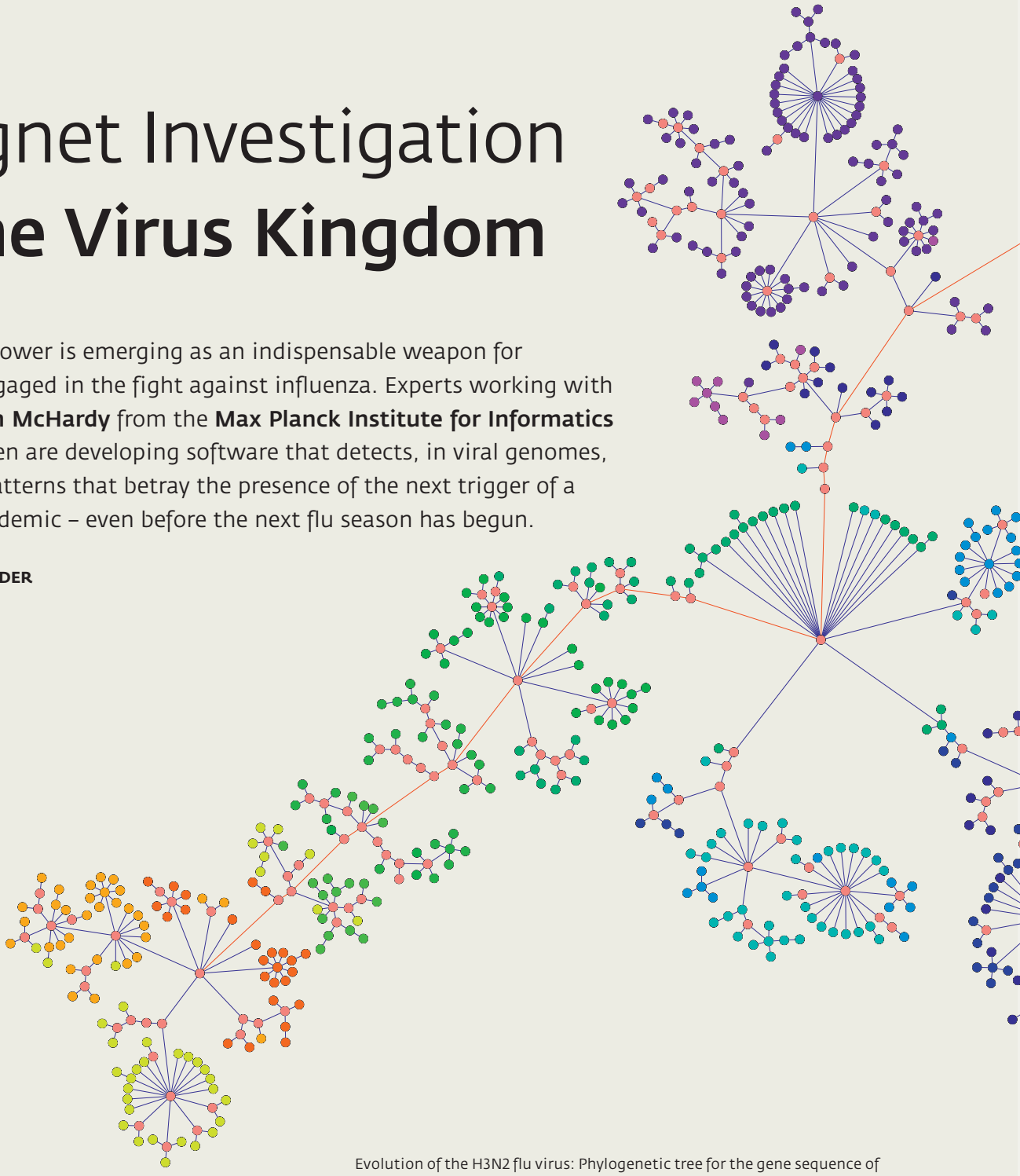
Dragnet Investigation **in the Virus Kingdom**

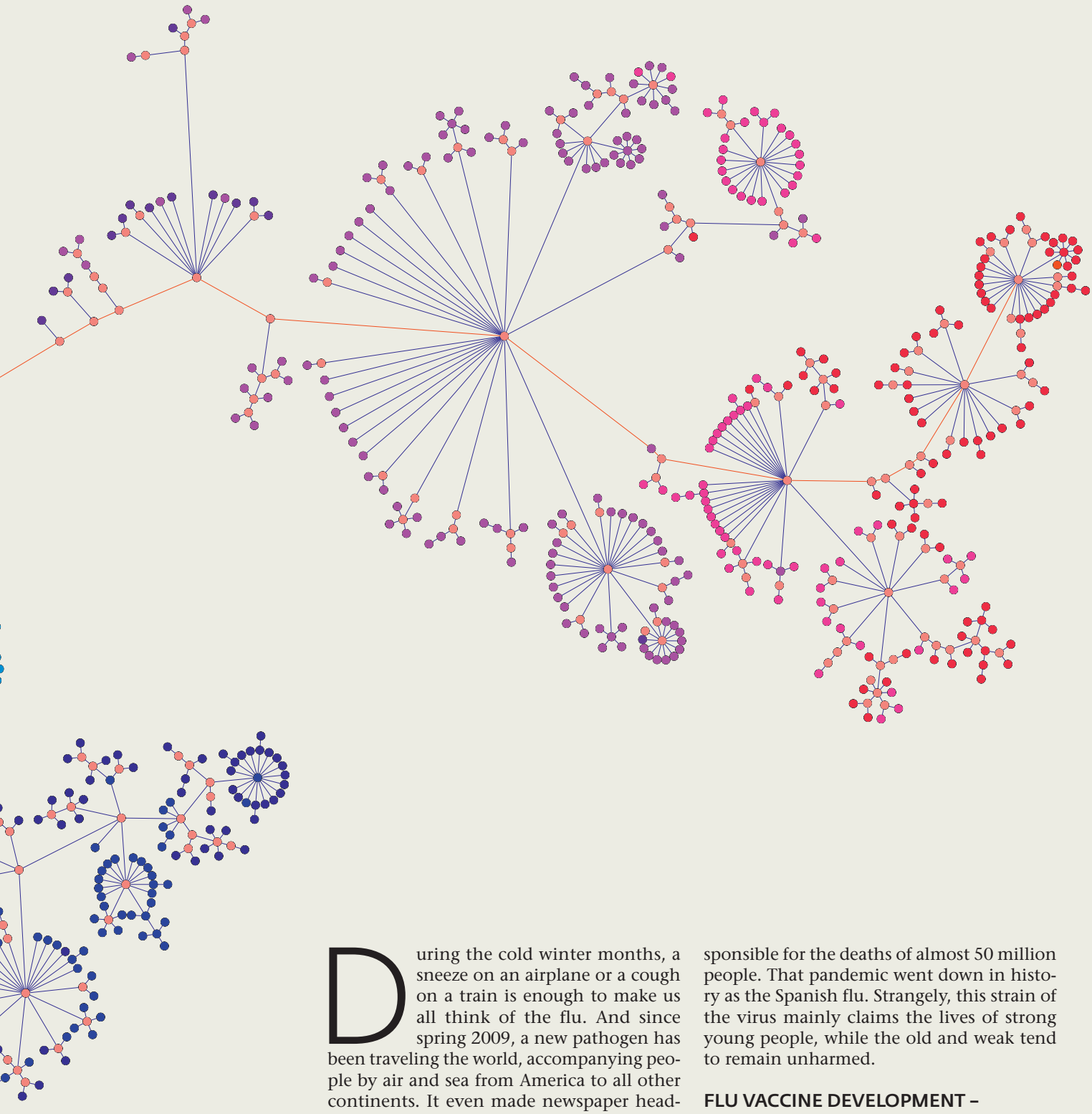
Computing power is emerging as an indispensable weapon for scientists engaged in the fight against influenza. Experts working with **Alice Carolyn McHardy** from the **Max Planck Institute for Informatics** in Saarbrücken are developing software that detects, in viral genomes, suspicious patterns that betray the presence of the next trigger of a global flu epidemic – even before the next flu season has begun.

TEXT **TIM SCHRÖDER**



Evolution of the H3N2 flu virus: Phylogenetic tree for the gene sequence of the surface protein hemagglutinin. The color of the end nodes indicates the time at which a particular variant was isolated. This enables scientists to trace the emergence of the different virus strains and predict possible new variants.

1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007



Graphic: MPI for Informatics

During the cold winter months, a sneeze on an airplane or a cough on a train is enough to make us all think of the flu. And since spring 2009, a new pathogen has been traveling the world, accompanying people by air and sea from America to all other continents. It even made newspaper headlines in fall 2009: the new H1N1 virus.

By mid-January 2010, 22,000 people around the world had died of the infection commonly referred to as the "swine flu". Just how many people have been infected is anyone's guess. What is certain, however, is that H1N1 has presented humanity with the first case of global influenza, the first pandemic of the 21st century. And yet, up to now, the virus has proven relatively harmless: in 1918, a predecessor of the new H1N1 virus was responsible for the deaths of almost 50 million people. That pandemic went down in history as the Spanish flu. Strangely, this strain of the virus mainly claims the lives of strong young people, while the old and weak tend to remain unharmed.

FLU VACCINE DEVELOPMENT – A RACE AGAINST THE CLOCK

For a long time, doctors had no explanation to offer for this. It was not until a few years ago that researchers were able to observe a similar phenomenon in monkeys. The pathogen clearly allows the immune system to "boil over." If a person is infected, his or her immune response releases an excess of infection messengers – inflammation-promoting substances that are supposed to help

Alice McHardy is head of the Computational Genomics and Epidemiology research group at the Max Planck Institute for Informatics.

» One aim of the bioinformaticians is that computers will one day be able to detect, in viral genomes, suspicious patterns that betray the presence of the next trigger of a global flu epidemic – even before the next flu season has begun.

fight the pathogen, but that attack the body's own tissues in the process. The immune systems of strong young people, in particular, simply produce too much of a good thing here.

Scientists are now very familiar with the flu virus. It has long been classified into different groups that can be clearly differentiated on the basis of genetic characteristics and typical protein structures. Large volumes of vaccine are produced against human flu viruses every year, with the intention of protecting against infection. However, sometimes the viruses mutate so rapidly and unexpectedly in some corner of the world that humans are unable to act against them in time. A new pathogen that is not targeted by existing vaccines spreads quickly. The fight against flu viruses is thus, above all, a race against the clock.

Will the scientists succeed in tracking down a new virus variant in time before it establishes itself globally in the next impending flu epidemic? Only then can manufacturers adapt the vaccines to the new pathogens before they trigger a major flu epidemic. The researchers usually win this race. Around every four years, however, the viruses outrace them.

The scientists from the Max Planck Institute for Informatics are thus arming themselves with computing power in the war against the flu virus. The experts working with Alice Carolyn McHardy are developing software that specializes in extracting secrets from the genetic material of viruses and bacteria. One aim of the bioinformaticians is that computers will one day be able to detect, in viral genomes, suspicious patterns that betray the presence of the next trigger of a global flu epidemic – even before the flu season has begun.

WHICH GENETIC CHANGE MAKES THE VIRUS DANGEROUS?

There are three types of influenza viruses: types A, B and C. The most significant are the influenza A pathogens, as these cause the major pandemics. For influenza A, there are many dozens of different flu virus strains that circulate among birds, but also among pigs. Occasionally, a virus emerges from these animal viruses that makes the transition to humans, like the current swine flu virus. The normal human flu viruses, in contrast, continue to develop from year to year, particularly in Southeast Asia. From there they spread throughout the world just in time for the annual flu season.

A flu virus is like a prickly ball whose spines are formed by the protein hemagglutinin. At the tips of the spines is a kind of lock structure at which the virus can deliberately dock onto the surface of animal and human cells. Whether or not the structures on the cell surface fit into the virus hemagglutinin and thereby gain entry into a cell depends primarily on the fine structure of this binding site. Should the two parts find each other, the disaster takes its course. The membrane of the host cell opens and the virus slips into the cell and releases its genetic strands in the interior of the cell.

The virus reprograms the cell into a compliant zombie and the cell becomes a virus production site. It obediently synthesizes virus components that are then combined to form hundreds of new viruses. At this point, a second important virus protein, neuraminidase, becomes active. The neuraminidase opens the cell membrane in such a way that the freshly produced viruses can pour out like soldiers from a troop carrier.

Viruses can carry different variants of the hemagglutinin (H) and neuraminidase (N) proteins on their membranes and are classified accordingly: H1N1 or H3N2, for example. But how do viruses from pigs transform themselves into potent human pathogens? First, through changes in their genetic material. Viruses replicate at breakneck speed. After just a few hours, infected cells release millions of new viruses. The genome must be duplicated for each new virus generation. Errors frequently arise in this process. Some genetic components (the bases) are copied incorrectly during the reading of the genomic template. Such mutations are occasionally fatal to the virus itself. Sometimes they are irrelevant. From time to time, however, the virus becomes really dangerous as a result of this process.

Photo: Manuela Meyer

Moreover, another characteristic of the flu viruses gives them an extraordinary capacity for adaptation: their genome is not composed of a single strand, but is cleanly packed in eight individual packages called segments. Like suitcases at an airport, it is entirely possible for such packages to become mixed up. This can happen if a cell is simultaneously infected by two viruses – a human virus and a pig virus, for example. Both viruses pour their genomic segments into the cell – a total of 16 segments that are replicated at breakneck speed.

A DENSE SEARCH GRID

From time to time, the following sequence of events unfolds: During the assembly of the new viruses, the reproductive apparatus steers a segment from the wrong virus into the offspring. This mixing of the genetic material is known as reassortment. This may well have occurred with the swine flu: different viruses contributed segments to the new virus and this changed its characteristics in such a way that it became an agent of disease that is also transmitted between humans.

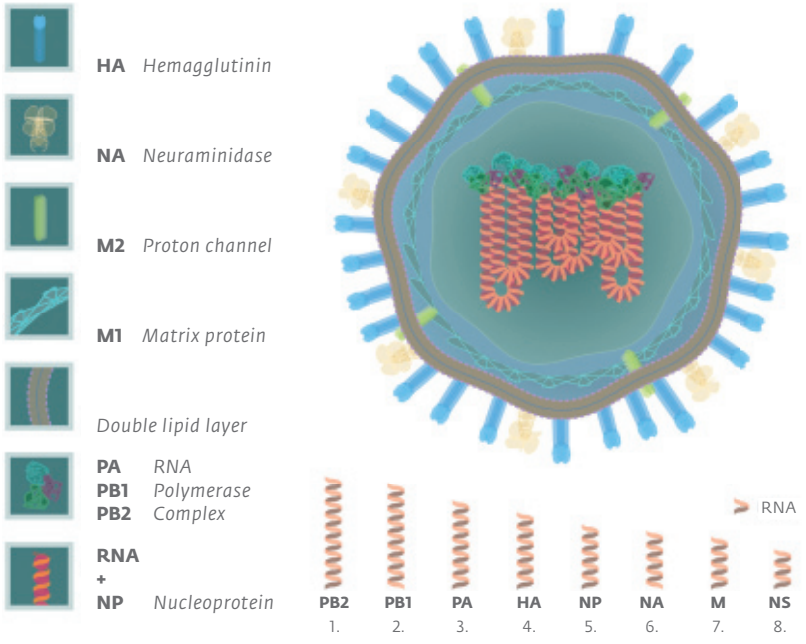
When the vaccine developed against the H1N1 virus is administered, the body produces antibodies that recognize and block the hemagglutinin on the virus. The virus can no longer dock onto the cells, and the infection is halted.

Researchers are constantly on the lookout for genetic changes in influenza A viruses that could pose a threat – not only in the new H1N1 swine flu virus, but also in the other known suspects, such as the classic human H1N1 and H3N2 viruses. The global search grid is dense. Blood samples are regularly taken from patients at 112 institutions and clinics in 83 countries on behalf of the World Health Organization (WHO). They are then genetically analyzed and sent to four large flu centers.

Every six months, a committee of experts convenes to examine the latest genetic analysis data to detect suspicious mutations or even reassortments. The experts try to identify from the gene sequences how the virus' form, its proteins, the hemagglutinin and the antigen will change to enable the virus to circumvent the current vaccine. They try to predict what form a new vaccine that will help fight such a virus should take if the virus in question were to pose a threat. Accordingly, the vaccines are developed on an ongoing basis and adapted to the modern influenza A types. But this does not always work.

"Up to now, the exact links between the genetic changes and the emergence of a new flu strain have not been fully understood," says Alice Carolyn McHardy, head of the Independent Research Group Computational Genomics & Epidemiology. "For example, why do some reassortments cause viruses to trigger the illness, while others do not? Does this require other genetic characteristics?" In an attempt to find answers to these questions, McHardy has transformed the computers in her office in Saarbrücken into prognosis tools whose task it is to unravel the complex interactions at work here.

She uses statistical learning techniques for this purpose. "These methods are able to relate the most wide-ranging data sets with each other and to track down hidden connections between them." To begin with, the learning method is fed with known data, such as the genetic information from viruses that triggered flu epidemics in the past. To this is added the information about when and where the virus showed up or the form taken by the protein structure of the hemagglutinin. Based on a process akin to a dragnet op-



Schematic diagram of the influenza A virus. The virus is surrounded by a lipid membrane from which various proteins protrude. With the help of the membrane proteins hemagglutinin (HA) and neuraminidase (NA), the pathogens can dock onto their host cell and leave it again. However, these proteins are also the pathogen's greatest weakness: they can be blocked by the immune system's antibodies. The viral genome consists of eight RNA segments to which different proteins are attached in the viral particle.

Graphic: MPI for Informatics

» Researchers collect soil samples by the shovelful and analyze the entire genetic material of the microbial inhabitants. The hope is that promising genes will be found that contain information about new superproteins.

eration, the computer becomes acquainted with the typical perpetrator profiles of disease-causing viruses.

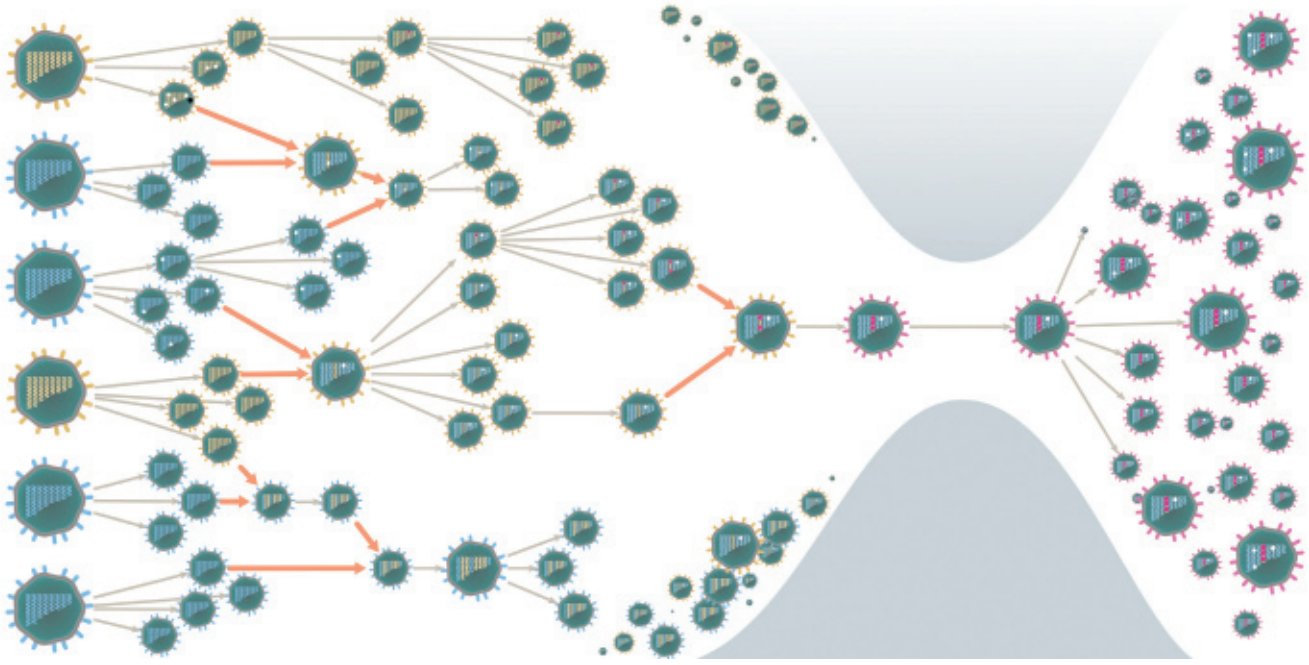
MORE RELIABLE PROGNOSIS FOR THE DOMINANT VIRUS

When the method has been trained extensively with the existing data sets, things really get going. McHardy feeds the computer with the genetic information of an influenza A virus whose level of danger is unknown. The learning method then compares new data with what is already known – that is, it orders the information in such a way that the current input values approximate the learned patterns as closely as possible. The computer then produces numerical values that answer the following questions: To what extent do the new gene sequences resemble the structures of a successful virus? How likely is it that there will be an outbreak of the disease?

McHardy uses support vector machines (SVM) and other statistical learning techniques for this work. These transform information into data points: numerical values that hang like stars in the firmament of a vast number space. When new data is entered into the system, the SVM assigns the new values to the floating numbers – the closer they are to a trained value, the more similar the data.

The Saarbrücken-based researchers are as yet unable to provide current forecasts for the new flu season, as the programs are still being optimized. Nevertheless, when McHardy recently entered the data for a well-known flu virus into the SVM for test purposes, it duly responded "very likely to be infectious."

"But we don't just want to find out whether a new virus variant will become established in the future and become the flu pathogen of the season," says McHardy. "We also want to be able to forecast when this will happen." The WHO experts are currently trying to predict for one year the new virus strain that will dominate and thus spread throughout the world as the agent of



The variety of flu viruses: Different strains of influenza A viruses (yellow and blue virus particles), which replicate in the human body (white arrows), circulate among the population. Numerous different genetic variants arise as a result of the redistribution of the genotype between two strains of the virus (reassortment, orange arrows) and mutations (white dots). With time, the human population becomes immune to these variants; this is represented in the diagram by the narrowing bottle neck. Virus strains with mutations that give rise to altered surface proteins (red dots) gradually become more common until a variant eventually arises against which the immune system is, in many cases, powerless. The new variant can then spread widely, triggering infections and creating similar new strains.

Graphic: MPI for Informatics

Fighting the flu with computer power: The scientists from the Computational Genomics and Epidemiology research group on their quest for the trigger of future flu epidemics. Back row, from left: Christina Tusche, Kaustubh Patil, Johannes Droege; front row, from left: Lars Steinbrück, Alice McHardy und Sebastian Konietzny.

the disease in the ensuing flu season. They currently miss this target in one out of four cases. When this happens, a different variant of the virus becomes established than the one they expected, and then there is no well-matched vaccine available for it.

McHardy's goal is to increase the reliability of this annual forecast. Flu arises in the northern hemisphere primarily from November to February, when weather conditions are cold and damp. In the southern hemisphere, people are affected by the illness in the southern winter months, between May and October. Such data are also fed into the SVM and other learning methods to be able to understand the viruses. McHardy has no doubt that it will one day be possible to produce a more reliable forecast. The software program Geno2Pheno, with which doctors can determine how quickly resistance to AIDS drugs develops, was already developed at her institute (MaxPlanckResearch 4/2005, p. 21 ff.).

Moreover, McHardy has an abundance of computing resources to draw on in her work, as GISAID, the world's largest virus gene database, is located in the cellar of the institute. For a few years now, researchers from all over the world have been feeding the results of the genetic analysis of influenza viruses into this database, which other researchers can access free of charge. "It makes complete sense to pool the data about the viruses, as this helps us to understand their nature as a whole," says McHardy. "Viruses are small with a very manageable genome. Researching causal relationships with them is thus a straightforward matter."

Viruses are one of McHardy's interests. The other is the genome of bacteria, which she also aims to decode with the help of support vector machines and related learning processes. The researcher specializes in imposing order on metagenomes – a wild mixture of genetic sequences identified from a variety of different microorganisms. A metagenome does not represent the genetic material of a single individual, but of many organisms – sometimes even thousands of them. But what is the point of studying this chaos?

THE SEARCH FOR SUPERPROTEINS

It is a known fact that extreme habitats, such as hot springs, for example, spawn special life forms – for instance, bacteria that can thrive in water as hot as 120 degrees Celsius. Scientists and industry expect that studying these bacteria will give rise to new substances, such as heat-resistant proteins. These could be used, for example, in the production of cosmetics or foodstuffs, namely for production processes that require high temperatures. For this reason, scientists are now searching on land and at sea for such extraordinary new microbes.

The simplest solution would be to be able to breed the unicellular organisms in the laboratory and investigate them thoroughly there for new substances. However, many bacteria do not thrive in test tubes. Researchers have thus been rolling up their sleeves for some time now, collecting soil samples by the shovelful and immediately analyzing the entire genetic material of the microbial inhabitants. The hope is that promising genes will be found that contain information about new superproteins.

The problem, however, is that metagenomic analysis usually provides thousands of minute genetic fragments, of which only a few can be assigned to an organism. This is where McHardy's method, which the researcher has already fed with the genetic fragments of known bacterial groups, comes into play. In this instance, the support vector machine was trained especially in a characteristic of the bacterial genome: short, recurring sequences of bases, known as oligomers, such as the base sequence ACTGAT. Interestingly, certain oligomers are characteristic of the genome of different bacterial groups, just like a fingerprint.

These oligomers arise not only in one, but in several locations of the DNA

Photos: Manuela Meyer (2)

Grand Prismatic Spring is the biggest thermal spring in Yellowstone National Park. Hot springs like this host bacterial communities that are particularly adapted to the extreme environmental conditions of this habitat. However, many of the bacterial species cannot be cultivated individually under laboratory conditions. For this reason, scientists are now jointly analyzing the genetic material of all of the microorganisms found in such habitats.

strand. As a result, oligomers offer a highly suitable means of imposing order on the metagenomic puzzle. When the support vector machine had learned which oligomers are associated with certain bacterial groups, McHardy fed new metagenome sequences of unknown origin into the system, for example sequences from microbe-rich sewage sludge. Once again, the data hovered past the learned numerical values in the vast number space. The experiment was successful: "Based on the characteristic oligomers, the program was able to assign many of the short metagenome sequences to certain bacteria."

DIGESTION WITHOUT ANY METHANE EMISSIONS

This process is called "binning" and involves the assignment to the correct bacterial group – the respective "bin" – fragment by fragment. In some cases, the statistical learning method can allocate up to 90 percent of the genetic fragments correctly based on the oligomers. How well the method works depends, ultimately, on the volume and quality of the training data. The computer scientists in Saarbrücken only occasionally achieved values of between 30 and 40 percent.

Among the most intriguing metagenome studies in which Alice Carolyn McHardy has participated is the analysis of bacterial communities from the intestines of termites and digestive systems of the Australian wallaby, a bush kangaroo. Both species digest wood and release hydrogen, the molecule that humans aim to one day exploit on a mass scale in the production of fuel cells.

The distinctive feature of the digestive process of these insects and mammals is that, unlike bovine digestion, almost no climate-damaging methane gas is released during hydrogen production. If scientists could succeed in imitating this process in the laboratory, it may be possible to develop a completely new environmentally friendly method of hydrogen production.

The results of the analysis are promising. McHardy's SVM was able to assign the crucial metagenome fragments to different microorganisms. As a result, it is now known which bacteria in the animal intestine participate in the miracle of climate-neutral hydrogen production – and, moreover, which proteins and metabolic processes lead to hydrogen. "Work is now being carried out to identify from the samples the microorganisms that are involved, and to analyze them further," says McHardy.

The metagenome analysis is still in its early days. Many habitats have thus far been studied only at a very basic level. "And we always need some preliminary knowledge to train our methods," says the Max Planck researcher. "But the advantage of the SVM method is that its training requires only small volumes of data," she adds. And McHardy has already shown what can be achieved. The search for the secrets contained in genome sequences continues. Thanks to the hunt in extreme habitats, such as the Arctic, where bacteria thrive in minus temperatures, metagenomics is gaining in significance. And thanks to bioinformatics, some other new, out-of-the-ordinary discoveries are bound to soon join the wallaby intestinal flora. ◀

GLOSSARY

Pandemic
A pandemic is understood to be a disease – usually an infection – that spreads across countries and continents in a relatively short period of time.

Reassortment
Reassortment is the mixing or redistribution of genetic information between two similar viruses. For this to occur, the two virus types must replicate in the same infected cell, and their genome must consist of several segments.

Metagenome
A metagenome is the entire genetic repertoire of a habitat – a cliff in the mountains, for example, or the edge of a hot spring.

Neuraminidases
Neuraminidases are enzymes that are firmly anchored in the membranes of many viruses. They act as door openers, so to speak, in that they open the cell membrane, thus enabling the freshly produced viruses to pour out and proliferate.

Photo: iStockphoto