max planck institute of biochemistry

Public relations

Press Release, May 27, 2019

dr. christiane menzfeld

phone: +49 89 8578-2824

pr@biochem.mpg.de
www.biochem.mpg.de/news
@MPI_Biochem

Learning from spectral experience

The protein composition of cells depends on the cells' function and current condition. Mass spectrometry (MS) can determine the identity and quantity of the proteins found in a sample. However, the data analysis of this method is time- and resource-intensive. Researchers at the Max Planck Institute of Biochemistry (MPIB) have collaborated with data science specialists from Verily in the USA to develop a machine learning approach – continuously self-improving algorithms – to facilitate the analysis of mass spectrometry data. Their results, which simplify MS applications and also led to the discovery of new chemical patterns in proteins, are published in the journal *Nature Methods*.

The final trip of many old cars leads them to a junk yard where they are deconstructed into potentially salvageable parts. By looking at the entirety of car components, a trained and experienced employee of the junk yard may be able to reconstruct the identity of the scrapped car. Mass spectrometers (MS), which identify and quantify proteins in a sample, are like molecular junk yards. The proteins are first broken into smaller fragments – the peptides. The information on the identity and the abundance of the peptides can be recorded as mass spectrometry spectra. To reconstruct the information on the proteins in the analyzed sample, the characteristics of these spectra are then compared to previously recorded libraries – a process that requires a lot of computational power.

Machine learning supports data analysis
In cooperation with Verily, the life sciences company of Alphabet, researchers at the MPIB have now developed the model *DeepMass:Prism* to facilitate the interpretation of mass spectrometry spectra., They used machine learning to train algorithms to "translate" proteins into MS spectra. The translation of these abstract data is challenging for artificial intelligence and best performed by "deep learning" algorithms. Similar deep learning approaches are used in the automatic translation of languages. But rather than translating from English to German or *vice versa*, *DeepMass:Prism* is trained to translate between proteins and the spectra that are usually generated in MS analysis.

Am Klopferspitz 18
82152 Martinsried

Tel.: 0049 (0)89 / 8578 - 0
Fax.: 0049 (0)89 / 8578 - 3777

www.biochem.mpg.de
www.mpg.de

MAX-PLANCK-GESELLSCHAFT

"The key to success in this project was the fusion of our expertise in mass spectrometry with Verily's expertise in deep learning, particularly in the fields of biology and life sciences.", says Jürgen Cox, independent group leader at the MPIB. Their program *DeepMass:Prism* was trained with more than 60 million peptide spectra from publicly accessible data bases. The program recognizes patterns from the training spectra and applies them to the analysis of new samples.

The computational biologist Cox highlights that *DeepMass:Prism* improves different applications of mass spectrometry. One potential use of MS is the characterization of samples whose composition is entirely unknown. The new algorithms can increase the number of peptides that are identified in this approach. Alternatively, large groups of samples with a similar general composition can be compared regarding individual differences in the protein quantity. For instance, blood samples from patients generally have a similar protein composition, but it is important to detect altered protein levels to diagnose diseases. "This is where our *DeepMass:Prism* has made the greatest strides", says Cox. "Rather than experimentally determining the reference libraries to which the samples are compared, the model can now predict them – a shortcut that saves a lot of time and resources. "

Finding needles in the peptide haystack

The more than 200 cell types in the human body are characterized by the presence of different proteins, but also by the differences in the abundance of identical proteins. The different quantities of proteins are particularly challenging for MS analyses. Jürgen Cox explains the importance of measuring protein quantities with an automobile analogy: "When you completely disassemble a car, the piles of parts can look quite similar. Therefore, knowing the abundance of certain parts can help in the identification. When you find 6 cylinders in a pile, you know that it can not be a car with a four-cylinder engine." Similarly, finding only three tires in a pile shows a potential damage of a car. The same principle can be applied to the analysis of cells or tissues. Diseases can cause certain proteins to be more or less abundant than in healthy control samples.

Many diagnostic procedures rely on the measurement of proteins in patient samples by mass spectrometry. "We need highly accurate MS to discover new biomarkers – indicators of disease. Sometimes, even a small variation of a certain biomarker can signal disease progression, therefore the prediction must be precise and reproducible.", says Peter Cimermancic, Senior Scientist at Verily. With *DeepMass:Prism*, the researchers strongly improved the correlation between the predicted spectra and the actual measured spectra. He is optimistic that the model will lead to the development of new diagnostic tools.

Am Klopferspitz 18
82152 Martinsried

Tel.: 0049 (0)89 / 8578 - 0
Fax.: 0049 (0)89 / 8578 - 3777

www.biochem.mpg.de
www.mpg.de

MAX-PLANCK-GESELLSCHAFT

Even though *DeepMass:Prism* was not trained with chemical knowledge, it discovered new chemical rules that determine how the peptides break into smaller fragments. "The previous library-based approach could only reproduce what it already knew. *DeepMass:Prism* is able to generate new knowledge by combining information and drawing its own conclusions. This is a very exciting finding", says Cox, "it's like a junk yard employee who understands where a certain part of the car is installed, even though he has never seen this type of car before. The predictions by *DeepMass:Prism* have led to the identification of a new kind of interaction within proteins. We believe that this discovery is only the beginning of what deep learning can do for research in life sciences." *DeepMass:Prism* will be available for download on Google cloud. [CW]

Original Publication:
S. Tiwary*, R. Levy*, P. Gutenbrunner*, F.S. Soto, K. Palaniappan, L. Deming, M. Berndl, A. Brant, P. Cimermancic and J. Cox: High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods*, May 2019 (*equal contributions)
DOI: https://doi.org/10.1038/s41592-019-0427-6
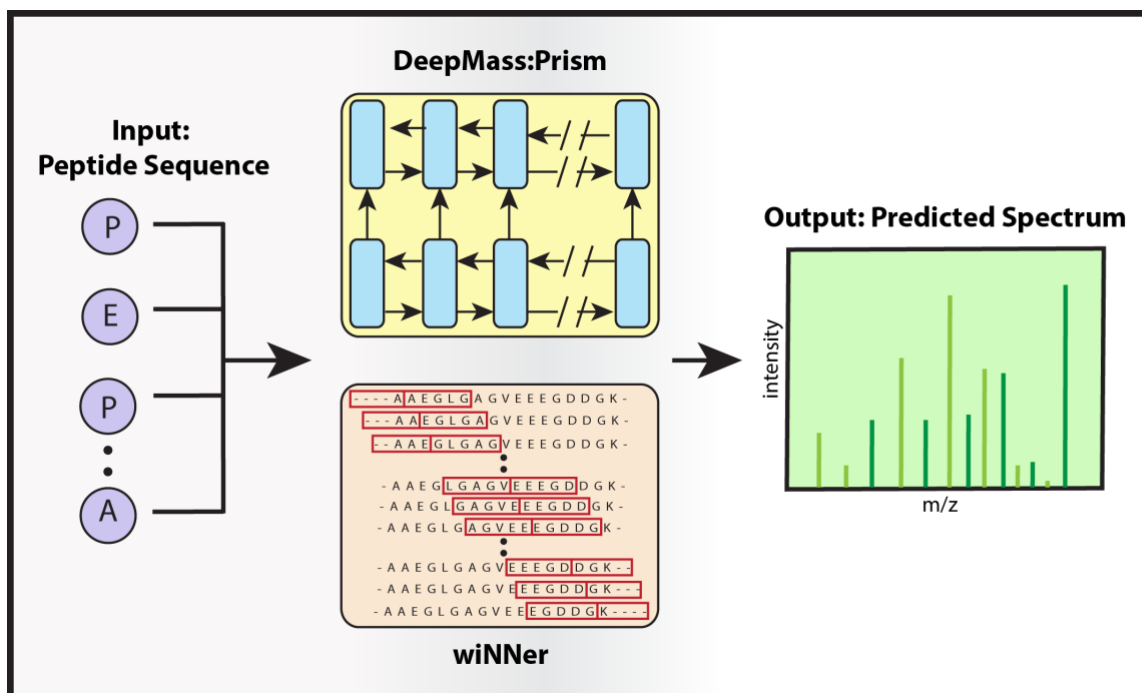
Caption:
DeepMass:Prism, a deep learning approach using bidirectional recurrent neural network (RNN) architecture for the prediction of fragment intensities.
Shivany Tiwary © MPI of Biochemistry

Am Klopferspitz 18   Tel.: 0049 (0)89 / 8578 - 0   www.biochem.mpg.de
82152 Martinsried   Fax.: 0049 (0)89 / 8578 - 3777   www.mpg.de

MAX-PLANCK-GESELLSCHAFT

About Jürgen Cox

Jürgen Cox studied physics in Aachen and received his PhD from the Massachusetts Institute of Technology, USA. After a Postdoc at the Technical University of Munich, he became a Senior Scientific Consultant for the bioinformatics company Genedata. In 2006, he joined the MPIB as a senior scientist and has been the independent leader of the research group "Computational Systems Biochemistry" since 2014. Cox was appointed as Honorary Professor of Proteomics at the University of Copenhagen, Denmark, in 2013.

About Peter Cimermancic

Peter Cimermancic received his BSc degree in Biochemistry from the University of Ljubljana, Slovenia, in 2009. He was awarded the Howard Hughes Medical Institute Predoctoral Fellowship at University of California, San Francisco where he received his PhD in Biomedical Informatics in 2014, under the supervision of Prof. Andrej Sali. He developed and applied computational methods to several problems in proteomics, genomics, drug discovery, and structural biology. Currently, he is a Senior Scientist at Verily Life Sciences, utilizing machine learning for life science and health-related challenges.

About the Max Planck Institute of Biochemistry

The Max Planck Institute of Biochemistry (MPIB) belongs to the Max Planck Society, an independent, non-profit research organization dedicated to top-level basic research. As one of the largest Institutes of the Max Planck Society, about 800 employees from 45 nations work here in the field of life sciences. In currently about 35 departments and research groups, the scientists contribute to the newest findings in the areas of biochemistry, cell biology, structural biology, biophysics and molecular science. The MPIB in Munich-Martinsried is part of the local life-science-campus in close proximity to the Max Planck Institute of Neurobiology, a Helmholtz Center, the Gene-Center, several bio-medical faculties of the Ludwig-Maximilians-Universität München and the Innovation and Founding Center Biotechnology (IZB).
http://www.biochem.mpg.de

Contact:

Prof. Jürgen Cox, PhD
Computational Systems Biochemistry
Max Planck Institute of Biochemistry
Am Klopferspitz 18
82152 Martinsried/Munich
Germany

E-mail: cox@biochem.mpg.de
www.biochem.mpg.de/cox

Dr. Christiane Menzfeld
Public Relations
Max Planck Institute of Biochemistry
Am Klopferspitz 18
82152 Martinsried/Munich
Germany
Phone: +49 89 8578-2824
E-mail: pr@biochem.mpg.de
www.biochem.mpg.de
Twitter: @MPI_Biochem

Am Klopferspitz 18
82152 Martinsried

Tel.: 0049 (0)89 / 8578 - 0
Fax.: 0049 (0)89 / 8578 - 3777

www.biochem.mpg.de
www.mpg.de

MAX-PLANCK-GESELLSCHAFT