagnostics, mostly when we worked a number of years ago on extending the sensitivity of antibody detection of proteins to low femtomole or even attomole amounts with multi-photon detection methods. At the time we satisfied ourselves that proteomics methods have sufficient sensitivity for effective medical molecular diagnostics and recent developments in areas like quantum dots certainly confirm this. Specificity for large numbers of antibodies and very large variations in the

cellular abundance of different proteins are crucial questions in efforts to use antibody-based proteomics for global monitoring of cellular function, especially for analysis of the quantitative aspects of highly coupled networks. These are difficult problems that are mostly skirted in diagnostics applications since highly specific antibody detection of panels of limited numbers of proteins can be optimized. I think the real question about the future of proteomics in molecular diagnostics has to do with uniqueness. If potentially critical aspects of cellular function such as PTMs and their coupling to dynamic spatial organization of cells are largely invisible to the very efficient detection methodology of genomic methods, can proteomics monitoring of limited numbers of proteins provide detection of features that are unique to specific diseases and not otherwise detectable? It is already pretty evident that single proteins/genes are often only moderately helpful and that panels are needed. At present this seems to be an open question, but I think there are good prospects for proteomics. I would predict that the key once again lies in the intermediate regime of functional networks mentioned above. We need diagnostics that efficiently monitor the functioning of the intermediate subsystems, maybe even independent of the details of genetic variation, infectious agent or changes accumulated over a lifetime. This is potentially where a critical symbiotic coupling between diagnostics and therapy really comes into play and where proteomics could be crucial.

1. Godovac-Zimmermann, J. The structural motif of b-lactoglobulin and retinol-binding protein - a basic framework for binding and transport of small hydrophobic molecules (1988) Trends. Biochem. Sci. 13, 64-66
2. Roos, M., Soskic, V., Poznanovic, S. & Godovac-Zimmermann, J. Post Translational Modifications of Endothelin Receptor B from Bovine Lungs Analysed by Mass Spectrometry (1998) J. Biol. Chem. 273, 924-931
3. Baqader, N., Radulovic, M., Crawford, M., Stoeber, K., Godovac-Zimmermann, J. Nuclear-Cytoplasmic Trafficking of Proteins is a Major Response of Human Fibroblasts to Oxidative Stress (2014) J Proteom Res, 13, 4398-4423