The **Protein Puzzle**

The human body consists of tens of thousands of proteins. What's more, these occur in several variants whose concentration in the organism can change over time. **Matthias Mann** from the **Max Planck Institute of Biochemistry** in Martinsried therefore needs clever algorithms and a lot of computing power for his research. His goal, after all, is to decode the entire human proteome – that is, the full set of proteins in the human body – for the benefit of medical science.

TEXT **TIM SCHRÖDER**

Matthias Mann's lab in Martinsried is as tidy as a hospital intensive care unit. The glass walls and doors provide a clear view, and several identical-looking workstations are set up around the room. At each station, a robotic arm hanging from the ceiling picks up small plates containing samples and places them in the instruments. Technically speaking, these systems represent the essence of what Matthias Mann has been developing over the years: machines capable of processing and analyzing proteins at breathtaking speed.

Mann is interested in these biomolecules because they are involved in almost all of an organism's biological processes. Proteins are made up of an amino acid molecule chain that folds into complex three-dimensional patterns. Some – like keratins in skin and hair cells – serve as structural substances. Other proteins, known as enzymes, speed up metabolic reactions: they convert fat into energy, for example, or make oxygen available to cells as a source of energy. If we subtract the body water content from body weight, most life forms consist of 50 percent protein. Without proteins, there would be no life on earth.

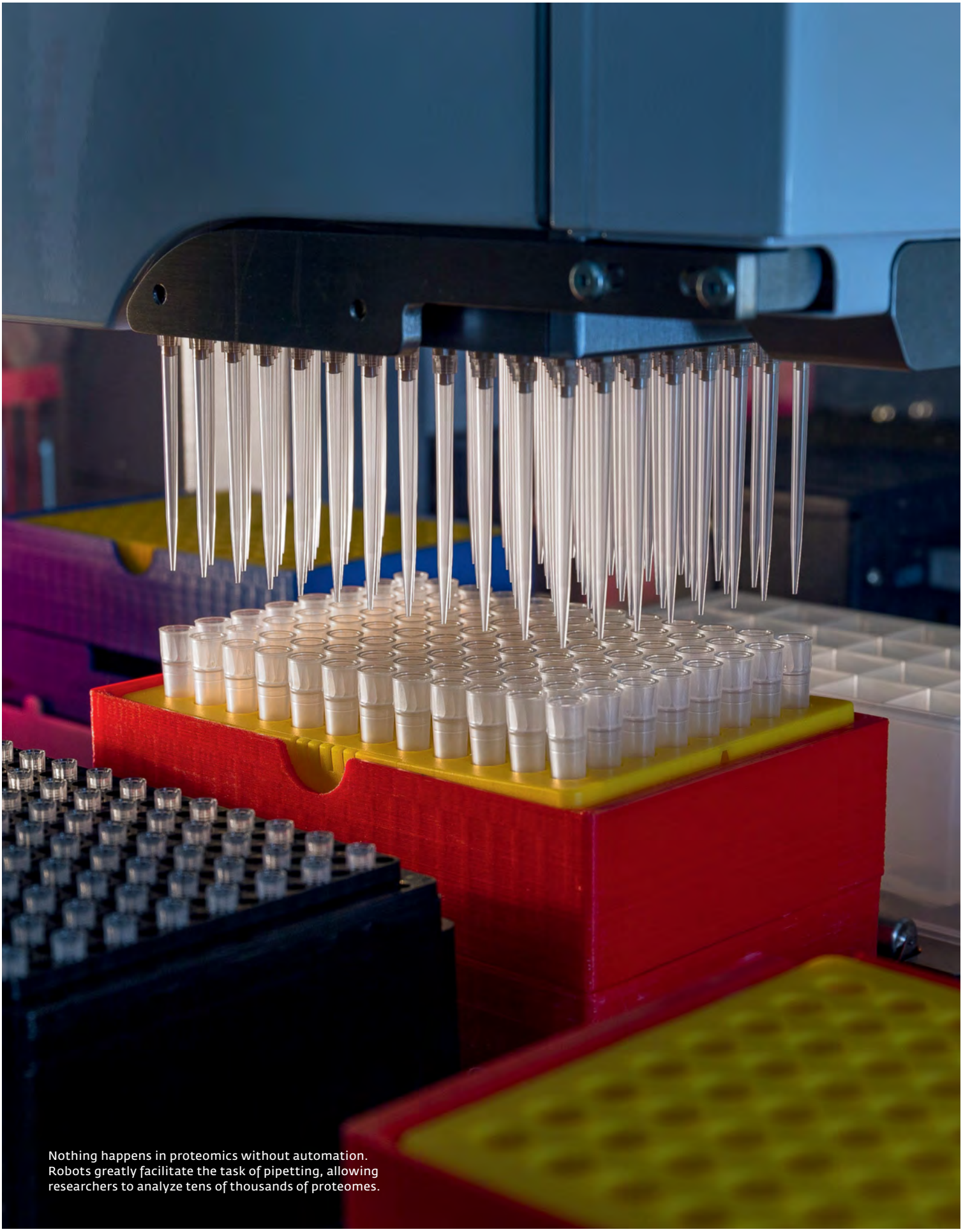
A MATURE TECHNOLOGY

Matthias Mann initially studied physics and mathematics, but he has been investigating the world of proteins since the 1980s. "It has taken two decades to develop the technology that allows us to analyze proteins in a reasonable manner," he says. "We've reached the point where we can actually apply the technology – and now things are getting really interesting!" The complex nature of protein analysis became evident when the human genome was sequenced in 2001. The scientists working on the Human Genome Project identified around 20,000 genes that encode the blueprints for around the same number of proteins.

Only gradually did it become clear that these proteins occur in many variants. Once a gene has been read, parts of the messenger RNA, which serves as a protein template, can be cut out. This gives rise to RNA molecules of various lengths, each of which is translated into a different protein. Numerous proteins, in turn, must be trimmed before they can be used as finished molecules, while chemical tags are appended to others. All in all, there are hundreds of thousands of protein variants that interact in a finely orchestrated choreography.

In addition, whereas an organism possesses the same genes throughout its lifetime, the protein composition varies according to cell type. Some proteins occur in large quantities and in every cell, while others occur only in trace amounts and only in certain tissues. >

Photo: Axel Griesch



Nothing happens in proteomics without automation. Robots greatly facilitate the task of pipetting, allowing researchers to analyze tens of thousands of proteomes.

Under the watchful eye of Heiner Koch, researchers Florian Meier and Scarlet Beck (left to right) analyze protein samples using the six mass spectrometers set up in Matthias Mann's (far right) laboratory alone.

Photo: Axel Griesch

Different proteins are active depending on the tasks the metabolic machinery is carrying out. "If we want to know how the metabolic machinery works or what its momentary state is, we must be able to analyze the protein profile in a tissue as well as how it changes," Mann says.

Scientists have always realized how important proteins are for body processes. Now, however, a growing number of researchers have become interested in studying the complete set of proteins in the body. This eventually resulted in the emergence of the research area we now call proteomics. But enormous quantities of data are required to analyze the protein composition of any life form. A huge volume of data must also be collected for evaluating and interpreting the results. Proteomics thus relies on sophisticated data processing.

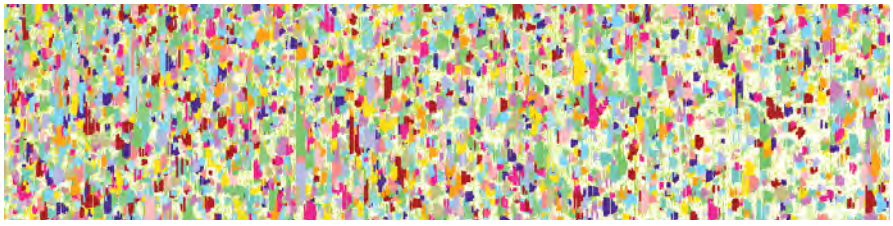
A major problem in analyzing proteins is that they are very sensitive. As anyone who has ever boiled or beaten an egg knows, the three-dimensional structures of proteins collapse when they are heated or mechanically stressed, and they clump together. It was therefore not possible to analyze proteins with the conventional method of mass spectrometry. Mass spectrometers are used to analyze samples containing unknown constituents, for instance to detect toxins in tissue samples.

DEFLECTION IN AN ELECTRIC FIELD

Before proteins can be investigated in a mass spectrometer, they must be converted into charged particles, for example by bombarding them with electrons or other charged particles, which converts them into electrically charged ions. Only electrically charged molecules are deflected from their path through the electric field of the mass spectrometer. The amount they are deflected depends on their charge magnitude and molecular weight, and this information allows scientists to infer a molecule's identity.

However, conventional mass spectrometry with ionization is too harsh a method for the sensitive proteins. In the lab of his doctoral supervisor, John B. Fenn, at Yale University, Matthias Mann, together with other colleagues, developed a gentler solution in the early 1980s. With the help of trypsin, a digestive enzyme, they snipped proteins into approximately ten-amino-acid-long fragments – so-called peptides. They then sprayed the peptides through a fine tube, imparting an electrical charge to them. Using this electrospray ionization method, they were able to analyze proteins for the first time in a mass spectrometer – a method for which Fenn was awarded the Nobel Prize in Chemistry in 2002.

The time had now come for computer scientists, as it is almost impossible for a human being to deduce the original proteins from a peptide mix. Together with his colleague Jürgen Cox, Mann developed an analytical program called MaxQuant, which can compare many thousands of peptide fragments with information contained in international databases. In addition to the



A section of the protein spectrum of a cancer cell: The proteins are first cut into peptides of varying molecular weight (shown in different colors), then separated from each other in chromatographic columns and subsequently analyzed in a mass spectrometer.

molecular weights of every conceivable peptide, the databases contain information about the protein to which each fragment belongs. MaxQuant compares the data from the mass spectrometer with the content of the databases and reconstructs the protein composition of a sample from the results.

CHROMATOGRAPHY IN MINIATURE FORMAT

Nevertheless, electrospray ionization and MaxQuant together were unable to overcome all the obstacles of protein analysis. For example, they could not detect proteins that are present in a sample only in trace amounts. This was because the researchers needed a relatively large liquid sample to separate the proteins from the other components by chromatography before the proteins are analyzed in a mass spectrometer. As a result, the trace proteins were excessively diluted and could no longer be detected.

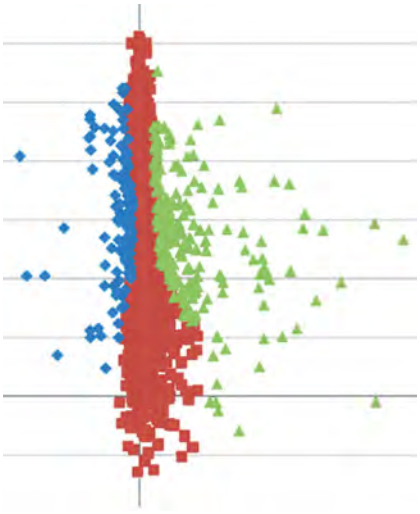
Mann miniaturized the tubes used in this type of chromatography down to a diameter of just a few nanometers. With these tubes, he needs only minute quantities of a sample: "Our nanochromatography enables us to obtain sufficiently high concentrations of proteins that are present in the sample in vanishingly small amounts."

By combining electrospray ionization, MaxQuant and nanochromatography, Mann succeeded in doing what had previously seemed impossible: decoding an organism's entire proteome. In 2008, the researcher analyzed the proteome of an entire organism, identifying the 4,399 proteins in a yeast fungus. He and his Max Planck colleague Frank Schnorrer achieved their latest breakthrough in 2016, when they decoded the fruit fly proteome. They discovered that the tiny insect has around 10,000 proteins. "By comparison, there are around 13,000 proteins in the mouse brain," Mann explains, and only 10 percent of them are limited to specific cell types.

Mann's method has since become the standard in proteomics laboratories around the world. Without the help of

The electrically neutral proteins are first cut into smaller peptide fragments. In order for the peptides to be deflected along different trajectories by the electrical field of the mass spectrometer, they must be electrically charged. This is accomplished by the electrospray ionization method, in which the peptides are given an electric charge in a metal capillary tube and then sprayed out at the tip of the tube.

Photos: MPI of Biochemistry (top), Axel Griesch (bottom)



Baker's yeast was the first organism whose complete proteome was analyzed. The unicellular organism can reproduce both sexually and asexually. The graph shows the ratios of proteins in both reproductive modes. Some proteins, such as protein attractants, are amplified during sexual reproduction (blue), while others are produced mainly during asexual reproduction (green). Still others are unaffected by the reproductive mode (red). Although all cells have identical complements of genes, they often use a completely different protein repertoire to carry out their specific functions.

computers, its use would still be unthinkable. Cox and Mann have refined the MaxQuant software so that they can now determine not just the identity, but also the quantity of proteins in a sample. Thus, multiple samples from a patient can be compared to determine how the concentration of a specific protein changes with time.

However, before proteomics can find application in hospitals, the methods have to be speeded up even more. "We're currently optimizing our system's workflow to allow us to analyze as many samples as possible. A year from now, we expect to be able to analyze 100 proteomes a day," Matthias Mann predicts. Researchers will then be able to study, for example, how a patient's protein concentration changes during the course of a day or as a disease progresses.

The scientists are already able to compare groups of people to determine differences in the metabolism of sick and healthy patients. To this end, Mann, together with doctors at Copenhagen University Hospital, studied how the proteins of obese people change during an eight-week diet.

NOT EVERYONE RESPONDS TO A DIET IN THE SAME WAY

The body reacts to obesity as if it were an inflammation, producing proteins that are typical of inflammatory reactions. The researchers wanted to know whether the quantity of inflammatory proteins decreases the same amount in all patients during a diet. Mann and his colleagues analyzed more than 1,000 proteomes and determined the quantity of inflammatory proteins using the MaxQuant software. They found that inflammatory proteins don't decrease at the same rate in everyone, even if they follow the same diet. In other words, not everyone responds the same.

Proteome analysis is complicated by the many variants that proteins can occur in. The protein ubiquitin, for example, binds to aging or defective proteins, initiating a breakdown process during which the protein is gradually dismantled. Moreover, many proteins are activated by tagging them with a phosphate molecule – a process known as phosphorylation.

Mann was able to show that an organism's day-night rhythm crucially depends on the phosphorylation state. "There are an enormous number of protein variants whose importance we still don't understand. Moreover, entire groups of proteins can take on different states. But it is precisely these changes in a patient's proteome that are decisive when it comes to treating diseases," says the scientist, who therefore thinks very little of some of today's diagnostic tests.

As an example, he mentions the PSA level, which can be an indication of prostate cancer, but which is controversial owing to its unreliability. "Such tests show the presence or amount of a single protein. Based on what we now know, though, that isn't enough. In the future, we will rely much more heavily on an individual's proteome to gain an overview of his or her health status," Mann says.

Another program developed by his colleague Jürgen Cox is expected to help: Perseus, as the program is dubbed, uses the statistical protein data from MaxQuant to conduct a big data analysis. The software accesses international databases containing the accumulated fund of knowledge about proteins – for example, where specific proteins occur and what it means when the metabolic system increases production of certain proteins. Perseus also takes existing knowledge about diseases into account.

The proteome-based diagnosis and treatment of diseases is still in its infancy. It would be extremely difficult to find early signs of malignant skin cancer in the proteome because the tumor is still very small in the early stage, meaning that very little of the telltale protein is released. Such minute amounts

Graphic: Max Planck Institute of Biochemistry/Matthias Mann

Analyzing data on the computer (clockwise from front left): Jan Rudolph, Jürgen Cox, Camila Duitama, Pavel Sinitcyn and Art Carlson – and playing a relaxing game of chess during a break.

can't be reliably detected even with the help of nanochromatography.

Nevertheless, proteomics has clearly advanced remarkably since the turn of the millennium. After the human genome was decoded in 2001, many start-up companies began offering proteome analysis as a service to clinical researchers. In light of such methods as nanochromatography and electrospray ionization, it's clear that the technology of the time was utterly inadequate.

Accordingly, the results proved useless for routine clinical practice. Complaints soon followed. Many startups disappeared from the scene, and the term proteomics became a mere buzzword. "Our new techniques have brought us a giant step forward, but things are only now really getting off the ground," says Matthias Mann. How fortunate it was that he didn't focus exclusively on computer science and physics, and instead took an early interest in the biological questions that his research raised. As a result, he can now help harness the potential of proteomics for the benefit of biology and medical science. ◀

TO THE POINT

- **For proteome analysis, the proteins must first be cut up into peptide fragments. Only sophisticated computer algorithms can reconstruct the original proteins from the huge volumes of data generated.**
- **The MaxQuant program accesses databases that serve as repositories of knowledge about peptides and proteins.**
- **Using the Perseus program, researchers analyze information from databases on the occurrence and function of proteins. This sheds light on the role of individual proteins in disease processes.**

GLOSSARY

Protein modifications: The number of proteins an organism produces can be several times greater than the number of its genes. This enormous diversity results from changes after a gene has been read (transcription) or after messenger RNA has been converted into a protein (translation). In a process known as alternative splicing, for instance, sections of a messenger RNA molecule are cut out or moved, resulting in a number of different gene products. Small molecular tags, such as phosphate and sugar residues, are subsequently added to alter the function of proteins. When a gene encodes several proteins, or when an amino acid chain is subsequently cleaved into several proteins, multiple proteins are produced from a single gene. In humans, up to ten different proteins can be traced to a single gene.

Proteome: It is now estimated that the human body contains between 80,000 and 400,000 proteins. However, they aren't all produced by all the body's cells at any given time. Cells have different proteomes depending on their cell type. There are thus at least 250 different proteomes corresponding to the 250 cell types in the human body. The proteome depends on many factors. For example, different proteins may be produced depending on an organism's age, diet and state of health. The protein composition is also affected by environmental influences such as medications and pollutants.

Photo: Axel Griesch