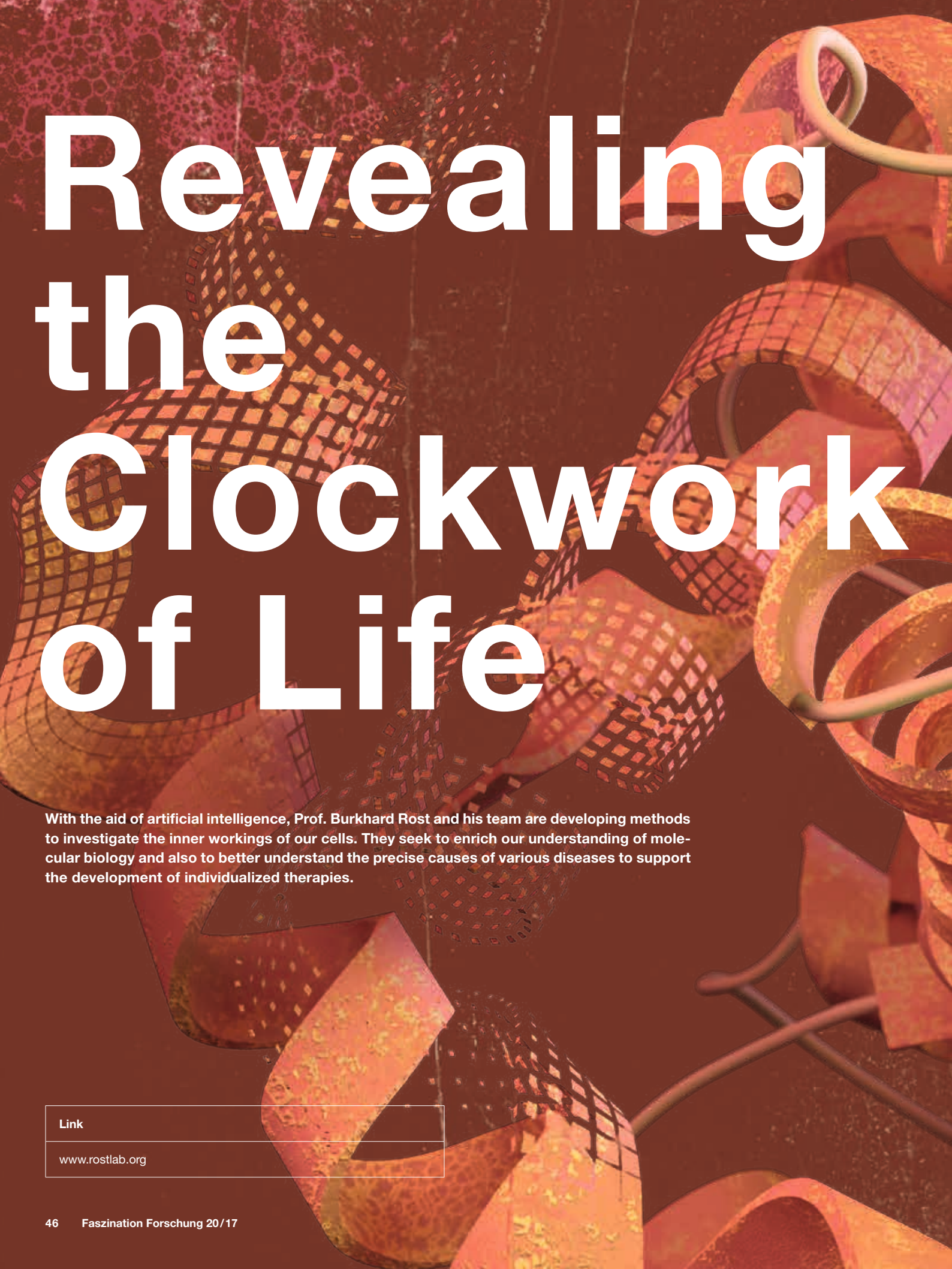
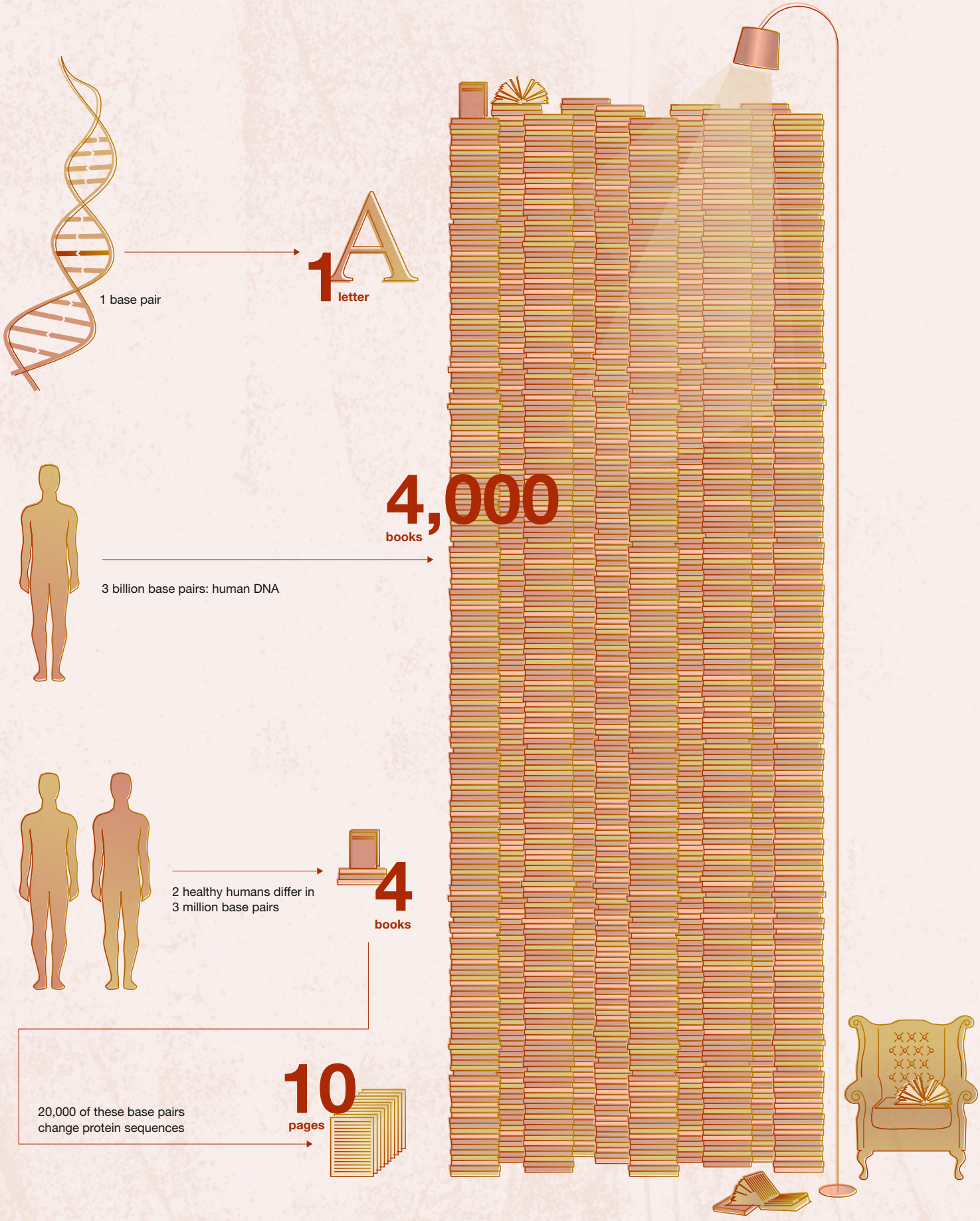
Revealing the Clockwork of Life



With the aid of artificial intelligence, Prof. Burkhard Rost and his team are developing methods to investigate the inner workings of our cells. They seek to enrich our understanding of molecular biology and also to better understand the precise causes of various diseases to support the development of individualized therapies.

Link
www.rostlab.org

Graphics: ediundsepp (source: RCSB PDB)


1 base pair
1 letter
4,000 books
3 billion base pairs: human DNA
2 healthy humans differ in 3 million base pairs
4 books
20,000 of these base pairs change protein sequences
10 pages

Brigitte Röthlein

Diagnosesysteme für die medizinische Werkstatt

Prof. Burkhard Rost und sein Team arbeiten daran, wichtige Aspekte der Proteinfunktionen aus den Informationen vorherzusagen, die im Genom stecken. Sie benutzen dazu einerseits die Daten über Proteine, die sich im Laufe der Evolution unterschiedlich entwickelt haben, andererseits genetische Varianzen, die sich zwischen gleichartigen Lebewesen zeigen. Auf diese Daten wenden sie neben der Statistik Methoden der Künstlichen Intelligenz an, insbesondere des Maschinenlernens, der neuronalen Netze und andere. Es geht in der Regel darum, dass der Rechner aus einer Menge bekannter Beispiele durch Versuch und Irrtum in vielen Iterationsschritten selbständig ein Modell entwickelt, das diese Daten so genau wie möglich simuliert. Ist das mit ausreichender Zuverlässigkeit geschehen, kann man anschließend neue, noch unbekannte Daten eingeben und anhand dieses Modells bewerten. Außerdem benutzen die Forscher strukturelle Informationen über die Proteine, um daraus zusammen mit anderen Daten Rückschlüsse über deren Funktion zu ziehen.

Der Bioinformatiker vergleicht seine Arbeit mit der Suche nach Problemen bei einem kaputten Auto: „Bevor man ein neues Auto kauft, kann man sich die Schadensstatistiken für jedes Fabrikat ansehen und damit die Wahrscheinlichkeit verringern, dass man ein Schrottauto erwischt. Sobald man aber ein spezielles Auto hat, nützt dieses Vorgehen bei einem Schaden nichts mehr. Dann muss ein Experte ein Diagnoseprotokoll Schritt für Schritt abarbeiten und mögliche Fehler ausschließen, bis er am Ende die Ursache für den Schaden gefunden hat.“ Entsprechend sieht er den Nutzen seiner Arbeit für die Medizin: Neben grundlegenden Erkenntnissen ist es das langfristige Ziel, die Abläufe in der Zelle zu verstehen. Mit diesem Wissen lassen sich die genauen Ursachen für Krankheiten verstehen und man kann daraus im Idealfall individuell zugeschnittene Therapien für jeden Patienten ableiten. □

“You discover more about biological function from a detailed description than from a number. That’s why I love protein structures.”

Burkhard Rost

Graphics: ediundsepp (Data: TUM)

Each of our cells is a veritable masterpiece. About 20,000 different proteins work together as molecular machines in each of those, ensuring that each cell receives energy and can perform its specific functions; can move, nourish and renew itself, reproduce, protect against enemies, and send and receive signals. “A cell is fairly crammed with proteins – it is as dense as a solid,” explains Burkhard Rost, Professor of Computational Biology & Bioinformatics at TUM. “The image of a cell as a sack full of liquid in which proteins swim around is completely misleading. It is better to imagine it as a type of clockwork mechanism, with moving parts interacting and interlocking similar to the tiny cogs inside a watch.”
The genetic information is at the heart of all of this activity. It is present in the nucleus of each cell as part of the genome, in the form of a chain comprising around three billion base pairs, and the cell uses it as assembly instructions for its proteins. These usually have an extremely complex design, folding themselves in a particular way to form intricate structures such as spheres, branches, spirals, stalks and channels. ▷

Many beliefs held 15 years ago viewed as errors today

Since the British scientific "virtuoso" Robert Hooke discovered in the mid-seventeenth century that living beings are made up of tiny "units of life", which he termed "cells", scientists have been keen to gain an understanding of their interior and their inner workings. When the human genome was decoded in the 1990s, it was first thought that all secrets had finally been revealed. In fact, the real search had only just begun. It quickly became clear that, alongside the actual genes providing assembly instructions for proteins, the genome also contained significantly more segments, whose purpose remained a mystery. So these were initially deemed "junk" and disregarded. Meanwhile, we now know that they also hold information, albeit of a different kind.

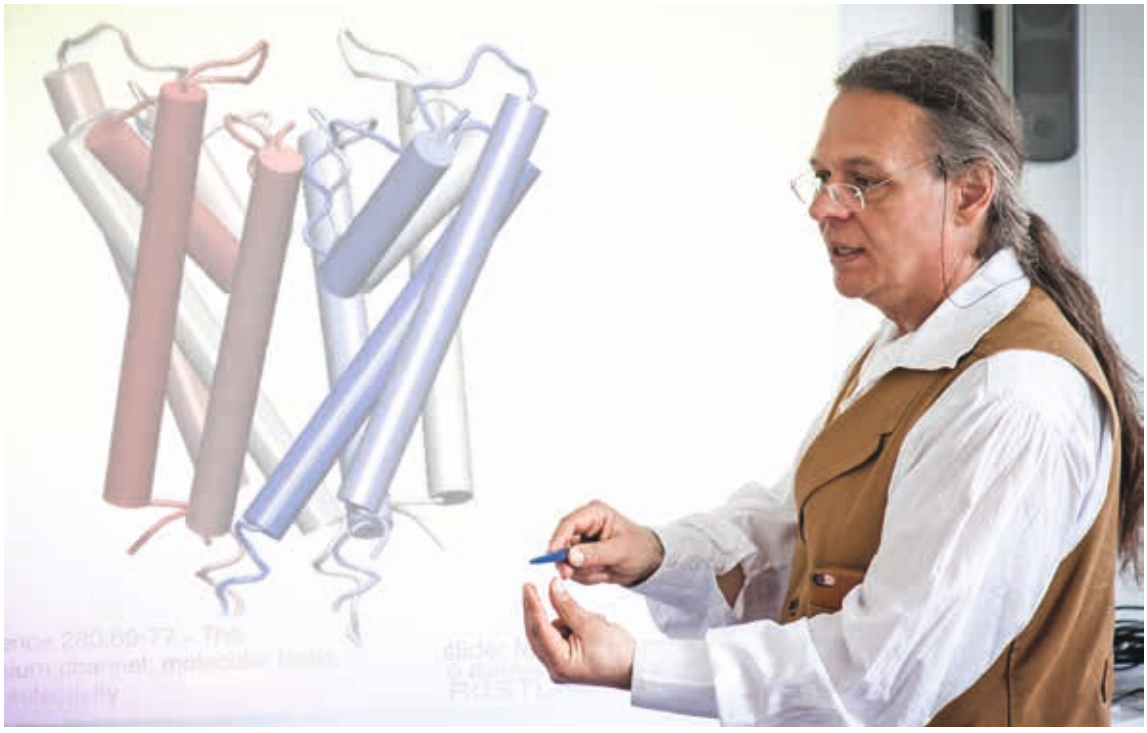
The proteins themselves also proved puzzling, however. As more and more human genomes were sequenced, the variation between genes was found to be significantly greater than was first assumed. As Burkhard Rost, Professor at Columbia University in the City of New York at the time, recalls: "Two people differ from one another in around 20,000 amino acids – the building blocks of proteins. On average, there is one difference in every gene. That is absolutely astonishing and we certainly didn't expect this 16 years ago. Back then, it was thought that there was a single reference genome, and all individual genomes could be expressed as small deviations from this reference genome. But we now know that the differences between any two unrelated people are too substantial for this."

So the researchers started to dig. If 20,000 variations in the genome between two healthy people produce no visible difference, it is probably safe to assume that these changes can be regarded as neutral. On the other hand, mutations had also been identified that adversely affect people's health, such as in sickle-cell anemia, which results from a point mutation on chromosome 11, i. e. by changing one single amino acid in all 20,000 proteins a disease is caused. Such conditions are known as "rare diseases". In contrast to infections or major, widespread diseases such as cancer or cardiovascular disorders, these are triggered by an early genetic defect. Angelman, Marfan and Treacher-Collins syndromes are examples here, as are Down syndrome and progeria – and the list runs into the hundreds. Each of these conditions is rare in its own right, but taken as a whole, they affect a good five percent of the population.

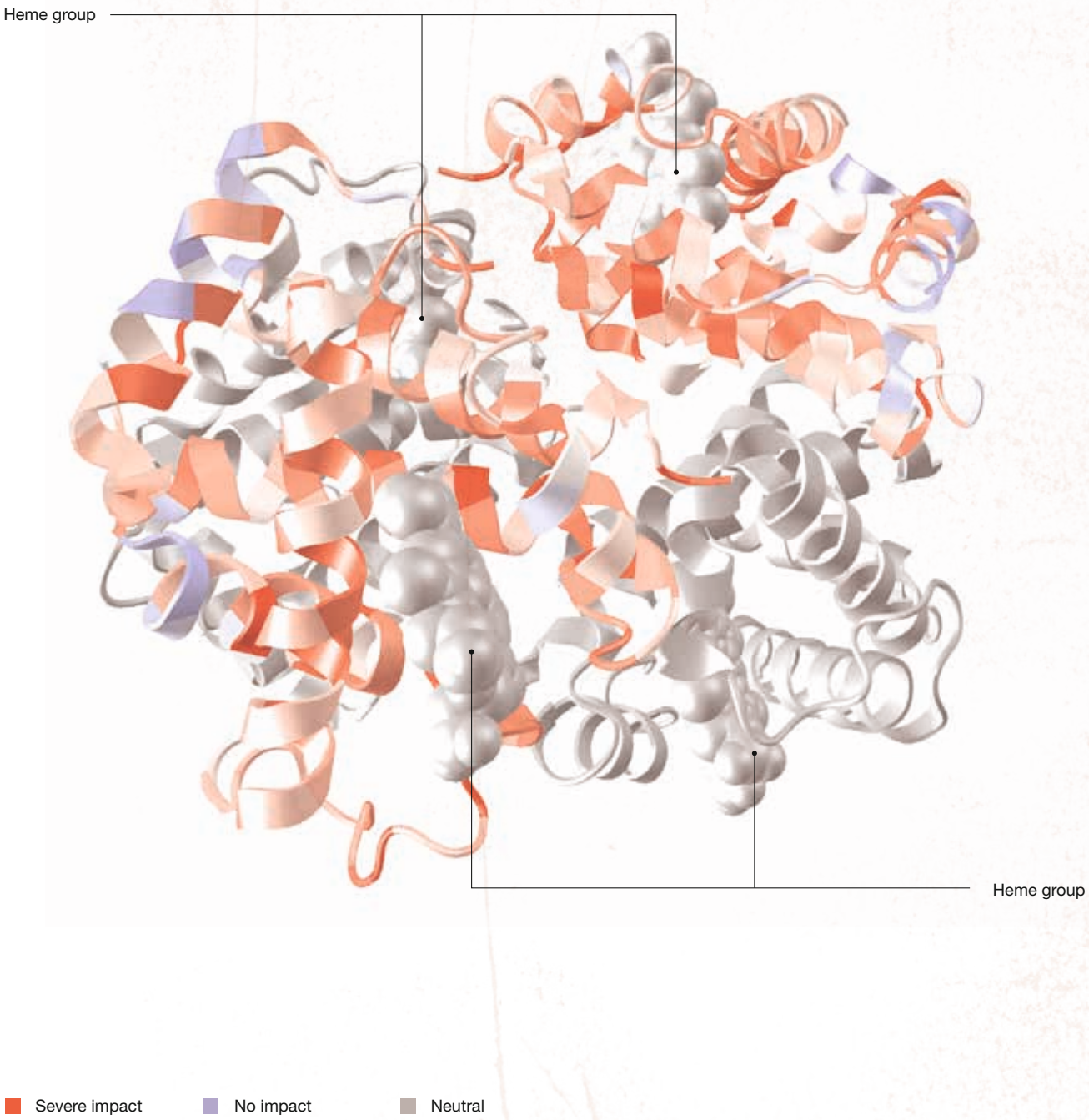
When looking at the effects of genes, these two scenarios represent the extremes: either no impact at all or a disease due to a single mutation. The reality is, however, more complex – as a rule, the outcome of genetic variations lies somewhere in between, with most diseases arising from several mutations that interact.

Artificial intelligence comes into play

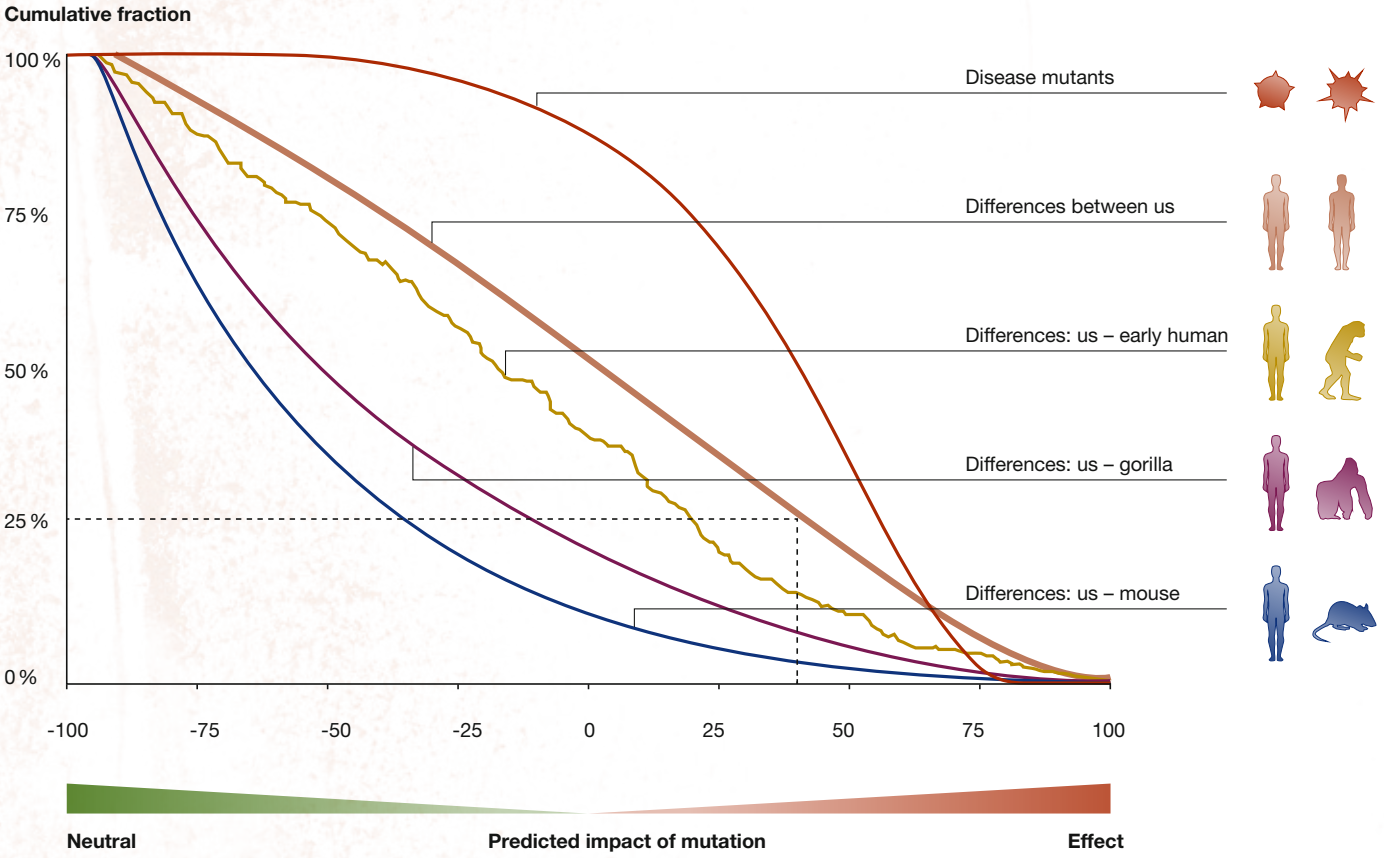
The question then was how to determine – or even predict – which variants in the genome have which effect? Burkhard Rost had a breakthrough idea in this regard 25 years ago, long before the variations had been identified and the ▷



Using tools like machine learning, researchers can predict whether mutations affect the function of a protein, depending on their location in the genome. This image shows the 3D structure of beta hemoglobin. One can clearly see the ring-like heme groups which bind to oxygen and thus transport the gas through the body. In the areas colored grey, mutations have a neutral effect. They change the protein sequence, but do not affect its function. Mutations in the red areas, such as the areas near the heme binding sites, have a significant impact on the protein function. Genetic differences in the blue areas have no effect at all.



Picture credit: ediundsepp, Graphics: ediundsepp; TUM/Rost; Schafferhans



The impact of mutations predicted with the help of machine learning: A strong effect (red) would result in a clearly detectable difference or in a disease, neutral (green) would have no effect at all. Of the 20,000 mutations in which humans differ and which can change protein sequences, about half are neutral, and half have an effect. About 25 percent of these have a strong effect (dashed line). Humans differ from their ancestors in a much higher number of mutations, but these have a much lower effect. On the other hand, mutations that are known to cause diseases are correctly predicted to have a very strong effect.

corresponding data made accessible. His suggestion back then was to compare the gene with its evolutionary relatives: "Suppose there is a particular enzyme in humans that also occurs in chimpanzees. And this protein has the same function in both cases. Maybe we then determine that the ape enzyme is not entirely identical in appearance to ours. But since there is no difference in function, we can assume that has no bearing. From this, we conclude: these variations have no effect; they are neutral – they alter neither structure nor function. If, however, we compare the relevant enzyme with that of mice, we might find that there are differences in function there. So by looking at the evolutionary family of this enzyme, we can see which variations have an impact and which do not. We can now feed this evolutionary profile information into a machine learning method or neural network."

Both of these options fall under the umbrella of artificial intelligence (AI) or machine learning (ML), which entails teaching computers to identify patterns without explicitly programming them in beforehand. Based on a number of known examples, the computer uses an iterative, trial-and-error process to autonomously develop a model that simulates this data as accurately as possible. If the outcome is sufficiently reliable, new, previously unknown data can then be entered and the model used to evaluate it. 1988 saw the first publications by researchers describing how they obtained information using neural networks. The real breakthrough came from combining AI with the evolutionary information, five years later in 1993. This method quickly spread. Today, Burkhard Rost's research team alone uses over 50 different AI methods to create 10 other AI methods that determine what effect variations in an amino acid might have.

The Munich-based researchers are not only focusing on the sequence of bases in the genome here. For instance, they are also analyzing structural information – that is, data relating to the form of the relevant proteins, primarily derived from X-ray crystallography. To do this, biologists have isolated and ▷

Graphics: ediundsepp (source: Y. Mahlich et al (2017) Scientific Reports 7, Article number: 1608), Picture credit: ediundsepp

Prof. Burkhard Rost

A passion for proteins

"I'm always excited to discover that something I believed was wrong," declares Burkhard Rost, one of the very first bioinformaticians. Like all his colleagues, at the outset he thought that the DNA fragments between genes were junk, for instance. Whereas now researchers know that they have other functions.

Burkhard Rost works on predicting the function and structure of proteins and genes, with a particular focus on forecasting protein interactions and the effect of variations in individual amino acids. His research findings are intended to foster a better understanding of protein, gene and cell function. Additionally, he aims to enable earlier detection of diseases and more effective treatment. His research group specializes in connecting artificial intelligence and machine learning with evolution.

After studying physics, history and philosophy at Giessen and Heidelberg universities, Rost received his doctorate at the European Molecular Biology Laboratory (EMBL) in 1994. He had already developed the first Web server for predicting protein structures in 1992. Following research visits to EMBL and the European Bioinformatics Institute in Cambridge (UK), as well as a brief period in industry at LION Biosciences in Heidelberg, he took up a professorship at Columbia University (New York City) in 1998. In 2009, he accepted an appointment to the Chair of Bioinformatics at TUM, where he holds an Alexander von Humboldt Professorship. He is a member of the New York Academy of Sciences and has been President of the International Society for Computational Biology since 2007.

crystallized the protein, before passing X-rays through it. They can then use the resulting diffraction patterns to draw conclusions – sometimes highly detailed – concerning the form of the protein. "International programs for structural genomics have made huge efforts to get closer to the aim of assigning a 3D structure to every protein," reports Burkhard Rost. "But we are still a long way from understanding them all. In humans, 3D structures have been experimentally determined for fewer than a quarter of all proteins to date."

A picture worth more than a thousand words

Nonetheless, this structural biology information can yield fascinating insights, since proteins work like tiny, three-dimensional machines. There are three different ways to describe them: first, in terms of their biochemical activity; second, by biological function; and third, by localization – that is, where their activity is sited, as in the mechanism of a clock. As Burkhard Rost emphasizes: "You discover more about biological function from a detailed description than from a number. That's why I love protein structures."

Here again, machine learning can tell us quite a lot about the possible 3D structure of proteins that have not yet been analyzed. Researchers can develop models showing how they interact with other known proteins. Another example entails investigating where in the cell the relevant protein must be located in order to perform its role: in the nucleus, where it facilitates gene transcription, or in the outer membrane, where it might function as an ion channel, or as a signal protein between the two?

Sometimes even a section with no discernable function can have a specific task. Here, researchers talk about "disorder", meaning parts of a protein that simply do not fold. "You can think of these pieces as the padding in Amazon parcels – they don't fold either," as Rost explains the phenomenon. "These areas are there to stop anything interacting with the protein. But the molecule can also use these fragments to scan their surroundings, and if something approaches, it emits a signal."

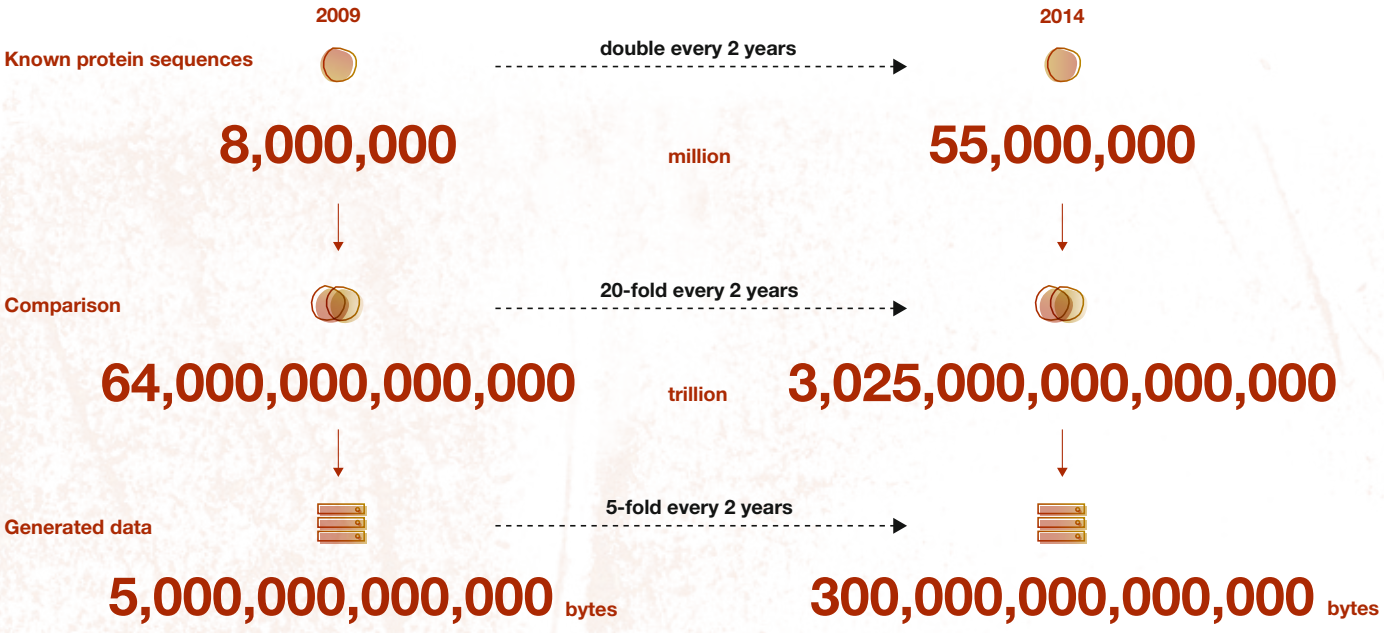
Ion channels are particularly complex. These are small pores in a membrane, able to open or close to allow something to pass through or block it. Few of these proteins, in particular, have revealed their exact structure to date, since they cannot be crystallized without their surrounding membrane. But even they can be analyzed using artificial intelligence: "If we identify a channel that does something quite different in bacteria compared to humans, for instance, we can ask why. Where are the similarities between the two; where are the differences?", describes Rost. "We also look at the structure: which part lies in the membrane; what signal causes the channel to open and close? Controlled by which amino acids? We can find this out through experiments, but we can often predict it too. To do this, we are investigating the entire ion channel family."

The difficulty with the data

The 55-year-old bioinformatician likens his work to searching for the problem when a car breaks down: "Before you buy a new car, you can look at the breakdown statistics for each make and thus reduce the probability of ending up with a "lemon". But as soon as you have an actual car, this process is of no use – it can't help you narrow down the reason for a break-down. In that case, an expert needs to work through a diagnostic protocol step by step, excluding potential sources of failure until they finally identify the root cause. Healthcare policy to date resembles the quest for breakdown statistics, but not really the search for actual causes of disease." Rost is determined that the latter should receive more support.

No matter what method is used, one thing is certain in bioinformatics: The more information you can feed into the models, the better they become. That is why researchers depend on obtaining as much high-quality data as they can. But there is a major problem here: "I am in favor of the free exchange of data," states Rost. "There are many issues we cannot address if the information gathered is not shared. Restricting data access on legal grounds is a mistake, from my point of view. We need to reach the stage where people understand: We can only move forward if I make my data available, even if other people then know I have a particular condition."

Brigitte Röthlein



Graphics: ediundsepp (source: TUM)

Big data is at the center of bioinformatics, as the discipline generates huge amounts of data which multiplies every few years.