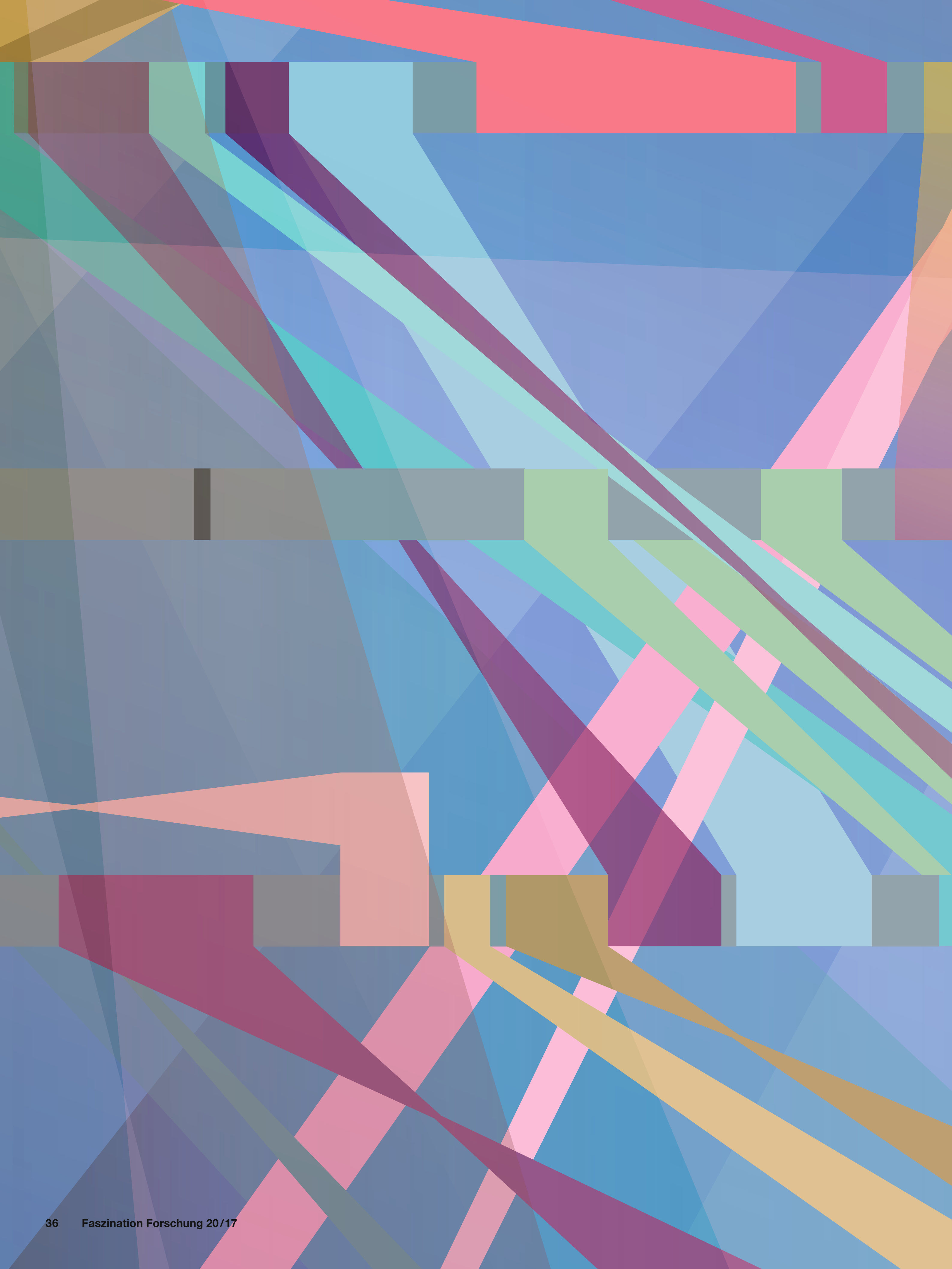
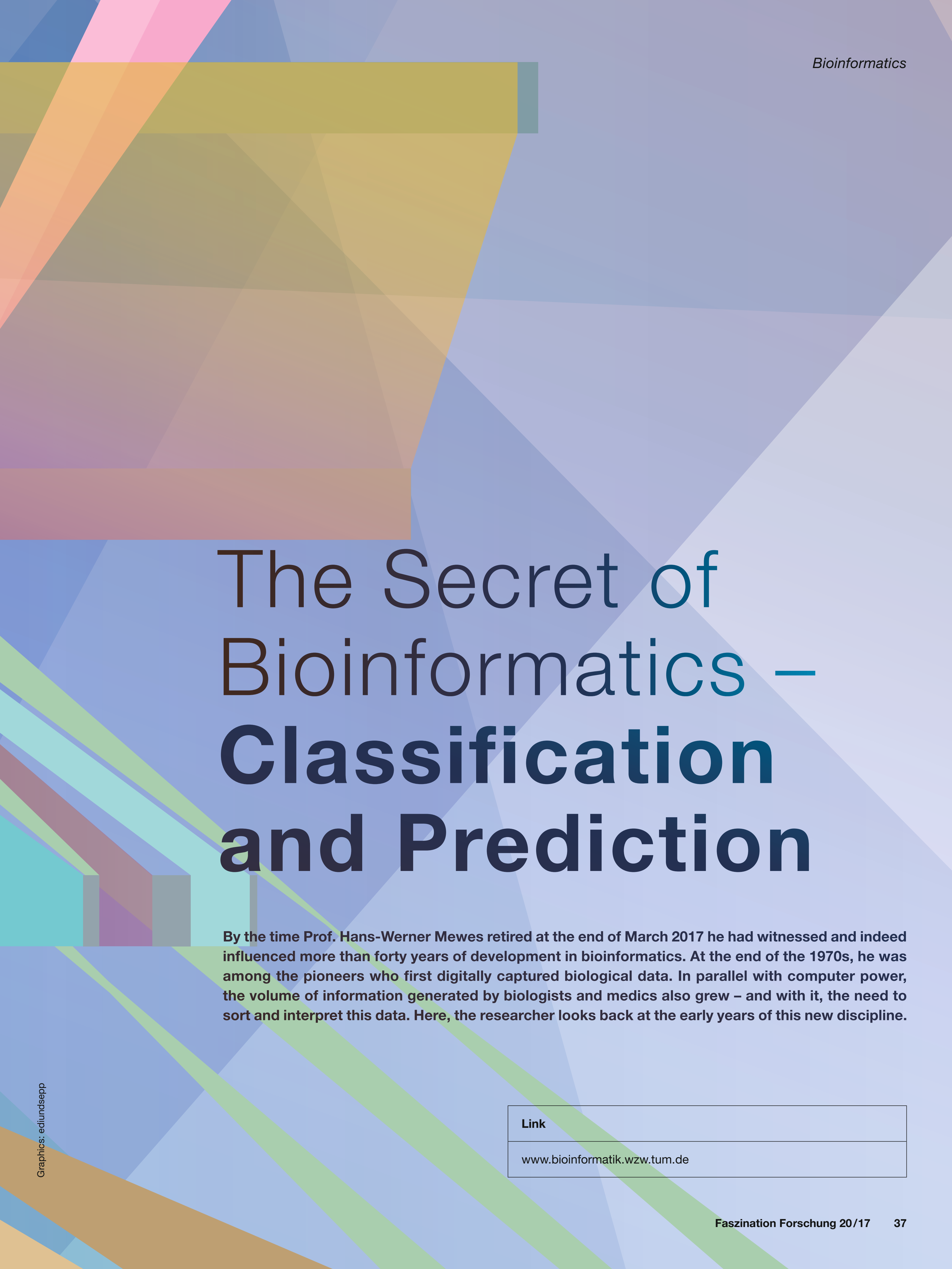




Graphics: ediundsepp

The Secret of Bioinformatics – **Classification and Prediction**

By the time Prof. Hans-Werner Mewes retired at the end of March 2017 he had witnessed and indeed influenced more than forty years of development in bioinformatics. At the end of the 1970s, he was among the pioneers who first digitally captured biological data. In parallel with computer power, the volume of information generated by biologists and medics also grew – and with it, the need to sort and interpret this data. Here, the researcher looks back at the early years of this new discipline.

Link

www.bioinformatik.wzw.tum.de

Brigitte Röthlein

„Die Kunst der Bioinformatik besteht aus zwei Dingen: der Klassifikation und der Vorhersage“

In den Lebenswissenschaften ist heute Forschung ohne Datenverarbeitung nicht mehr möglich. Durch die immer schnellere und preisgünstigere Sequenzierung von Genen bzw. Proteinen gab es in den letzten Jahren eine Explosion von Daten, deren positive Folgen im Erkenntnisgewinn in der Praxis und Klinik erst jetzt allmählich sichtbar werden. Um sie auszuschöpfen, benötigt man Techniken der Künstlichen Intelligenz, etwa maschinelles Lernen oder künstliche neuronale Netze. So ist ein neues Fachgebiet entstanden, die Bioinformatik.

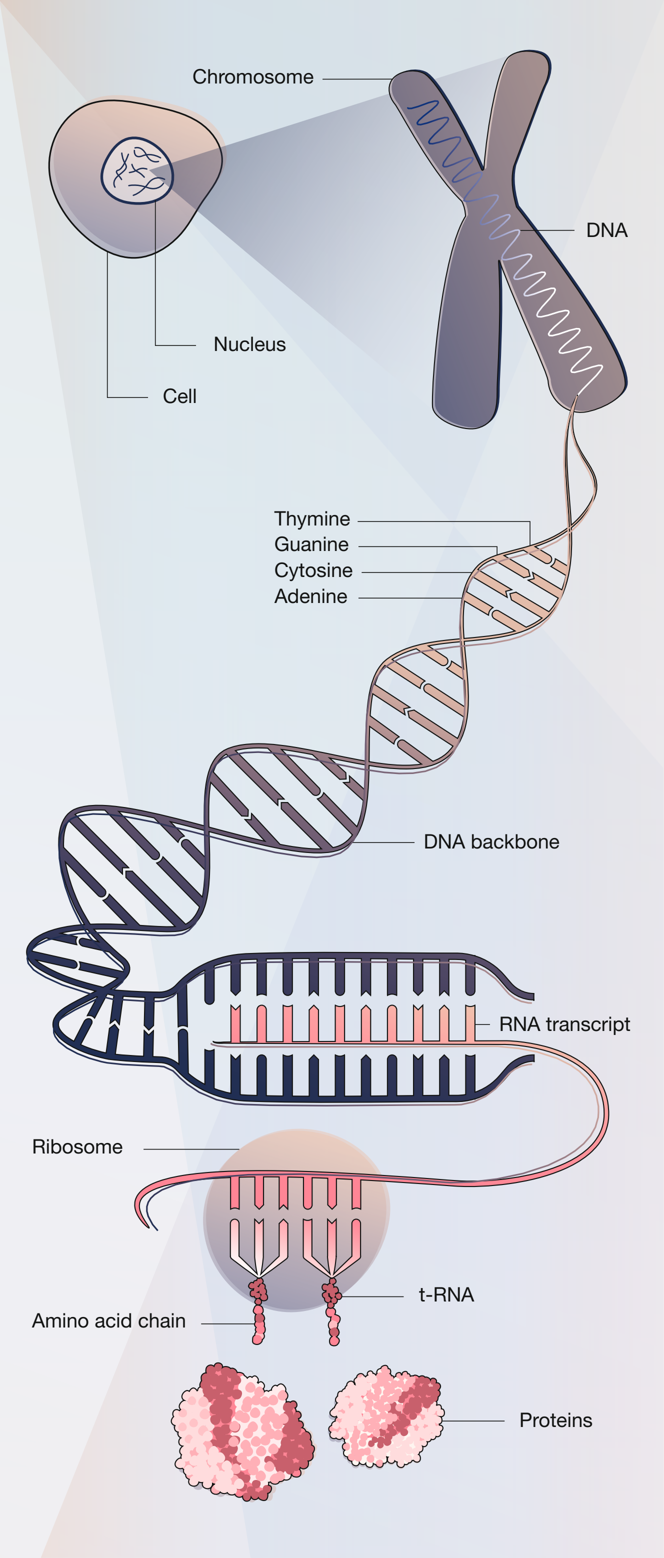
Prof. Hans-Werner Mewes, der zu den Pionieren in dieser Disziplin gehört, setzt sich seit Jahren mit der Umwandlung von biologischem Wissen in berechenbares auseinander. Er leitete bereits 1989 beim Hefegenomprojekt, einem der ersten Projekte zur Entschlüsselung des kompletten Genoms eines Organismus, das Team, das die bioinformatische Aufbereitung der Daten von den insgesamt rund 600 beteiligten Wissenschaftlern aus über 100 Laboren übernahm. Zuerst mussten dort die analysierten Fragmente anhand überlappender Abschnitte korrekt zu einem durchgehenden DNA-Strang zusammengefügt werden. Danach stellten die Forscher die Sequenzdaten und die zugehörigen Informationen systematisch in computerlesbarer Form dar und organisierten sie als Datenbank. Sie entwickelten damals zur Darstellung des Hefegenoms eine Oberfläche, die eine einfache symbolische Visualisierung der Chromosomen erlaubte. So wurde es durch vielfältige Verknüpfungen von Datenelementen möglich, mit Browsern durch das Erbgut zu navigieren.

Die bioinformatischen Verfahren wurden im Laufe der Jahre immer ausgefeilter und spezifischer auf die Probleme zugeschnitten. Man erstellte Funktionskataloge, das heißt, man versuchte zu klassifizieren, welche Aufgaben welcher DNA-Abschnitt jeweils übernimmt. Denn die Kunst der Bioinformatik besteht aus zwei Dingen: der Klassifikation und der Vorhersage. Mit der zuverlässigen Vorhersage von biologischen Eigenschaften aus Daten werden viele Experimente überflüssig, andere können so erst interpretiert werden.

Um die Übertragung akademischer Lösungen in die industrielle Praxis voranzubringen, gründete Mewes zusammen mit Kollegen 1997 die Firma Biomax, die solche Lösungen anbietet. Seit 2002 gibt es in München das Studienfach Bioinformatik in einer Zusammenarbeit zwischen TUM und LMU. Seither wurden schon rund 600 Bioinformatiker ausgebildet, viele haben promoviert, einige sind Professor(innen) geworden, viele arbeiten im Ausland und in der Industrie. □

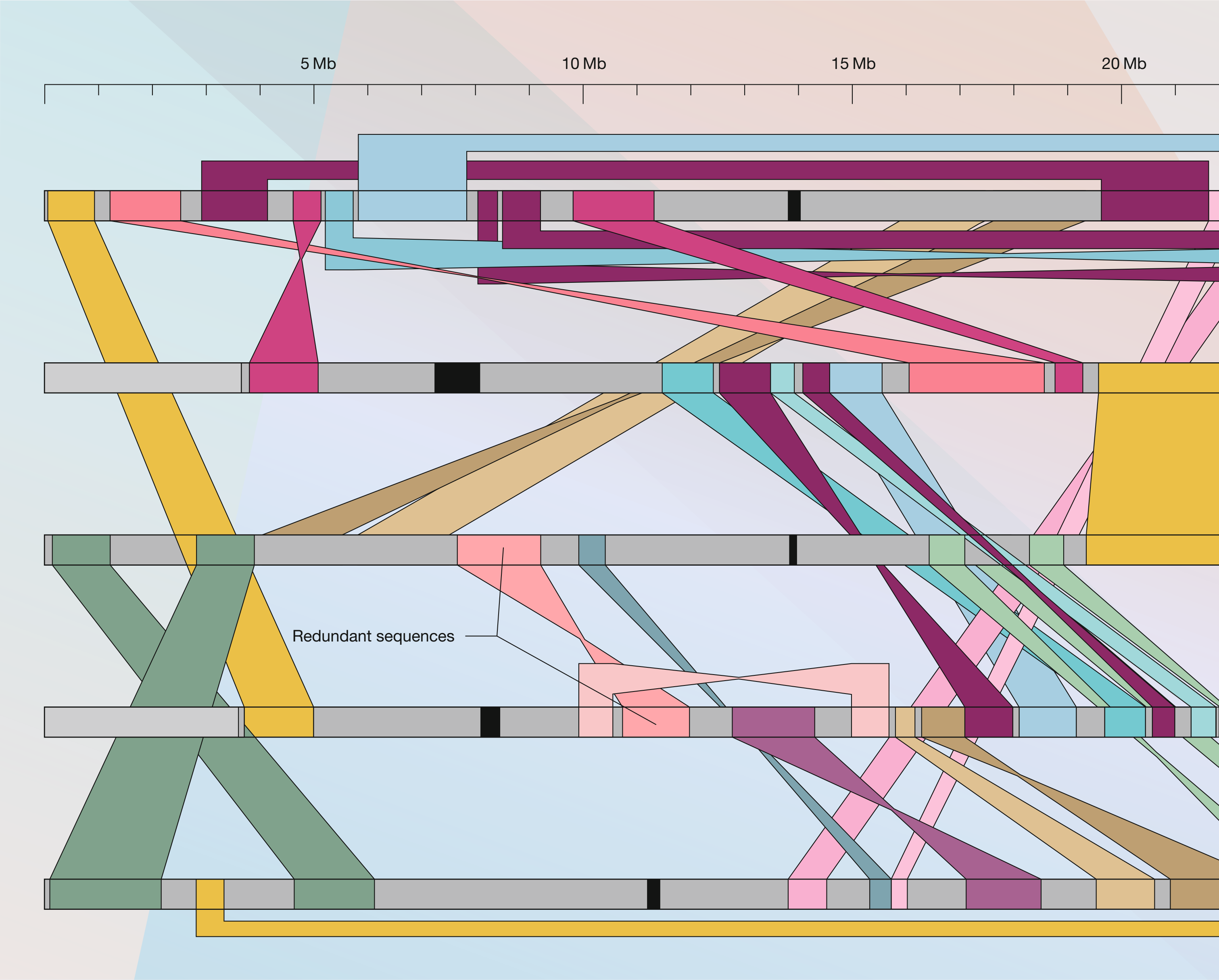
Prof. Mewes, you were among the very first bioinformaticians. Could you tell us how it all began?
Well, the field of bioinformatics is really still in its infancy. It is directly linked to the massive data streams produced by biology today – our role being to sort, merge and interpret it. Personally, I began working in this area in 1979 at Heidelberg University. There we had what was – at the time – a very expensive computer with 10 kilobytes of storage capacity, and I programmed that to digitally record measurement data. That was a very exciting step. Fortunately, we were dealing with very modest volumes of data back then.

But then all of that changed?
Yes, very quickly indeed. The core challenge facing bioinformatics today is the sheer volume and complexity of available information. Over the last thirty years, sequencing technologies have become more and more advanced, and this has created huge streams of data. The first step here is to determine the order of the four bases that form the building blocks of life. We represent them as strings consisting of four letters: G, A, T and C, designating the bases guanine, adenine, thymine and cytosine.
The process began in the 1950s with protein sequencing – then a matter of a few hundred amino acids – and continued in the seventies and eighties with the sequencing of DNA in bits and pieces, not to assemble whole chromosomes. By then, we were already dealing with millions of base pairs. We then selected certain model organisms, already extensively studied from a biological perspective – like yeast, for instance, whose 12 million base pairs encoding 5,800 genes were identified by 1996. Finally, at the end of June 2000, the first full sequence of a human genome was published, consisting of over 3 billion base pairs coding for around 20,000 genes. Today, huge amounts of data are generated by other methods too, such as proteome analysis. Fortunately, at the same time, computer power has also grown at an explosive rate – and every lab can now access the relevant databases over the Internet. ▷



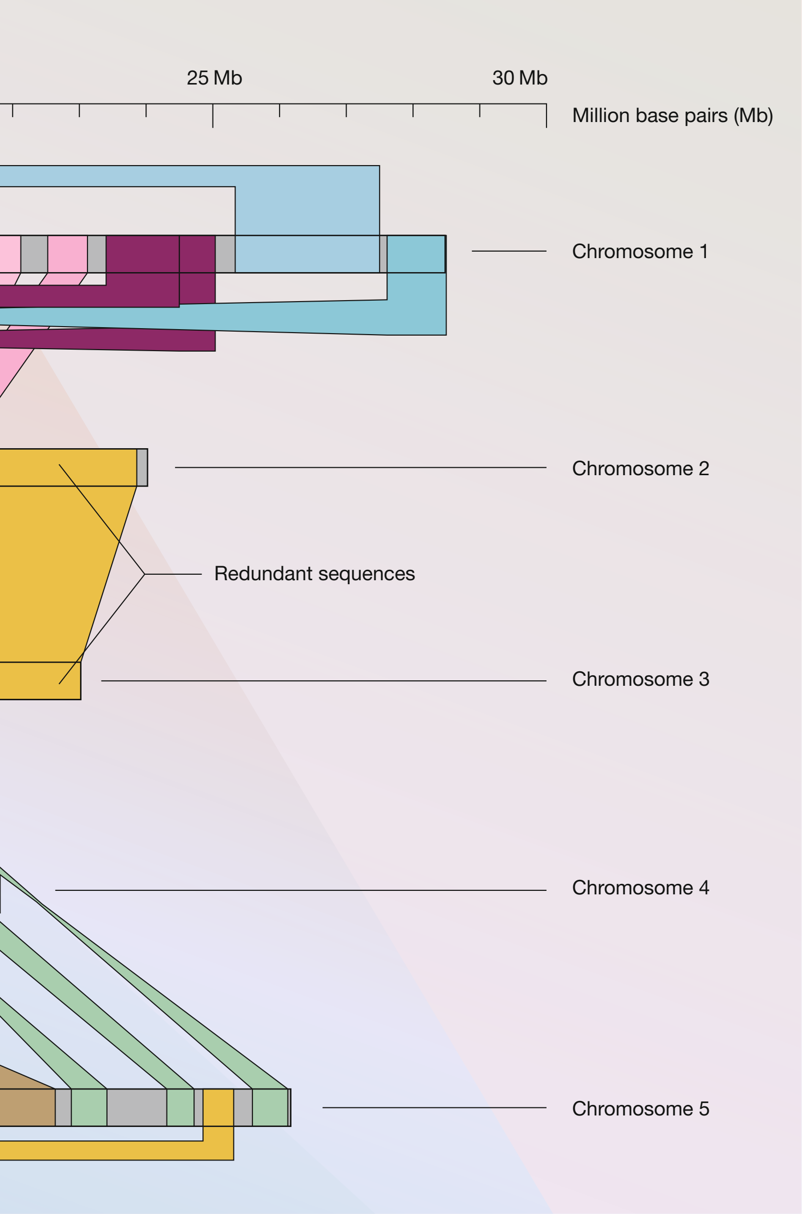
The amino acid sequence stored in the genes is the blueprint for every creature's proteins. To start protein production, DNA is unpacked from the chromosomal proteins and transcribed into messenger RNA. Ribosomes – the protein factories of the cell – link the individual amino acids together to form a protein following the blueprint of the messenger RNA.

Graphics: ediundsepp (source: TUM), Picture credit: Jooss



What did these developments mean for you personally? Well, in the 1980s, I was engaged in collecting protein sequences and annotating them with information about the origin, properties and content of the data. In the subsequent yeast genome project, which started in 1989, I led the team responsible for bioinformatics processing of data obtained from all the participating scientists – roughly 600 – across over 100 laboratories. First, we had to correctly assemble the analyzed fragments into a continuous DNA strand by means of overlapping sections. Then, we converted the sequence data and associated information systematically into computer-readable form and organized these findings in a database. Our 1997 paper on this was published in "Nature" and was widely acclaimed. To illustrate the yeast genome, we also developed an interface allowing simple symbolic visualization of the chromosomes. Numerous links between data elements thus enabled users to navigate the genome using a browser. The catalog grouping the yeast genes by function played an important role here and is still in use today.

But you didn't stop at compiling databases, did you? No. Once we had our catalog of 5,800 yeast proteins, I was convinced that we couldn't just publish it like that – it wouldn't make sense. So we started to produce a catalog of functions – in other words, we attempted to classify the tasks performed by each section of DNA. The secret of bioinformatics lies in two things – classification and prediction. Reliable prediction of biological properties based on data renders many experiments superfluous, while others can only be interpreted in this way.



“The core challenge facing bioinformatics today is the sheer volume and complexity of available information.”

Hans-Werner Mewes

Three years after sequencing the yeast genome, the genome sequence of *Arabidopsis thaliana* was unraveled. The genome of *A. thaliana* comprises 125 million base pairs stored in five chromosomes, shown here as gray bars. The sequence analysis revealed that several parts of the genome are redundant – the illustration shows these sequences as colored bands. This indicates that the genome of *A. thaliana* was duplicated several times over the course of its evolution. As members of the Arabidopsis Genome Initiative, Mewes and his colleagues were involved in assembling the entire genome and in localizing those genes which actually code for proteins.

Graphics: ediundsepp (source: AGI, Nature Vol. 408, December 2000)

And how do you go about doing that?
Methods for genome analysis have been developed to make the best use of the collected data. An entire repertoire of algorithms to search for patterns and similarities, translate DNA into hypothetical proteins and link genetic elements with knowledge gained through experiments can be applied to the bare sequences.

Did that not require a huge amount of computing horsepower?
It certainly did. To give an impression of the sheer scale of searches for similar patterns: you are comparing each sequence with all of the others. So the more there are, the more challenging that becomes – it takes an enormous amount of computing time. I think entire power plants were probably running just to work the computers for the Human Genome Project. We had two tricks up our sleeve here: The first was to use a high-speed format to store the data. And the second was to build a network enabling large numbers of Internet users to contribute to these endeavors. This type of crowdsourcing means that anyone who has the time and interest can help with specific digital tasks. The project was implemented by Thomas Rattei at TUM’s Department of Bioinformatics – he is now a professor at the University of Vienna. And there are still around 10,000 private computers involved in this program today, as everything is always being updated. Now, all you need to do is press a button to ask: “Where can we find similar sequences?” ▷

Why is it so important to search for similar gene sequences?
This is the essential groundwork we must cover before we can begin detailed analysis of the sequences to establish the function of individual genes. In this way, we try to predict which functions a protein might have or whether a mutation could trigger a particular rare disease. Research efforts investigating who has a tendency toward obesity and who does not are another example. Gene regulation plays a major role here, which is why it is so difficult to link genetic predisposition to disease.

What methods do you use for this?
Along with many other groups, we use machine learning techniques to search for patterns in large and complex datasets – such as DNA information and disease progression data – in order to identify patient groups responding to a specific treatment, for instance. While neural networks learn from our brain patterns, machine learning involves the computer autonomously developing a model that simulates the data as accurately as possible. This is an iterative process using trial and error, based on a number of known examples. If the outcome is sufficiently reliable, this model can then be used to evaluate therapeutic decisions and individualize treatment.

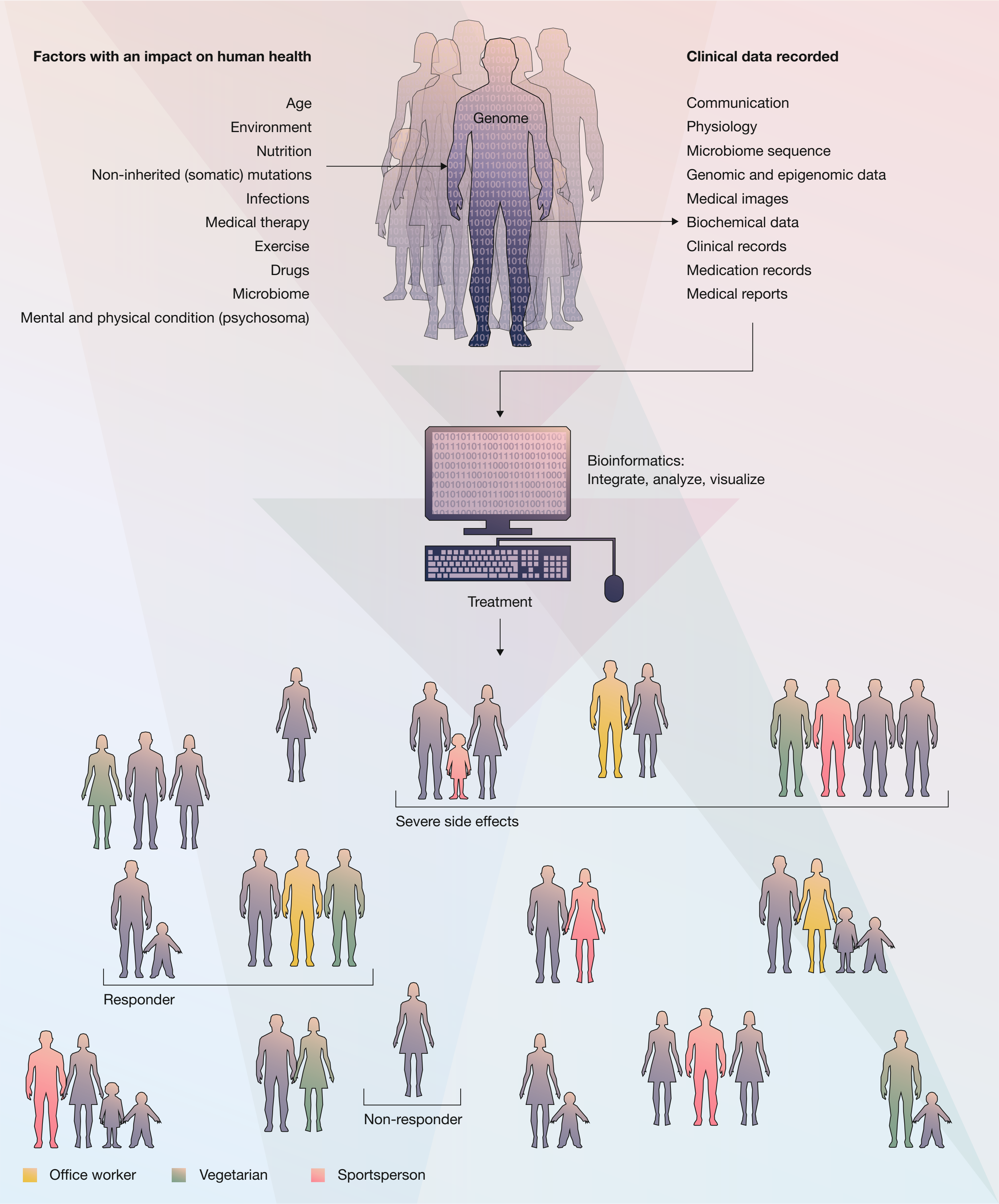
What other methods did you apply?
In another major step, we set out to determine how proteins interact with one another – which of course feeds into network biology. To do that, we began compiling a catalog of protein-protein interactions a long time ago.
After the turn of the millennium, the idea of systems biology then emerged. This approach aims not only to explore the interactions, but also to establish full computational models to replicate the actual processes. These could then be used to essentially simulate experiments on the computer. However, systems biology struggled with the complexity of the processes; and in practice it was only possible to develop relatively small models, which did not allow the bigger puzzle to be reassembled. As a result, this approach lost some of its appeal.

So how, specifically, does bioinformatics contribute to research?
Bioinformatics is a highly interdisciplinary field. We are not the only ones responding to computational problems. Obviously we do have to master computational techniques, but what the bioinformatician sets out to understand is: What is the biological research question? Which method do I apply? What data does this then generate? How can I convert that data? And how do I then interpret it in the context of the biological question being researched?

Are there already tools for this on the market?
Yes, several, but there is a problem: In the academic space, research focuses specifically on the data needed to answer the next research question, on an ad hoc basis. So you have a team of people who can handle the data and find answers to biological questions, and the end result is a publication. The effort involved is huge, but there isn't really a professional software package or solution to support this. So the market doesn't actually have standard tools that can be universally used for these projects. As it stands, if you want to use bioinformatics in industry, you have to hire specialists, but you will end up with an isolated solution that is not easy to sustain.

In other words, there is no real technology transfer?
It is still very limited, since academic solutions cannot be directly commercialized. That is why, together with colleagues, I founded the Biomax company in 1997, to offer such solutions. But the market beyond large institutions remains tough. Even in the clinical sector, computing power is primarily used for accounting and not set up to use diagnostic data to find the best possible therapy.

Where are the shortcomings, in your view, and what improvements would you like to see?
I would like the data to be used by doctors and in the clinical environment, and I want its application to benefit patients. Patient care produces enormous amounts of information. ▷



How patients react to medication depends on a large number of variables, for instance genome, overall health, lifestyle and environment. Bioinformatics develops tools to gather, analyze and understand complex medical data with the aim of predicting whether a patient will or will not respond to a certain therapy or whether he or she will experience severe side effects from it.

Graphics: ediundsepp (source: TUM), Picture credit: Jooss

It is estimated that a medium-sized hospital with 100 to 200 beds generates 600 terabytes of data per year – and the use of advanced diagnostics keeps adding to that. But this data mine is currently untapped. Yet it could be a source of important findings, allowing closer analysis of the factors involved in complex diseases such as diabetes or psychiatric disorders. Based on the success of treatment, it should then be possible to identify specific patient groups with different reactions to medication: responders, non-responders, poor metabolizers and super-responders. This knowledge could be used to optimize treatment and save further costs.

And today it is possible to study bioinformatics as a degree course, isn't it?

Yes, the German Research Foundation (DFG) requested proposals for bioinformatics courses in 1999, and I would never have dreamed that I would one day become a bioinformatics professor. But the subject has been very well received and our graduates are snapped up. The course has been available in Munich since 2001, and we have already trained around 600 bioinformaticians. Many of them have also pursued doctorates and some have gone on to become professors. A lot of them work abroad and in industry, with highly interdisciplinary careers. *Interview conducted by Brigitte Röthlein*

Prof. Hans-Werner Mewes

A pioneer of bioinformatics

As a bioinformatician, Hans-Werner Mewes is an interdisciplinary researcher, combining his grounding in chemistry with expertise in biology, medicine and computer science. Following his school-leaving exams (German Abitur) in 1969 in Marburg/Lahn, he studied chemistry at the University of Marburg, receiving his degree in 1978. He then took up roles first at the University of Heidelberg, then at the European Molecular Biology Laboratory (EMBL) in Heidelberg in 1983. Two years later, he moved to the Max Planck Institute (MPI) of Biochemistry in Martinsried, at the same time studying for his doctorate at the University of Marburg. On obtaining this, he became Director of the MIPS (Munich Information Centre for Protein Sequences at the MPI of Biochemistry), and was then made an honorary professor in the Faculty of Biology at LMU Munich in 1999.
Since 2001, Hans-Werner Mewes has been full professor of genome-oriented bioinformatics at TUM, based at the School of Life Sciences Weihenstephan. At the same time, he took over as Director of the Institute of Bioinformatics and Systems Biology at the German Research Center for Environmental Health. Since 2011, he has also been a faculty member of the TUM School of Medicine. From 2010 through 2014, Mewes was an active member of the Helmholtz Association Think Tank and, most recently, also served as spokesperson for the Helmholtz Graduate School of Environmental Health. Hans-Werner Mewes became professor emeritus on March 31, 2017. He is the founder of two companies: Biomax Informatics AG (with Klaus Heumann and Dmitrij Frishman) and Clueda AG (with Volker Stümpflen and Daniel Pinnow).

> *"What the bioinformatician sets out to understand is: Which experimental method was applied and what kind of data does it generate? How can I analyze that data to find signals? And how do I then interpret it in the context of the biological question being researched?"* Hans-Werner Mewes

Picture credit: Jooss