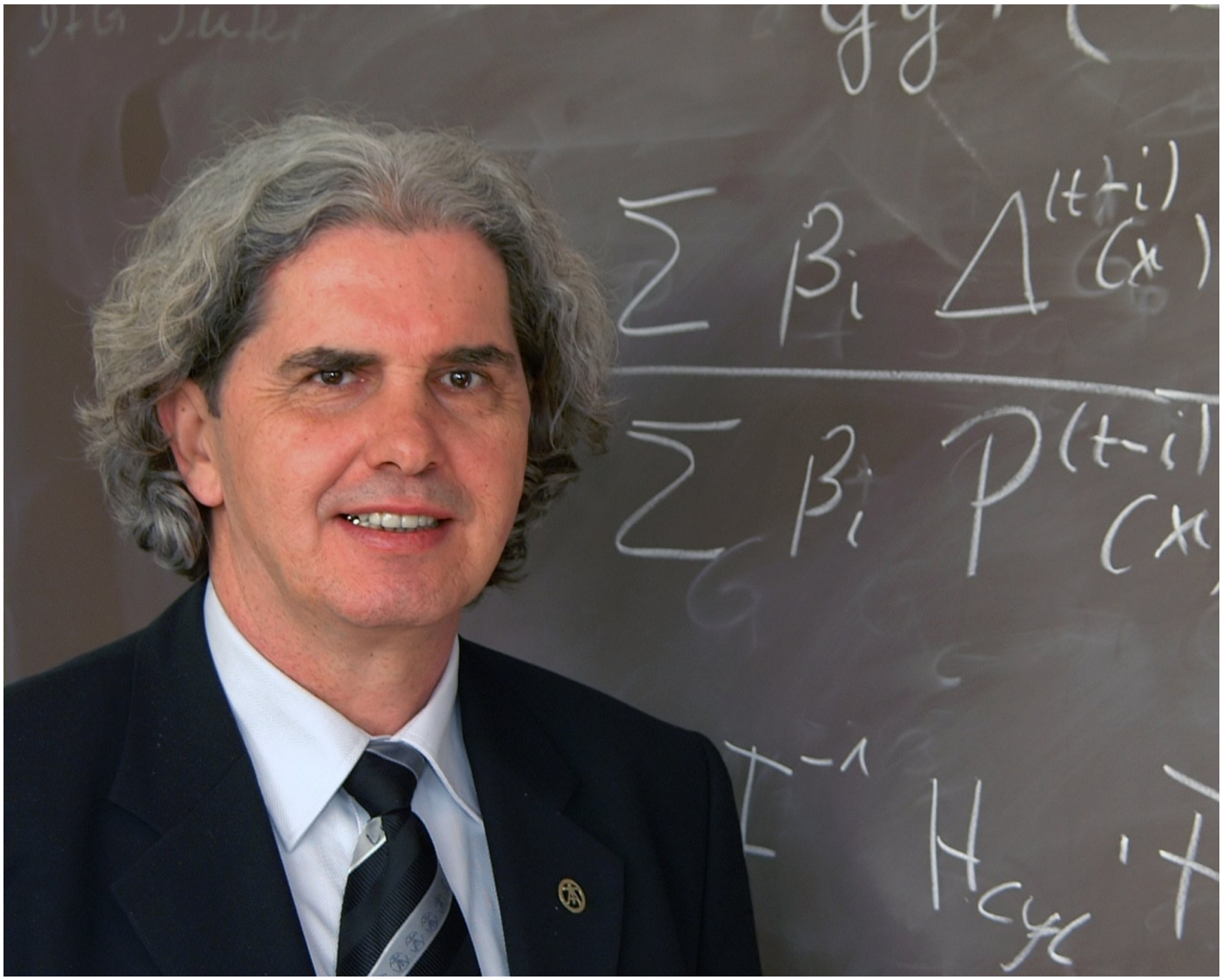
Healthcare industry BW

Molecular biology assisted by information theory

What do the Internet and mobile communication have in common with the division of yeast cells and viruses? Quite a lot, says Martin Bossert, professor at the Institute of Telecommunication Technology and Applied Information Theory in Ulm. The 55-year-old engineer coordinates an interdisciplinary priority programme of the German Research Foundation (DFG, SPP 1395 Information and Communication Theory in Molecular Biology) that applies information theory approaches to issues dealt with in biology.

Information-theoretical models can be used to make predictions on cellular behaviour – this is what Bossert is hoping for, and he has good reason to believe that it can be possible. He also believes that these models should make expensive measurements obsolete as well as help to identify errors and correct erroneous sequencing data generated by biologists.

Baden-Württemberg holds the lion's share



Prof. Dr. Martin Bossert is the coordinator of the new DFG priority programme.

This interdisciplinary approach is relatively new in Germany. In California and Great Britain, information theoreticians and life scientists have been working together for quite some time in purpose-built research buildings, explains Bossert. Bossert is one of only nine information theoreticians at German universities.

Information theory was developed by Claude Elwood Shannon at the end of the 1940s. The mathematician and electrical engineer, who died in 2001, did his doctoral thesis at the MIT in 1940 ("An algebra for theoretical genetics") using mathematical approaches to elucidate genetic aspects. Only a few years after Shannon had finished his doctoral thesis, the DNA double helix was discovered in 1953 and it became clear that the information of life was digital, and consisted of only four letters (the nucleotides G, A, T, C). In an attempt to make his approach appear less exotic, Martin Bossert explains that Gregor Mendel's laws of

inheritance have a lot in common with mathematics.

It was around 20 years ago that Bossert first discussed the similarities of information theory and genetics with some Russian colleagues, and did not return to the issue again until 2004. He hired an assistant to work on the complex preliminary work and this eventually led to a joint research proposal with his colleagues Joachim Hagenauer from the Technical University in Munich, Hans Peter Herzel (Institute for Theoretical Biology, Humboldt University) and the evolutionary biologist Michael Kühl (Institute of Biochemistry and Molecular Biology) from Ulm.

DNA is a file that contains a huge amount of information

"The actual process was not difficult," said Bossert referring to mobile communications and Internet systems of similar complexity. Bossert holds that cellular processes can be described in detail using information-theoretical methods. Regardless of all the (bio)chemical processes in a cell, Bossert believes that there is a kind of hard disk that contains all the working instructions. By "hard disk", Bossert means DNA, a long sequence of four letters. In information theory, this would be an information-containing file consisting of a tetravalent alphabet.

"The same mechanisms in cell biology and communication"

Bossert finds it very exciting that telecommunications engineering and information theory are faced with the same issues as biologists. "Biologists and information theoreticians alike deal with information that comes from somewhere, be it in the form of a sequence of letters or symbols," said Bossert. This is the same in cells and on the Internet. With regard to cellular processes, it is important to know where the process starts, e.g., where the transcription of DNA is initiated and knowing which amino sequence codes for a protein. With regard to the Internet, the same question applies: where does the data package begin and where does it end? With both DNA and the Internet, it is important to know where the reading frame starts and where it ends. "We call this synchronisation," said Bossert.

Mobile phones and navigation systems would not exist without information theory

Bossert finds it difficult to understand why information theory is relatively unknown in Germany since it more or less has a huge influence on our daily life. Many of our communication media such as navigation systems, DVDs, CDs and mobile phones are shaped by findings made in information theory.

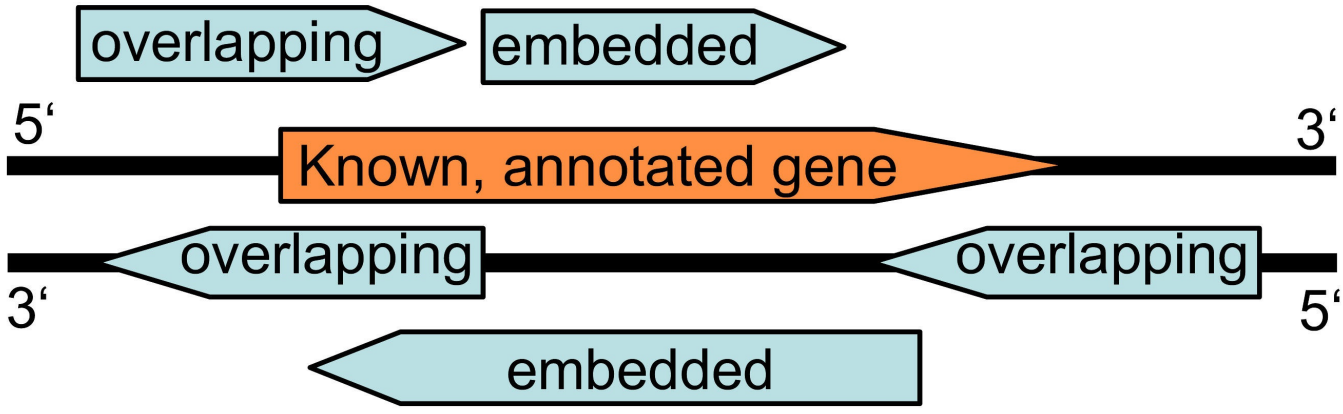
Information theory is based on probability theory and statistics and can make statements as to how much clearly defined information can be transmitted over a noisy channel. In other words, the theory establishes the number of bits (smallest information units) needed to represent the result of an uncertain event without any losses. Bossert explains that information theory is based on two axioms: for a given channel, it is possible to calculate how much reliable information can be transmitted. In addition, it highlights the number of bits that are required to reliably represent the result of a source without any losses. Without information theory there would be no email or mobile phones, explains Bossert.
The functions and conclusions of information theory are independent from the physical medium, which is why some of the more advanced theories can be applied to biology, cell biology and intercellular communication. "This was the basic idea behind this priority research project," explains Martin Bossert.

In contrast to bioinformatics which uses effective algorithms to extract specific features from huge quantities of data structures, information theory mainly uses models to explain observations. Bossert believes that bioinformatics alone is no longer able to deal with and analyse the huge amount of data produced by life scientists, and that communication-theoretical approaches are required to do so.

Information theoreticians discover guinea pig puzzle

Bossert believes that they would not have received the DFG grant if there had not been substantial evidence that information-theoretical models could be transferred to molecular biology. Joachim Hagenauer, an information theoretician from Munich, showed that mutual information can be used for phylogenetic classification, for example by comparing DNA sequences with each other. The mutual information refers to the strength of the statistical relationship between two random variables.

Scientists without previous biological knowledge calculated mutual information from the DNA sequences of animals stored in online databases and used information theoretic distance measures to assess the degree of the animals' phylogenetic relationship with each other. The phylogenetic classification obtained by these scientists was confirmed by biologists, although, surprisingly, neither information theoreticians nor biologists were able to solve the phylogenetic relationship of guinea pigs (P. Hanus, J. Dingel, J. Zech, J. Hagenauer and J. C. Mueller, Information theoretic distance measures in phylogenetics, Proceedings of the International Workshop on Information Theory and Applications, Jan. 2007, p. 421-425).

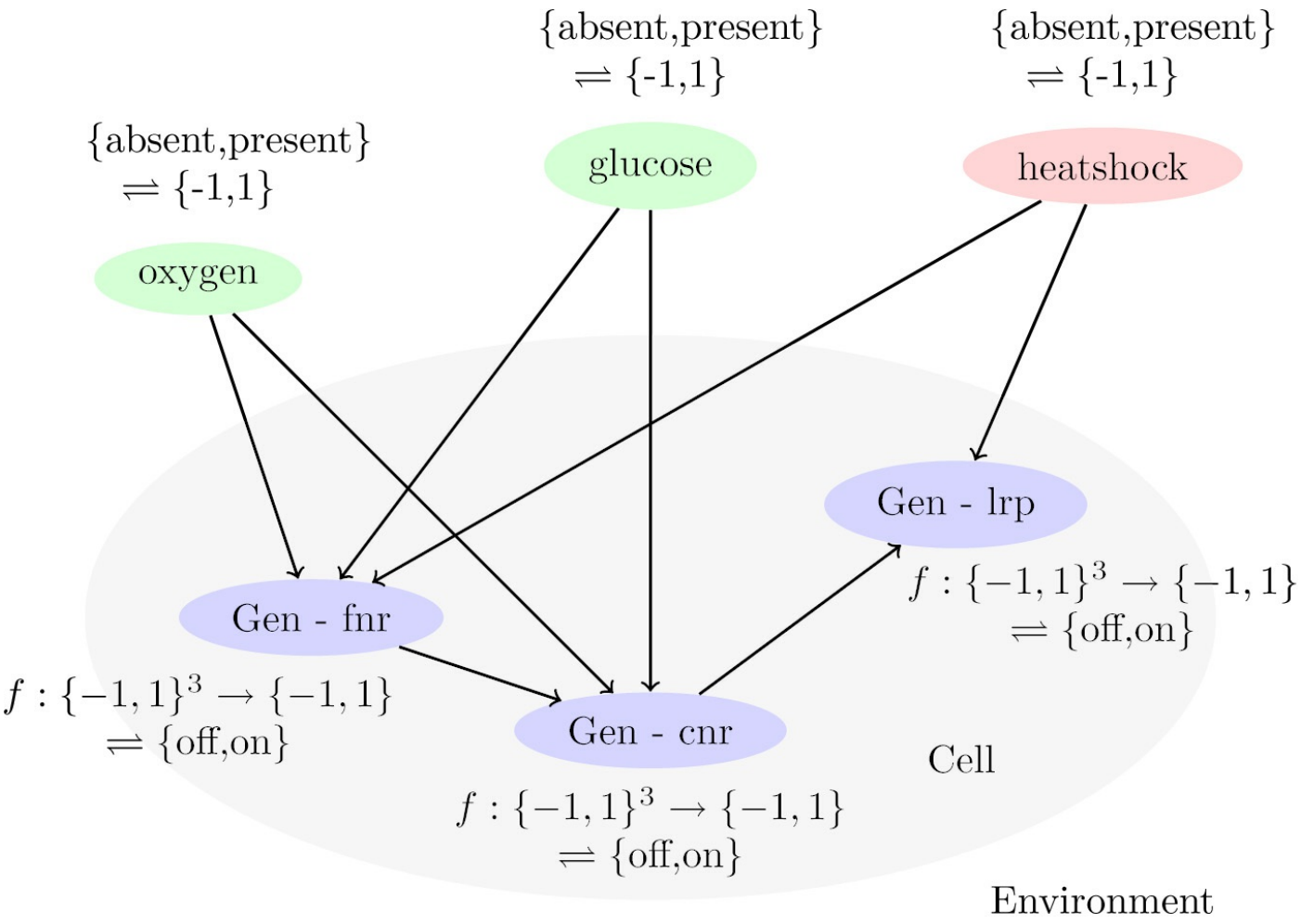


In close cooperation with molecular biologists (TU Munich), computer scientists (UKON Konstanz) and information theoreticians (TAIT Ulm), Bossert's team hopes to find new overlapping protein-coding DNA sequences and understand their basic mechanisms.

The phenomenon of overlapping genes

Besides coordinating the DFG project, Bossert is also involved in two subprojects. He is working with the microbiologist Siegfried Scherer (Department of Microbial Ecology, TU Munich) and the computer scientist Daniel Keim from the University of Konstanz. The researchers' goal is to investigate overlapping genes in prokaryotes. Recent publications suggest that there are more overlapping genes than has previously been assumed.

The researchers hope that this research will provide them with answers to questions about the frequency of overlapping genes and the genes' evolutionary development, in particular the genes of bacterial pathogens. The computer scientist will analyse the huge quantity of data acquired and the biologist will investigate whether certain DNA regions code for proteins. The goal of the project is to predict coding regions on DNA. The Ulm researchers bring their know-how on the coding theory (error correction code, random coding) to the project. The comparison of knowledge is based on molecular biological data and prognosis models; the researchers will then identify those DNA regions that have a high probability of coding for proteins. They hope that this method will help do away with expensive measurements.



The figure shows an example of a regulatory Boole's network for the metabolism of the bacterium Escherichia coli. The junctions represent genes and metabolites. Lines represent regulatory dependencies. The network is used to investigate the evolutionary adaptation of E. coli.

Prediction of E. coli reactions

In another project, Bossert is working with the biologist Georg Sprenger (Institute of Microbiology, University of Stuttgart) and the control engineer Oliver Sawodny (Institute of Systems Dynamics, University of Stuttgart). The Stuttgart researchers own an "Escherichia coli computer" that causes E. coli strains to undergo directed evolution. The information theorist knows about the substrates used and metabolic products produced. He will then calculate and use models to predict the processes and variables in the E. coli strain. Bossert explains that this is basically the same approach as the one used in a mobile radio unit when it receives a transmission, and subsequently seeks to discover the content of the transmission. Bossert and his partners from Stuttgart are aiming to calculate and model the protein concentrations and protocols used by E. coli bacteria to incorporate certain mutations in order to reach a particular evolutionary state.

It took a number of years before information theoreticians and biologists managed to understand each other. Each side had to learn about their partner's field of research before they were able to prepare a joint grant proposal for submission to the DFG.

Since the establishment of information theory 60 years ago, a huge range of statements, axioms, theorems, processes and algorithms have been produced. Bossert believes that the DFG priority programme will show that information theory can be applied to solving biological problems. Bossert knows from colleagues in other countries that the new interdisciplinary alliance of information theoreticians and molecular biologists will continue to exist beyond the funding period.