The First Maps of the **Human Proteome**

A group of researchers led by Prof. Bernhard Küster has published one of two initial comprehensive maps of the human proteome. Following the first complete sequencing of the human genome in 2001, a second key to unlocking the mystery of life has now been found with the mapping of the proteome – in other words, all the proteins in the human body.

Link
www.wzw.tum.de/proteomics https://www.proteomicsdb.org/

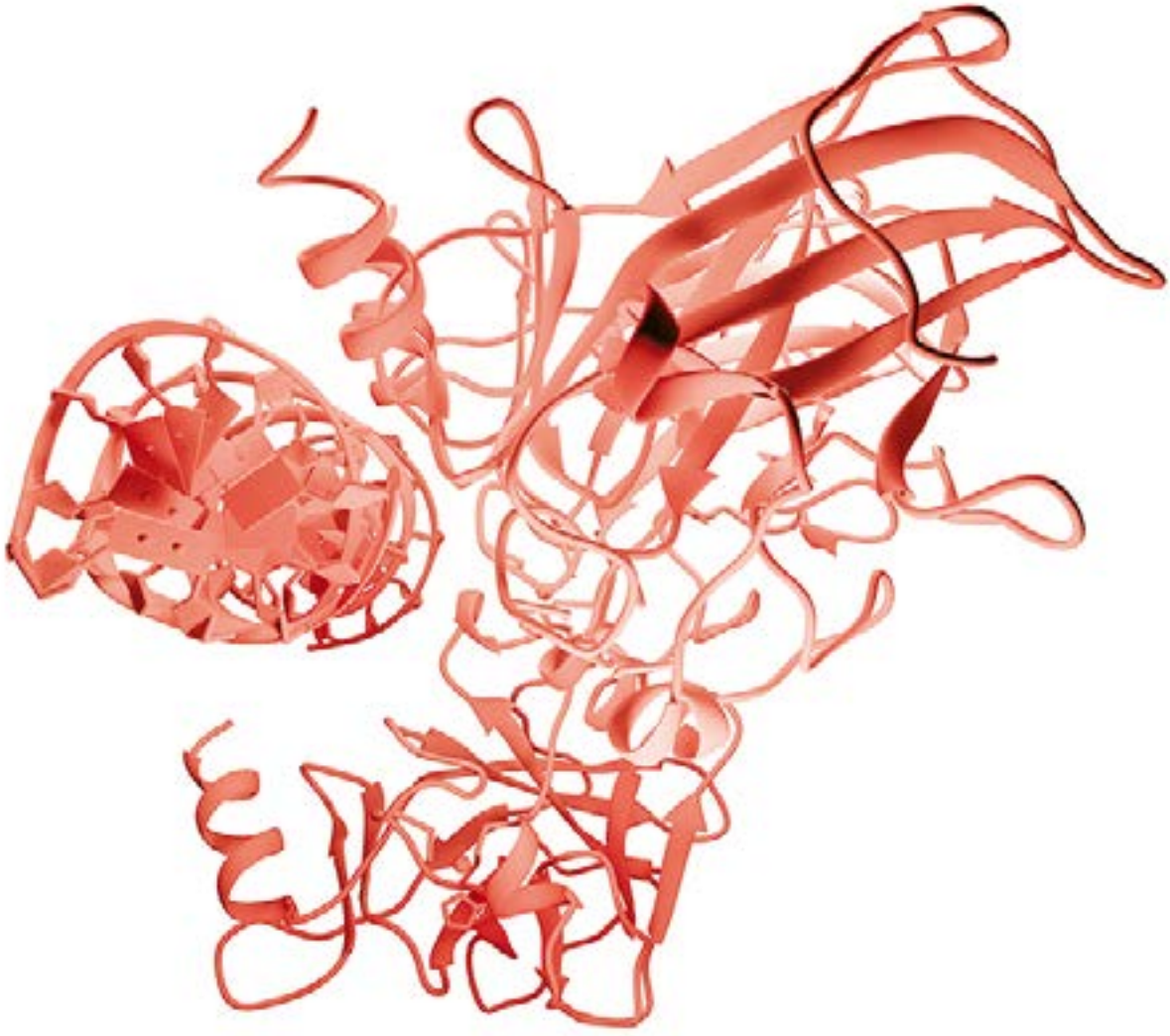
Die ersten Karten des menschlichen Proteoms

Eine Gruppe der TUM um Prof. Bernhard Küster hat im Mai 2014 zeitgleich mit einer Gruppe von US-Forschern eine der beiden ersten wirklich umfassenden Karten des menschlichen Proteoms vorgelegt – der Gesamtheit aller Eiweiße, die unser Körper bilden kann. Gene liefern den Bauplan für Proteine und die Münchener haben den Nachweis für 92% oder 18.097 der aus dem menschlichen Genom abgeleiteten Eiweißgrundformen erbracht. Diese Grundformen sind allerdings erst der Anfang, da Menschen über eine Vielzahl von Mechanismen verfügen, um Proteine verschiedenen Bedürfnissen entsprechend abzuwandeln. Dennoch konnten die Forscher bereits eine Reihe fundamentaler Erkenntnisse aus den vorliegenden Proteinkarten gewinnen. So wurden offensichtlich hunderte Gene im Lauf der Evolution stillgelegt da für sie keine Proteine mehr zu finden sind. Gleichzeitig scheinen neue Proteine im Entstehen begriffen zu sein, die bislang gänzlich unbekannt waren.

Die Arbeit wurde vor allem durch zwei methodische Fortschritte möglich: Zum einen erlaubt die Massenspektrometrie heute Spezialisten, binnen weniger Tage das Proteom menschlicher Gewebe zu analysieren und dies zu Kosten von wenigen 1.000 Euro. Zum anderen ermöglicht eine von der Küster-Gruppe zusammen mit der Firma SAP entwickelte Datenbank der internationalen Forschergemeinschaft, ihre bislang in vielen Einzeldateien verstreuten Analyseergebnisse zusammenzutragen und gemeinschaftlich auszuwerten.

Im Fokus steht auch der medizinische Nutzen. So konnten Bernhard Küster und sein Team anhand ihrer Daten bereits die Wirksamkeit von Medikamenten aus dem Proteinprofil von Krebszellen vorhersagen.

Langfristig wollen die Forscher über das biologische Verständnis des Proteoms und des Genoms die personalisierte und zielgerichtete Therapie von Patienten weiter voranbringen.



Picture credit: RCSB Protein Data Bank

The protein p53 is encoded by the TP53 gene and functions as a tumor suppressor. It is also known as the “guardian of the genome” because it is intimately involved in repairing damaged DNA. In most human cancers, TP53 is mutated, rendering it functionless.

Decoding the Human Proteome

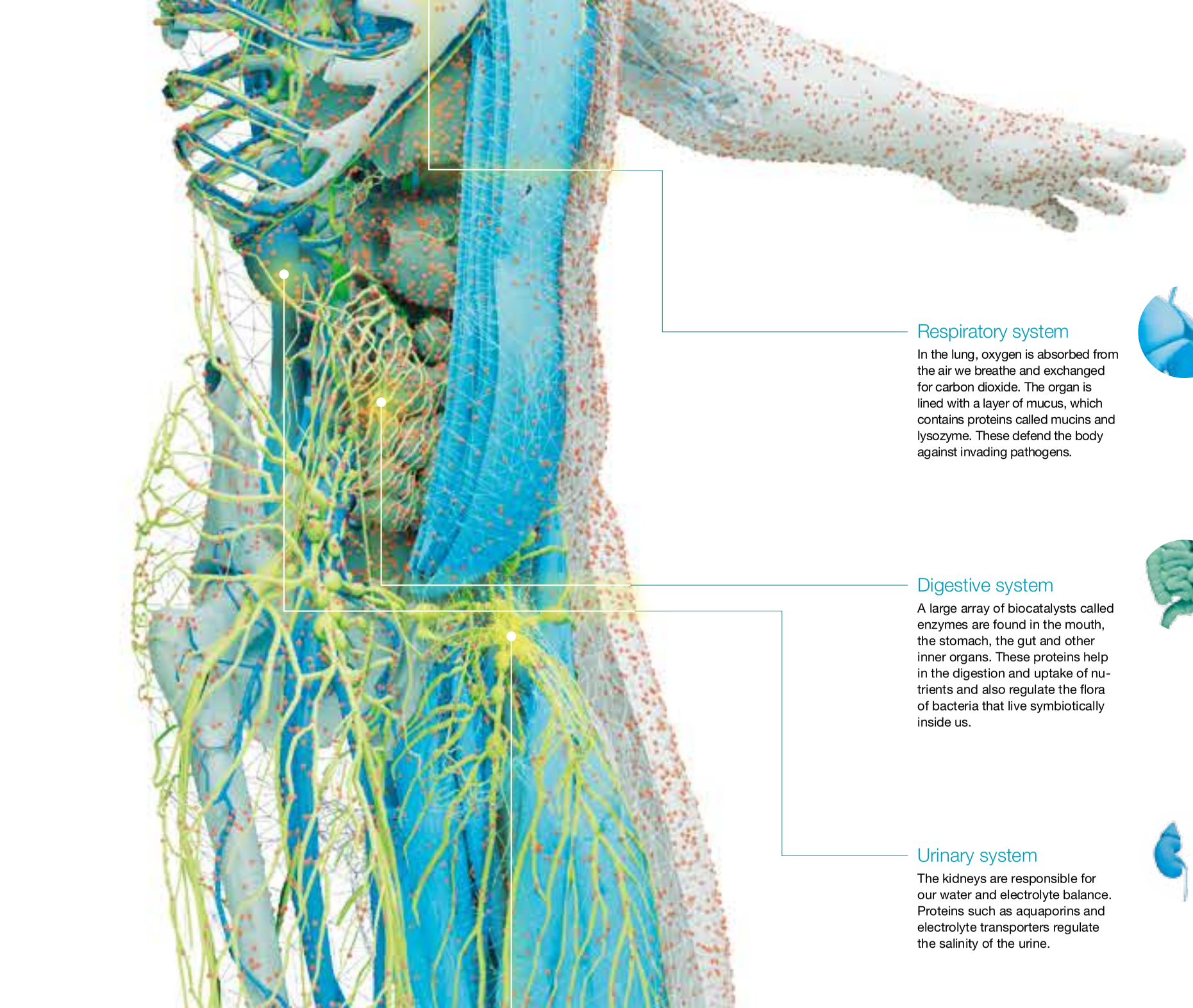
Each human cell contains the same genetic information, including 20,000 protein coding genes. It is the repertoire of expressed proteins – the proteome – that gives the many different cell types in our body the specific characteristics required to perform their diverse functions.

Nervous system

Nerve cells are wrapped in a blanket of myelin, which is required for the efficient transmission of neural signals. Myelin contains mainly lipids, but also proteins, such as the myelin basic protein.

Lymphatic system

Lymph nodes are a reservoir for immune cells and contain a spectrum of antibodies and immune proteins that guard human beings from pathogens, such as viruses and bacteria.



Respiratory system

In the lung, oxygen is absorbed from the air we breathe and exchanged for carbon dioxide. The organ is lined with a layer of mucus, which contains proteins called mucins and lysozyme. These defend the body against invading pathogens.

Digestive system

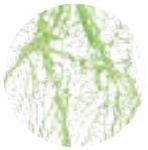
A large array of biocatalysts called enzymes are found in the mouth, the stomach, the gut and other inner organs. These proteins help in the digestion and uptake of nutrients and also regulate the flora of bacteria that live symbiotically inside us.

Urinary system

The kidneys are responsible for our water and electrolyte balance. Proteins such as aquaporins and electrolyte transporters regulate the salinity of the urine.

Reproductive system

Important hormones such as estrogen and testosterone are produced in the ovaries and testes. Interestingly, these two organs also show the broadest spectrum of protein expression of all organs in the body.



Circulatory system

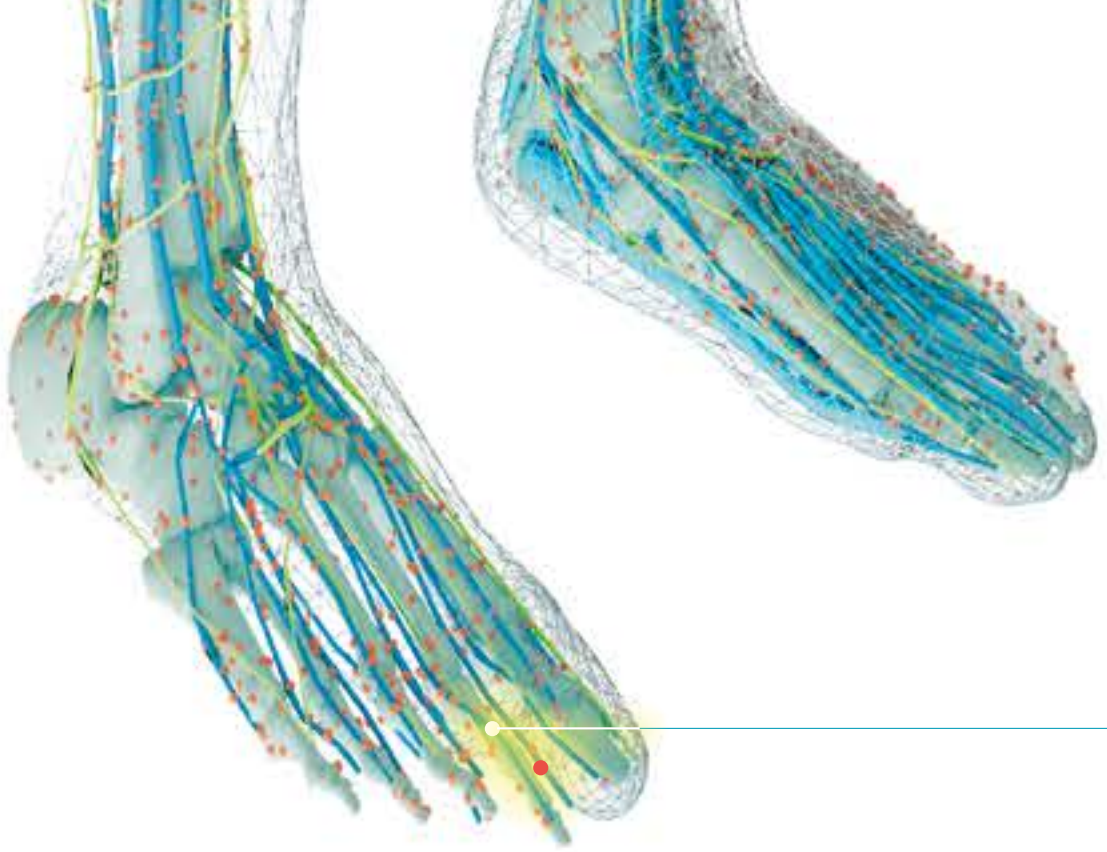
Hemoglobin is the protein that transports oxygen from the lungs to every cell of the body. There are about 5 liters of blood in adults, and oxygenated hemoglobin gives blood its distinctive red color.

Integumentary system

The skin is the largest organ in the body and contains large quantities of fibrous and waterproof proteins called keratins, which provide protection against all kinds of damage, including mechanical force or rain. Keratins are also the major constituents of hair and nails.

Muscular system

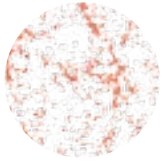
Actin and myosin are proteins that constitute a large part of the 5-6 kg of muscle proteins in the human body. These proteins give muscles the ability to contract and thus create movement.



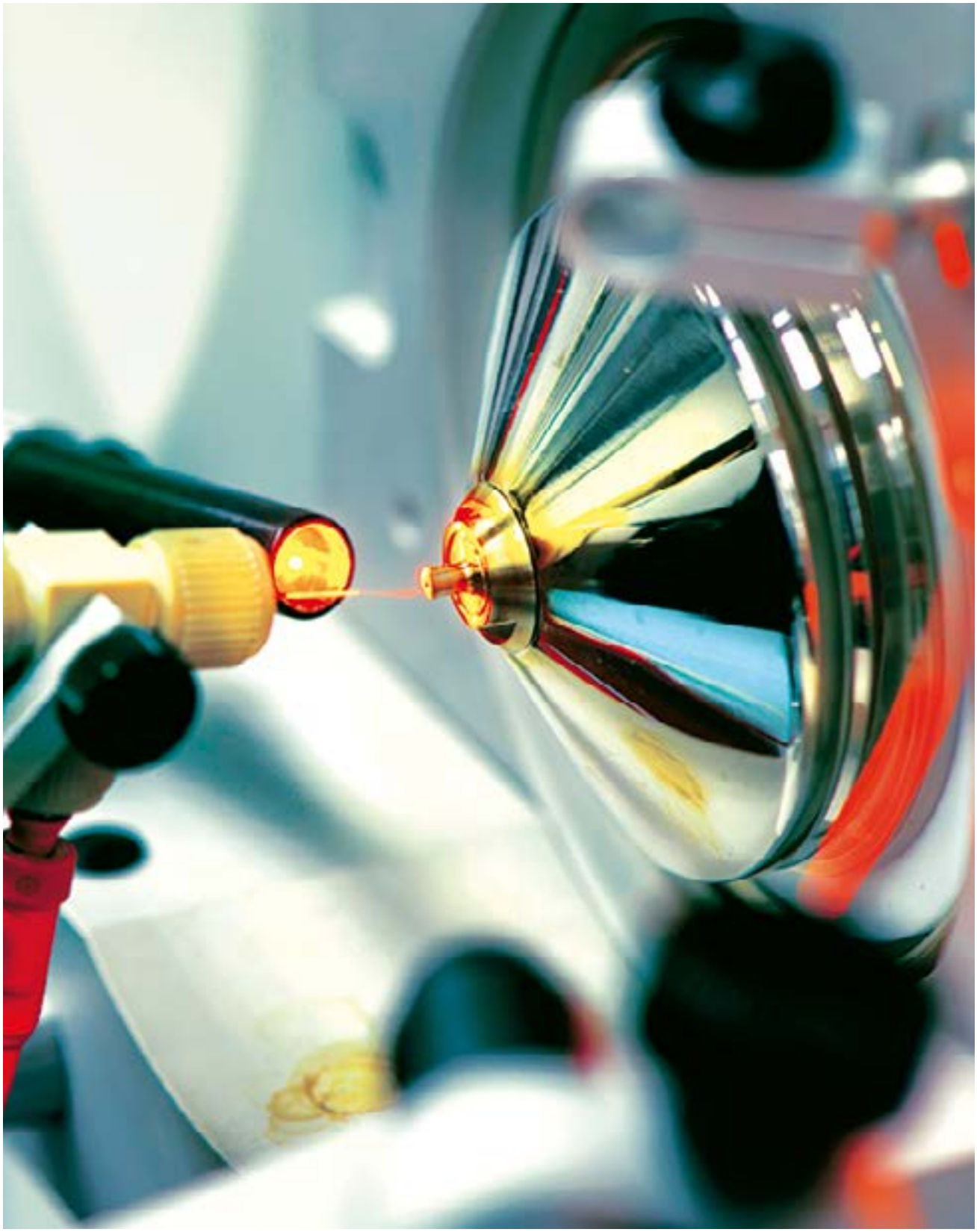
Skeletal system

Collagen is a structural protein that provides strength and flexibility to bones, tendons, ligaments and skin. A staggering 25 percent of all the body's proteins are collagen.

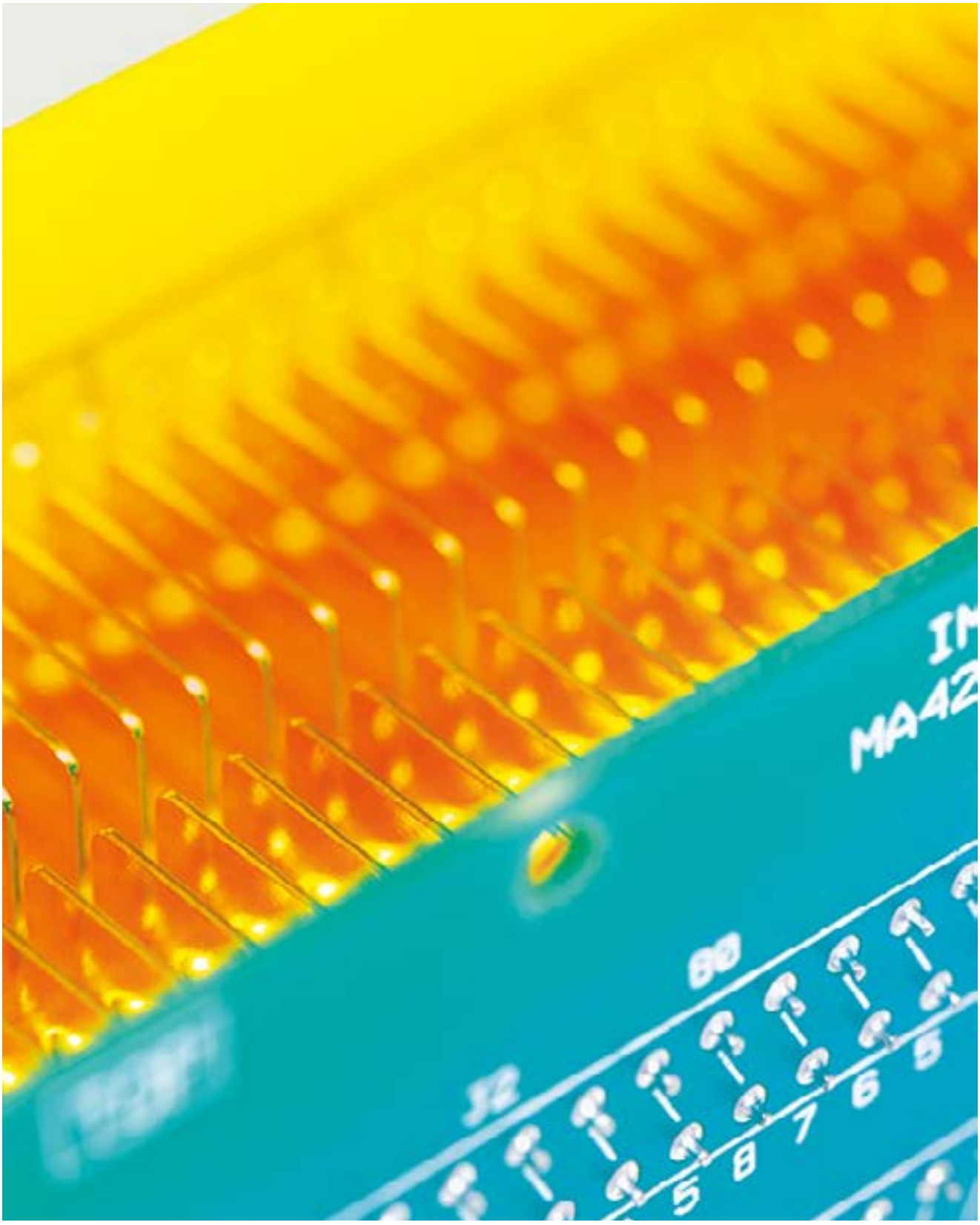
18,097 proteins



Picture credit: ediundsepp



Ionizing peptide samples for mass spectrometry. Guided via a thin glass capillary (left), the samples encounter high electric voltage and are transferred into an aerosol (electrospray ionization). The resulting ions enter the metal capillary (right) leading into the vacuum of the mass spectrometer.



Picture credits: Filser

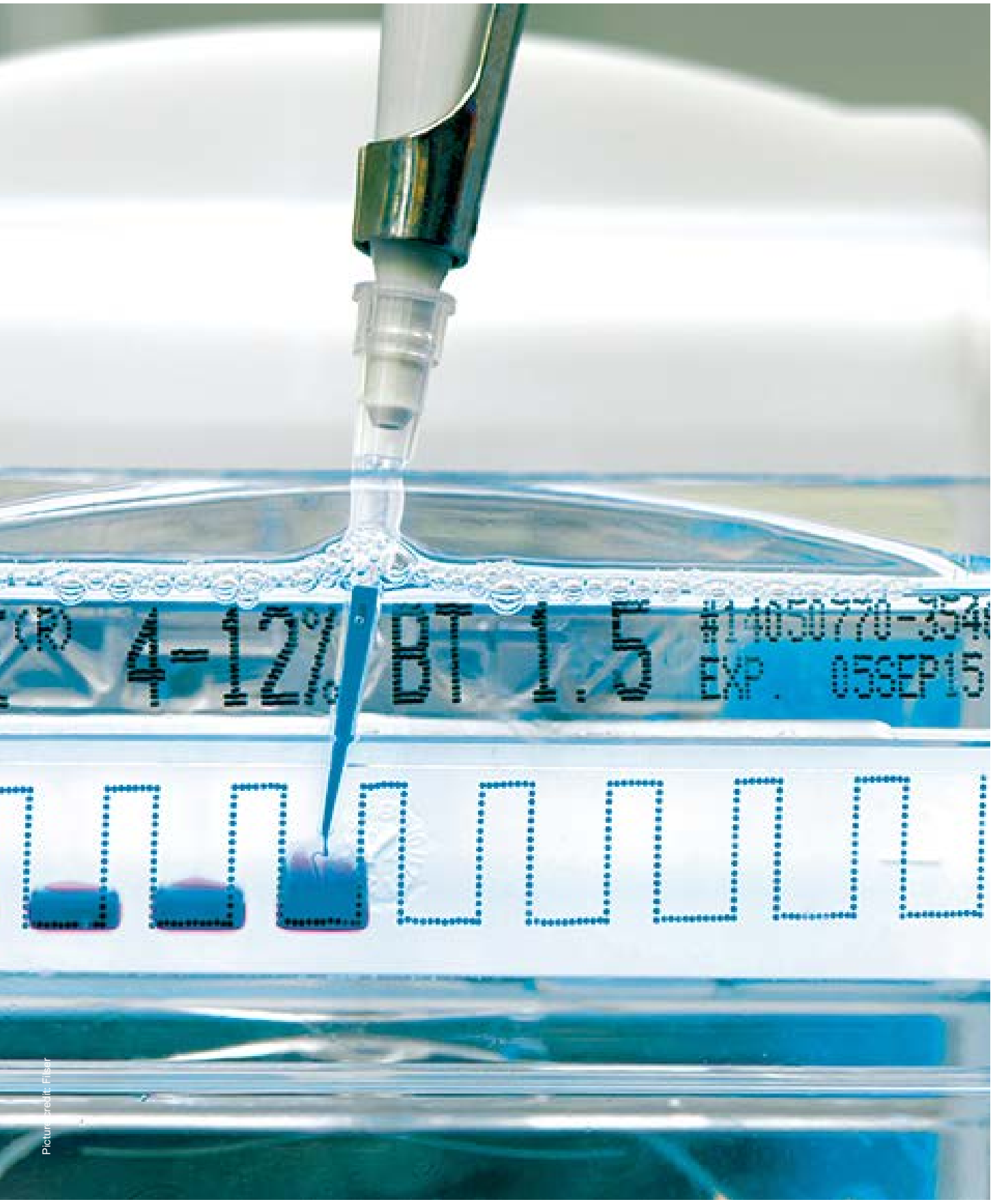
Inside the mass spectrometer, the ions are accelerated. The resulting ion beam enters a magnetic field, which deflects the individual ions according to their mass and charge. The above image shows a so-called stacked electrode, which is part of the optics that focus the ion beam.

Complex projects call for dogged determination – and Bernhard Küster, TUM Professor of Proteomics and Bioanalytics, has what it takes; his work is more complex than most people can imagine. At the end of May, the results of his research were announced to the world. According to Küster's article published in the scientific journal "Nature," the human organism is made up of at least 18,097 different proteins. A study published at the same time by a team led by Akhilesh Pandey from Johns Hopkins University in the US arrived at the slightly lower figure of 17,294 proteins. The two teams have produced by far the most comprehensive map of the human proteome to date. Bernhard Küster's group maintains that they have so far accounted for 92 percent of all proteins in the human body. "Nature" dedicated its front cover to the finding, confirming the human proteome as one of the great scientific discoveries of the year.

The human genome codes for the proteome

The term proteome, coined in 1994 by Australian Marc Wilkins, describes the entire set of proteins in an organism and is based on a more well-known word ending in "ome": the genome, or the genetic blueprint of an organism. As far back as the 1990s, the sequencing instruments of genome researchers were already running at full speed. By 2003, the DNA sequence of humans was more or less correctly decoded. Geneticists soon announced that they had also solved the human protein puzzle. In simple terms, they calculated that each gene encodes a protein, so the number of genes corresponds to the number of proteins. Based on this interpretation, the human proteome would comprise 19,629 proteins, all stored and named in freely accessible databases. Many, however, are just predictions derived from computer analysis of the human genome. Now, 15 years after the decoding of the genome, 2014 could well become known as the year the human proteome was decoded. Scientists have mapped the proteome and, as a result, arrived at a more solid and accurate prediction for the number of human proteins. ▷

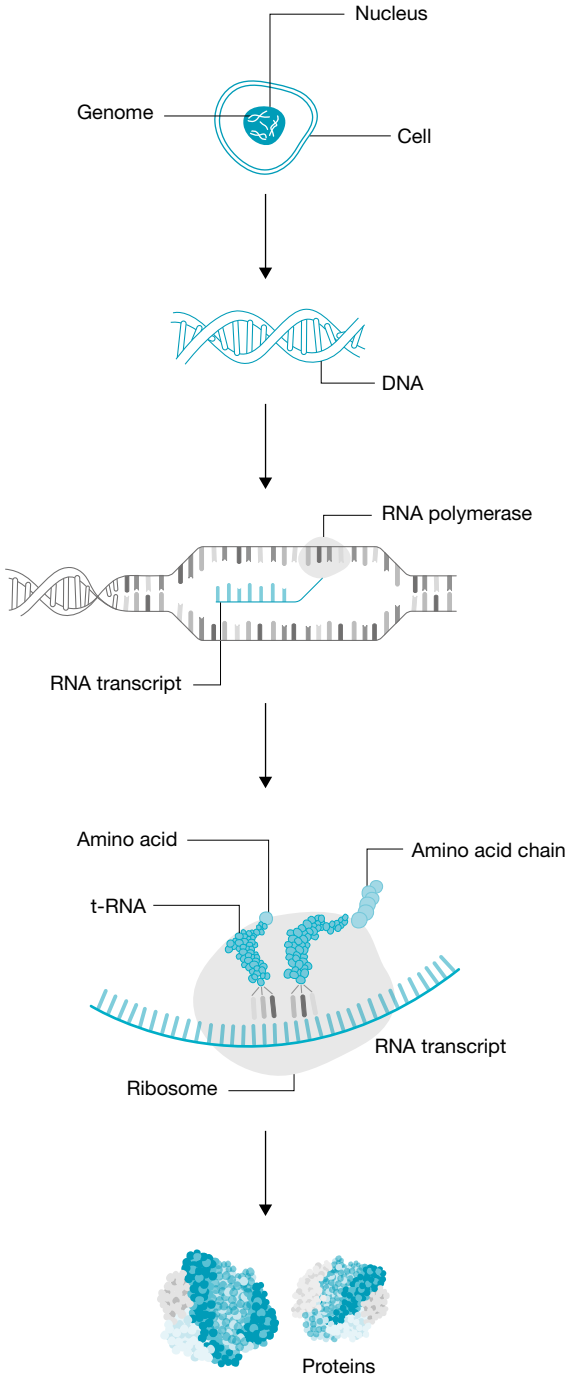
Bernhard Küster's group collected protein samples from 60 different tissues, 13 bodily fluids and 147 cancer cell lines. Here, a gel electrophoresis apparatus is being loaded with protein samples to separate them according to size.



Picture credit: Fiser

Genes and Proteins

Our genes contain the blueprint for the production of proteins. The process starts with the unpacking of the DNA of the chromosomes and its transcription into messenger RNA. At the level of the ribosomes – the protein factories of the cell – many individual amino acids are linked together to form a protein following the blueprint of the messenger RNA, but with all kinds of chemical modifications, if required.



The catch is that, unlike the genome, which barely changes in the course of a person's life, the proteome is highly dynamic. Proteins are constantly generated, transformed and broken down, depending on the organism's exposure to stimuli, environmental factors, diseases and drugs. The proteome is highly complex, as it reflects all facets of our life and our environment.

Back in 2001, an international research project called the Human Proteome Organization (HUPO) was set up along the same lines as the genome research project. The ultimate target is to analyze every protein in every tissue – including their changes over time and variants of the basic form. This undertaking would mean analyzing 500,000, possibly even a million proteins. "That is a lot of proteins," noted the journal "Nature," skeptical of the hugely ambitious plan in 2001. The researchers are still a long way off their ultimate target, but now two research groups have at least reached the first milestone. The basic number of human proteins has been determined, the variants of which will be identified in further research.

Mapping the human proteome

Progress in mapping the human proteome was aided by advances in mass spectrometry (MS). This method is as important to proteome researchers as sequencing instruments are for genome researchers. Küster has five mass spectrometers available at his chair, each one worth some EUR 750,000.

Over the last 18 months, these machines were tasked with drawing a map of the human proteome. Bernhard Küster's team analyzed 60 different tissues, 13 bodily fluids and 147 cancer cell lines. Most of the tissue samples were provided by TUM's pathology labs, and as Küster confirms, laughing: a number of saliva and even ear wax samples were acquired from his own people for analysis. After that, the researchers followed a straightforward work flow. First of all, the proteins from any biological source were broken down using a protease, an enzyme that cuts all proteins into pieces called peptides at precisely defined positions within the amino acid sequences. After a few hours of lab work, the reaction tube contains an impressive 100,000 peptides instead of the original approximate number of 19,000 proteins.

The next step is pre-separation. "Even the most modern mass spectrometers are not able to handle the complexity of 100,000 peptides all at once," stresses Küster. So a liquid chromatography technique is used to send around 50 different peptides every second to the measurement chamber of the directly connected mass spectrometer. "Modern MS devices can properly identify five to ten peptides per second in a repetitive process along the liquid chromatographic time frame of several hours," he explains. A mass spectrometer needs ions – it cannot measure neutral particles. To generate ions from peptides,

Picture credit left: ediundsepp, right: Filser

the scientists use electrospray ionization, a technique for which John B. Fenn won the Nobel Prize for Chemistry in 2002. Bernhard Küster explains, "That was a very important development because it allows us to ionize even large proteins and peptides without destroying them." Subsequently, and through controlled manipulation of the trajectory of the ions in electrical fields under high vacuum, the ionized peptides are separated in the mass spectrometer according to their ratio of mass to charge. At the end, the detector records a huge spectrum of signals from which the mass and quantity of each peptide can be determined. In a second analysis step, the device determines their amino acid sequence, and within one hour, the instrument can churn out data for over 20,000 peptides.

But what can be done with this huge volume of data? What goes together? Which peptides belong to which protein? The next step is to complete a giant jigsaw puzzle aided by information from the genome. "Our work would be much more difficult without the decoded genome. We need the amino acid sequences that can be derived from genes to piece together our peptide analysis data efficiently," explains Küster. For this, a computer must first virtually reproduce the protease breakdown of all of the predicted 19,629 proteins and, if needed, of other sequences present in the genome. Next, a mathematical algorithm compares the amino acid sequences determined in the measurements with those of the theoretically calculated peptides. One by one, the machine checks which puzzle pieces it can and cannot combine and assemble into proteins.

The biggest coherent data set of the human proteome

In this way, the Munich-based scientists identified some 80 percent of all human proteins. This work thus comprises the largest single data set of the human proteome. In addition, the team re-analyzed several dozen individual MS data sets on tissues and cell lines that other groups had uploaded to public databases. "The main starting point for our study was the idea that we could use the world's existing knowledge by bringing it all together," emphasizes Küster. Examples of such internationally available databases include PeptideAtlas and ProteomeXchange. The problem is the lack of software and computer architectures capable of performing comparative analyses of the terabyte-scale data volumes. Hence, some databases are currently little more than data dumps. Just over two years ago, therefore, Küster hit on the idea of collaborating with SAP to develop a new database and software. The result is ProteomicsDB – an all-in-one database and software system running on computers at SAP's headquarters in Walldorf to perform a comprehensive analysis of the human proteome. Küster sees the platform as a tool for all scientists engaged in human proteome research: "It's available to anyone." New data sets are continuously being added to ProteomicsDB – the Pandey group's sets have already been loaded – and in mid-September 2014, updated results were made public. At that stage, 18,248 of the 19,629 predicted human proteins were actually available, or 93 percent of the number forecast by the genome researchers. ▷

"If we knew the protein profile of, for instance, a tumor, we might be able to administer drugs in a more targeted way."

Bernhard Küster

18,097

human proteins were published by Bernhard Küster in May 2014. New data sets are continuously being added, and in mid-September, updated results were made public. At that stage, 18,248 of the 19,629 predicted human proteins were actually available.

Using proteins to develop cancer therapies: Küster and his team were able to show that the efficiacy of 24 cancer drugs on 35 cancer cells bore a clear correlation with their protein profiles. The photo shows cancer cells being removed from the liquid nitrogen tank in which they are stored.

Missing proteins

Some details in the huge trove of data confirm previous expectations. For instance, a core set of 10,000 to 12,000 proteins can be identified in most cell types and tissues. In addition, many tissues are characterized by the presence of specific proteins.

The proteome map also raises some fundamental questions for scientists, such as how to find the seven percent or so of all predicted proteins that have remained elusive until now. This is undoubtedly due, in part, to the detection limits of mass spectrometry, as proteins with just 100 molecules per cell or even fewer remain under the radar. Detecting highly insoluble proteins is also proving to be a challenge. To identify this last seven percent, the TUM scientists are appealing to the world's experts in the field for help – one of the sections in ProteomicsDB is called "Adopt a protein". "There are very good labs out there doing in-depth research on a small region of the map, and we invite them to add their data to ProteomicsDB," explains Küster.

It may turn out that some of the currently gray areas in the proteome map don't even exist. A number of genes are permanently inactive, having been switched off as humans evolved. One such inactive area can be found in the human nose. In order to smell, organisms from mice to

men need special receptors on the surface of sensory cells in the olfactory epithelium located far back in the nose. Based on the predictions, we should have 853 of these olfactory receptors. But Bernhard Küster and his group have been able to account for only one in four of these. Perhaps this is one reason why our sense of smell is much less well developed than that of, say, mice and dogs.

Previously unknown proteins

Another surprising finding is that the term "gene" might have to be interpreted more broadly than it has been to date. It is clear that our bodies create proteins whose genes have not yet been recognized by the genome researchers.

Our cells are home to some 10,000 large RNA molecules with functions that remain unexplained. Some are referred to as transcripts of unknown coding potential (TUCP). Others are called long intergenic non-coding RNAs (lincRNA). Some play a role in the regulation of stem cell properties and cell differentiation, for example. The attribute "non-coding" is not quite accurate, however, because the TUM team has identified 430 proteins originating from such non-coding lincRNA sequences. They were able to assign these by comparing the proteins found with amino acid sequences that can be derived from the lincRNAs. At present, no one knows why these "genes" are translated into proteins, but they may well represent evolutionary processes in which nature "tries out" new proteins with perhaps novel properties.

Protein patterns can determine drug efficacy

"Our map is only the starting point," says Küster. Scientists are sure to find more surprises while filling in the remaining spaces on the human proteome map and asking questions of the database. But what is the point of proteome research, apart from gaining a better understanding of basic biological principles? According to Küster, there are already indications of possible medical implications, as proteins are the targets for almost all medicines. It seems that certain protein patterns influence the effectiveness of such drugs. Küster and his team were able to show that the efficacy of 24 cancer drugs on 35 cancer cell lines bore a clear correlation with their protein profiles.

Says Küster, "This edges us a bit closer to even more individualized treatments for patients. If we knew the protein profile of, for instance, a tumor, we might be able to administer drugs in a more targeted way. This would also create a rationale for investigating new drug combinations and, generally, aligning treatments more closely with a patient's individual needs." *Bernhard Epping / Barbara Wankerl (TUM)*

Picture credit: Filser