Association for Computing Machinery
Advancing Computing as a Science & Profession

Ubiquity, an ACM publication
March 2016

The Rise of Computational Biology

An Interview with Prof. Thomas Lengauer
Max Planck Institute for Informatics
Saarbrücken, Germany

by Walter Tichy

Editor's Introduction

In 2000, Craig Venter and Francis Collins presented a draft of the human genome sequence. Venter and Collins were competitors in the race to the human genome. Collins headed the public consortium that was allocated almost $3 billion to sequence the human genome by the year 2005. In the late '90s, Venter applied a faster method, called "whole genome shotgun sequencing", which relied heavily on computers. This was a $300 million project. It accelerated the assembly of the genome significantly and the public human genome-sequencing project was soon completed in 2003. Today, the cost of sequencing the human genome has dropped dramatically, to below $1000. This drop by five orders of magnitude is much larger than Moore's law would predict.

Cracking the mysteries surrounding the genome holds enormous promise: healing cancer, understanding the link between gene variations and disease, tracing evolution, and even designing synthetic cells. DNA sequencing and analysis are now inexpensive enough to be done routinely. In fact, in the near future, the DNA of every newborn might be analyzed to check for genetic issues that require immediate therapy.

In this wide-ranging interview, we will hear from a pioneer in computational biology on where the field stands and on where it is going. The topics stretch from gene sequencing and protein structure prediction, all the way to personalized medicine and cell regulation. We'll find out how bioinformatics uses a data-driven approach and why personal drugs may become affordable. We'll even discuss whether we will be able to download our brains into computers and live forever.

Walter Tichy
Associate Editor

The Rise of Computational Biology

An Interview with Prof. Thomas Lengauer
Max Planck Institute for Informatics
Saarbrücken, Germany

by Walter Tichy

Walter Tichy: You received a Ph.D. in computer science from Stanford University in 1979 and then worked at Bell Labs for a time. In 1984 you became a professor of computer science in Germany, studying combinatorial algorithms for circuit design. You are a dyed-in-the-wool computer scientist. When did you switch to computational biology, or bioinformatics, and why were you attracted to this area?

Thomas Lengauer: When I was young I wanted to become a physician. I found out before I finished high school that two things stood in the way of this. First, I shied away from the constant exposure to human suffering that I would have had as a medic. Second, I stumbled over mathematical formulas in practically everything I read. Thus I decided against medicine and started studying mathematics and later computer science. When it became clear in 1987 that the human genome would be sequenced, I realized biology and medicine would now become quantitative. So I decided to turn to computational biology to fulfill my original research vision. It took me five years until I was able to open a lab in this field in 1992.

WT: What were the major milestones in computational biology and how did you contribute to some of them?

TL: Since computational biology is such an interdisciplinary field you cannot separate computational advances from those in experimental methods. As an example, the major breakthrough in the '90s was the whole genome shotgun approach. Originally, the human genome sequencing project aimed at carefully cutting the human genome at predefined places, such that [the] stretches [that] resulted were short enough to be able to be sequenced. Since we would know where we cut the genome we could easily piece it together. However, the

required lab technology proved to be a major hurdle. Since the genome is very long—three billion letters—and the stretches to be sequenced needed to be short—one to 200 thousand letters—a complex hierarchical scheme for cutting had to be devised. Simultaneously, for shorter genomes, like those of bacteria with just a few million letters, one could follow a different approach: One blasts the genome by a global shredding process into a random collection of short "reads" that are sequenced and then pieced together by computer. The ingenious idea by Gene Myers and James Weber in the mid '90s was to do the same thing with the much more complex human genome. That required complex algorithms, which essentially transferred the problem from the lab to the computer. That technology is how we sequence genomes today.

Similarly, when people learned how to measure cell-wide profiles of gene expression levels in the late '90s, the demand for analyzing and interpreting the relevant data led to a surge in applied statistics techniques that we attribute to the field of computational biology. Bayesian methods were introduced into the field along the line, for instance for inferring gene-regulatory networks from gene expression data.

I think this character of the field will prevail. New experimental techniques continue to demand new computational approaches.

In my lab, two major advances stand out. Both of them are based on existing experimental techniques. First, in the mid '90s, we provided the first software, our program FlexX, that could dock a flexible small molecule, for instance a drug, into a rigid protein structure in a matter of minutes. This was a breakthrough, because before computing this would take days or longer. With our program you could scan through large sets of thousands of molecules, searching for promising candidates for drugs that could be further tested in the lab. This result was the basis for our company BioSolveIT GmbH that develops and sells software for medicinal chemistry labs and the pharmaceutical industry. The second breakthrough was the development of a statistical technique for estimating the resistance of HIV to drugs, which I think we will get to a bit later.

WT: Please sketch the algorithm involved in the shotgun sequencing approach.

TL: Whole genome shotgun assemblers are very complex but the basic principle is as follows: The genome is shredded into pieces of different but—up to some small variation—specific lengths, say 2,000, 10,000 and 150,000 letters, respectively. This is facilitated by a random destructive procedure such as sonication (tearing apart the chromosomes with sound waves). Each fragment is then sequenced from both ends to a length of a few hundred letters (paired-end reads). (Today's most frequently used sequencing technology does not allow for more, even though there are new developments that are expected to overcome this limitation.) The collection of reads covers each letter in the genome several times, the more the better. In the original human genome assembly, that multiplicity was about eight. Today one can go substantially higher to, say, 30 or even 100. Then graph algorithms based on special constructions like the overlap graph or the de Bruijn graph (which I will not dwell on here) are used for assembly. The fact that we know the (approximate) distances between a pair of paired-end reads enables us to build genome scaffolds that span repeat regions (stutter text) in the genome (see Fig. 1). Like almost all complex genomes, the human genome is full of repeats, which poses the main complication for genome assembly. Statistical analyses and simulation studies have led to appropriate settings of the relevant parameters such as the collection of fragment lengths and the sequencing depth.

Such assembly algorithms are for de-novo genome assembly, i.e., for assembling a genome without a pre-existing reference sequence. While we have the genome sequences of many species today (more than 40,000), there are still major challenges, for instance, sequencing certain plants that have among the largest and most complex genomes. Once you have a reference genome, assembling other related genomes, e.g., the genomes of human individuals, is much simpler. Here you just align the fragments form the genome to be sequenced to the reference sequence. The major issue is here to distinguish biological variation from technical sequencing errors

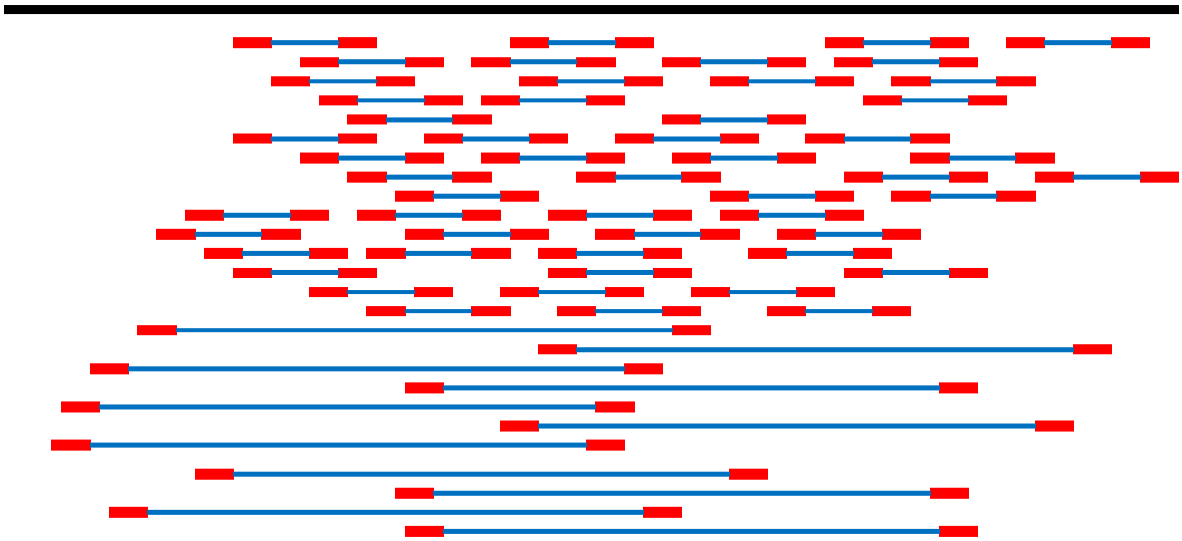

Figure 1: De-novo assembly of a genome sequence from read pairs originating from fragment libraries with two different fragment lengths. The black line at the top is the genome sequence to be assembled (unknown). The colored segments are the fragments. The sequenced ends of each fragment, which are also called "reads," are in red. The non-sequenced middle part is in blue. As can be seen by running a vertical scan line across the figure, each position in the genome sequence is contained in several fragments. In a real setting, every position in the genome sequence is covered by several reads, on average. The average number of reads covering a sequence position is called the coverage.

WT: How are variations detected? And how large are they from one human to the next?

TL: Genome variations take several forms. They can be changes of single letters in the genome sequence, so-called single nucleotide polymorphisms (SNPs). Two people typically differ in a single letter every thousand letters or so. That means they have a little less than 3 million different letters in their genomes. In comparison, the human and chimpanzee genomes are about 96 percent identical—still an amazing congruence. Then there can be deletions and insertions of whole stretches of genome sequences—of varying length—and also translocations of genomic sequences from one place to another in the genome. Life's organisms are highly optimized today, such that major changes in genome sequence are frequently lethal and thus not observed. Even single letter changes in genomic text—when they occur in critical places–can be lethal or confer critical diseases, as we know from the many monogenetic diseases

known in humans, such as sickle cell anemia, hemophilia, or cystic fibrosis. Human genome variants are detected by aligning the sequencing data coming from the individual to the now existing genome reference sequence. Here one has to be careful to distinguish read errors from the actual biological variations. Specific statistical methods have been developed for this purpose.

WT: Can you tell us how some of the other bioinformatics algorithms work, for example protein structure prediction?

TL: Protein structure prediction is a very good example of a bioinformatics algorithm. By now it is a fairly mature field. Principally, there are two ways of computationally predicting the three-dimensional structure of a protein from its amino-acid sequence. One is biophysical. We know an ensemble of protein molecules assumes structures that minimize the free energy of the whole ensemble. That means structures with high energy are rare and those with low energy are frequent. The exact statistical relationship is given by an equation that dates back to Boltzmann in the late 19th century. So the theory for computing protein structures is in place. However, exactly calculating the free energy of biochemical systems is outside the range of today's computing capabilities and to find its minimum is practically impossible.

Computational biology takes a different approach. The idea is to replace biophysics with evolutionary analysis. It turns out two proteins that are similar in sequence—say, at least 30 percent of the amino acids are identical in both protein sequences—are invariably related to each other evolutionarily and almost certainly have similar structure. Thus, if we want to know the structure of a protein we scan the complete database of proteins with known structure—more than 100,000 today—for a similar protein and model the structure after that so-called template protein structure. That will give us a good first approximation, which we then can refine with methods that minimize energy.

What if we do not find a similar protein with known structure? That is where the problem is not completely solved yet. But researchers have made great advances there, too. A group of computational biologists has devised a process by which one can find subtle evolutionary signals in large sets of related proteins—with at least 1000 members—none of which we know by structure, though. The basis of the method is there is a correlation between parts of the protein that are close in space. Such correlations can be detected if one has a sufficient number of diverse but related proteins. From these correlations one can deduce hypotheses about the

proximity of parts of the protein in space, which can then be used to put together a model of the protein structure. This method enables us to craft reasonable protein structure models in a growing number of cases.

This evolutionary approach is a major tool in computational biology. It is a pattern matching approach we use all the time in daily life as well. For example, when we reason about other people, their state of mind, mood, potential next actions, etc., we reason not from theory but from data. It is what we call "experience."

WT: There is talk about noncoding "junk" DNA. What is this, how much of it is there in the human genome, and does it have a function?

TL: It has become evident early on that [there is] only a very little of the whole genomic text codes for proteins, a bit above one percent. The function of the rest of the human genome has remained pretty much in the dark for a long time, even though one knew certain noncoding regions had to be involved in regulating gene transcription—the synthesis of an RNA copy of the gene, which is the first step towards making a protein. But then there was the stutter text in the genome, which makes up almost 50 percent of the genome. Apparently that text originated from self-replicating parts of the genome, which are partly of viral origin and to which one could not attribute any sensible biological function. So it was called junk DNA. In the last two decades, two central developments have happened. First, the human genome sequence became available, and with it the complement of all genes. Second, the multinational ENCODE project, which took about 10 years to identify all regions of the genome that are transcribed into RNA, found out more than three quarters of the genome are actually transcribed to RNA in some tissues. Many transcripts never make it into protein, but they are there as RNA and many of them (if not all) have a biological function. So the term junk DNA was clearly a misnomer and I have not heard it much lately. Already in the late '90s people found out RNA is a prime player in cell regulation—arguably more important than proteins. This finding was deemed so important that it received the Nobel Prize in 2006. Much of what the RNA does in the cell is not resolved yet, however. This is a focus of research in biology today.

WT: How are genes identified?

TL: As we have just seen, the notion of a gene has undergone a substantial transformation in the last two decades. Let us stick to protein-coding genes. I think one can consider this problem solved today. Identifying the gene means determining the protein-coding region and the entire sequence-encoded infrastructure the gene carries with it for steering its own transcription and regulation. This infrastructure consists of noncoding regions belonging to the gene. If the inspected sequence resembles that of a known gene, then one can identify the regions by aligning the two sequences. For this purpose one just scans through a comprehensive database of gene sequences for a similar sequence. If one has identified that sequence as a gene, one can map its infrastructure onto the gene under study and make succeeding refinements and corrections. If there is no known gene that is similar to the sequence under study, then one uses a statistical approach based on the fact that the different gene regions exhibit different characteristic sequence motifs. In recent times, the experimental method called RNA-seq, for sequencing RNA transcripts, gives us access to effectively all transcripts in a cell mixture. If one has the transcript one can make the connection to the genome by aligning the RNA sequence of the transcript with the genome sequence and thus locate the respective gene. As a further significant step in scientific progress, single-cell sequencing is in an advanced state of development.

WT: You made a major contribution to HIV therapy, please explain. Have people survived AIDS because of your work?

TL: Yes, I think one can rightfully say that, even though I myself cannot point to specific persons, because I do not meet the patients who are treated with our geno2pheno software. The problem with HIV is that the virus changes its genome rapidly. Each HIV patient harbors a unique cocktail of viruses, and that cocktail changes in response to the drugs applied. After some time, the viruses become resistant and the therapy fails. Our software is based on a large collection of data from clinics and labs throughout Europe on how this resistance develops. Data records comprise the genome sequence of the relevant HI virus, the applied drugs, information on the success or failure of the therapy, and some clinical measurements or, in a subset of the cases, a lab value measuring the level of resistance of the virus to drug applied *in vitro*. From the data we build statistical models that infer the level of resistance of the virus to the drugs that one might want to apply next. This information is given to the physician, who then puts together a new drug combination for the patient. It turns out using such a scheme is superior to selecting drugs by hand, especially for therapy-experienced patients whose virus

has acquired many resistance mutations. There is even a type of resistance HIV can acquire, where a statistical model is indispensible because there is simply nothing to be seen in the viral genome with the naked eye. The relevant statistical model is our blockbuster—a slight misnomer, because we offer our services at no cost. The technology we brought to the scene has helped advancing HIV therapy to what I like to call the "spearhead of personalized medicine." In this well-delineated field therapies are routinely tailored specifically to each patient with computer help.

By the way, significant advances in the field are frequently made in collaborations. In our case we are embedded in effective consortia at the German and European levels, comprising virologists, clinicians and computational biologists.

WT: What are the major challenges ahead for computational biology?

TL: In the '90s we considered genomes as something that pertains to species: "the mouse genome," "the human genome," etc. Our basic object of study was the genome sequence. We were busy assembling the building blocks—finding genes, translating them into protein sequences, and predicting protein structures. In the late '90s, with the advent of the technology for measuring gene expression profiles (transcriptomics, as we say today) the attention turned to the fact that while all our cells share the same genome, they differ in the gene complement they use. Those differences define tissues and mark the difference between healthy and diseased. At the same time, the interactions between the building blocks came into focus. We now started studying networks of interactions of genes and proteins in order to bridge the immense gulf between the cryptic genome sequence and the observed biological function. The focus on cell regulation has turned into what today is epigenomics—basically the science of exactly what happens in the cell nucleus. We are in the middle of this adventure, still assembling building blocks and elucidating the first interactions. We are far away from a global picture of the processes in the nucleus. That is one of the major quests I see, and we will only have mastered it once we have a computer simulation of the processes in the nucleus that can be configured with external parameters that will enable us to analyze the origins and progression of diseases. Of course, the nucleus is not everything. There is the cytosol, the other part of the cell, as well. The cytosol is the production line of the cell where proteins (proteomics) are synthesized and metabolism (metabolomics) and communication (signaling networks, interactomics) occur. At all levels we can now gather cell-wide data whose integrated analysis is a great challenge for computational biology. A further line of research that was

started early this century is based on the realization that, while we are all human, we do not have the same genome. The slight variations between our genomes—we are only different in roughly one in every thousand letters—are the basis of our individuality. They make us who we are, with all talents, shortcomings, predispositions etc. Especially in the context of diseases, studying genomic variation has become a major focus. It will lead us into the age of personalized medicine.

WT: Cell regulation seems to be a hellishly complicated subject. Can you please share an example of what has been discovered so far?

TL: This is another large field, which is addressed with terms like regulatory genomics and epigenomics. Regulatory genomics is the more general term that addresses all aspects of cell regulation. Part of it is, for instance, about what proteins bind when and where to the DNA , which comprises the infrastructure of a gene, in order to effect or inhibit transcription, duplication, and the like. Epigenomics is about how the cell nucleus is organized. It turns out the DNA, with a length of about two meters, is not simply stuffed into a volume of 1µm diameter, but it is chemically labeled and exquisitely packed into a structure made up of proteins. The whole complex is called chromatin. The relevant sequence information is called the epigenome. An epigenome is much more complex than a genome since it comprises not only the DNA sequence but also includes from half-a-dozen up to theoretically 200 different chemical labels. These labels influence the packing, and the packing determines which genes are accessible for transcription and which ones are tucked away, since the cell does not and should not use them. This is a highly dynamic process and it is specific to each tissue. Anything going wrong here means disease. We are participating in substantial German, European, and global consortia to sequence and interpret reference epigenomes from different tissues, healthy and diseased. The global research program spans a decade and we have just entered the second half. We are analyzing the putative epigenetic basis of diseases such as cancer, metabolic diseases, and inflammation. The research is partly directed at obtaining an overview, e.g., how is the chromatin organized in the large, and partly at putting together the parts list, e.g., what are specific epigenetic regulators and what is their function in specific settings?

WT: One hears a lot about the microbiome lately. What is this, what is its significance, and how does one compute it?

TL: The human organism is colonized by a very large number of microorganisms. One estimates that 10 times as many microbial cells colonize the body as the body itself has cells. The microbes have co-evolved with the human and both partners actually form a unified living system. Of interest here are mostly bacteria, followed by viruses and fungi. Sometimes such microorganisms enter a symbiosis with the human, such as is the case for bacteria in our intestine that help us digest our food and are essential for our survival. In other cases, microbes can cause diseases. If we want to understand the body in health and disease we need to take these microbes into account. They are suspected, or have even been proved to be involved, in diseases like inflammation—as in the case of certain gut bacteria, which play a role in Crohn's disease, a serious chronic inflammation of the gut—and cancer—such as the human papilloma virus, which can cause cervical cancer. The microbiome differs markedly in different environments in our body such as nose, mouth, colon, skin and genital tract—and it differs in association with lifestyle, e.g. diet, as well as in association with disease. By taking samples from the relevant bodily environment and exercising sequencing technology one can analyze the microbiome. This field is called metagenomics. The difference between a metagenome and a "normal" genome is that the metagenome is comprised of a mixture of many unknown species, e.g., bacteria that cannot even be isolated or cultivated. It is a great and yet unsolved challenge to identify bacterial genomes from such mixtures, but most often it is not even necessary—grouping the partial genome sequences identified into clades (groups of evolutionary related organisms) affords a metagenomic profile that is helpful for diagnosis, therapy, and studies of factors causative of disease. The respective computational challenges involve classifying bacterial genomes from microbial data—a process called taxonomic binning—and elucidating the relationship between the microbial profile and phenotypic traits such as disease.

WT: Will the promise of computational biology come true? What are the prospects for curing cancer, diabetes, and personalized medicine?

TL: It has been just about 20 years since biology has turned into a quantitative discipline, giving us the chance of understanding living organisms at the molecular level at a larger scale. We can now search systematically for the cause of diseases. We have just stuck our head out of the water. I am not in a position to predict what progress will be made in medicine and when it will happen. I fear press releases are generally overoptimistic. But now we have the chance of digging through this large mountain towards the promised land of personalized effective cures.

Progress will not be revolutionary, but come in small steps—focused insights on very specific diseases, which will eventually open up the path to new cures. As we have seen before, all of this will require lots of data; data that we as individuals need to be willing to offer to science. In turn, science must keep the promise of careful handling of these data. The time is ripe now for us to embark on this search and we have embarked on it.

WT: The problem with personalized medicine is that it is extremely expensive, simply because a therapy costs millions to be developed, but can be applied to only a small group of patients or a single person. Do you see any way around this problem?

TL: It is a ubiquitous trend that as medicine is developing it becomes more and more expensive. Up to now the major cost increase was attributed to expensive technology, for instance, for surgical gear or imaging apparatus. I cannot predict what will happen as we enter into personalized medicine. However, here I do not regard cost explosion as a given. Doing the measurements, employing the omics technologies, is going to be cheap. What about the small patient groups? Take the HIV example. The multitude of therapy options, which determine the reduction in the number of patients receiving the same therapy, does not result from an equally large number of drugs. In fact, about 30 drugs lead to more than a thousand viable therapies. While the development of each drug costs many million dollars, combining the drugs is not subject to substantial additional cost, neither for development nor for certification. I expect the same thing to happen for cancer in the near future.

WT: Raymond Kurzweil and others think the technical singularity will allow us to download our brains into computers and live on inside computers. What do you think about extra-biological life?

TL: This question really transcends computational biology, because this field is about using informatics technology to understand life, not about creating life with computers. I will try to answer, nevertheless. There are two separate issues. First: Will we be able to download our brains into hardware and will this be something we want to do? My answer is clearly "No." Neurologists have increasingly come to realize that the brain and the body are exquisitely and inextricably intertwined. Much of our brain is designed to provide maps of our body. The brain originally is a tool for survival. Its ultimate purpose is to guide the behavior of the physical individual towards a state of well-being. This purpose has been grounded very early in the

history of life. It goes back even to organisms that do not have a brain. In order to fulfill this purpose the brain has access to myriads of sensors, a much larger variety than we are consciously aware of. Severing the brain from the body would mean depriving it both of all these data and of its prime purpose. We would not only have to download the brain into a computer but the body, as well. Otherwise our notion of well-being would be completely compromised. I do not foresee such a state nor do I see any sense in aspiring to it.

The second issue: Can we create extra-biological life? That depends on how you define "life" and how you define "extra-biological." Researchers in the recently emerged field of synthetic biology are engineering systems by using building blocks of life or by substantially modifying evolved cells, either for scientific reasons or for engineering purposes. Depending on the approach, I would say you obtain either an engineered system that does not live or you modify a living organism. The latter we have done for several decades via gene technology and metabolic engineering and previously for thousands of years via breeding. These procedures have manifested a fundamental shift in the development of life by supplementing evolution with rational design. However, in my view they do not transcend biological life.

Creating life with no reference to life as it has evolved on earth is a completely different issue that rests on what we mean by "life." Usually life is defined by criteria involving replication, energy consumption, metabolism, movement, communication, and others. Related criteria may be able to be fulfilled in a virtual world inside the computer leading to a transformed and arguable notion of life. I find it harder to imagine a physical implementation of artificial life with no reference to biological life or to see any point it.

WT: One last question. If a student wants to go into bioinformatics, what and where should he/she study?

TL: There are three options: (1) Start with informatics and then specialize in bioinformatics, (2) start in biology and then specialize in bioinformatics, or (3) study bioinformatics right from the start. There are advocates for all three approaches. I strongly prefer approach no. 3. Informatics and biology are very disparate subjects. Informatics is an engineering discipline, which is characterized by rational design. Biology is ruled by evolution, which, in some sense, is the opposite of rational design. A good bioinformatician needs to epitomize both approaches in a balanced fashion. Where can you study bioinformatics? Of course there are many good universities today. The field has taken off substantially since the turn of the millennium. May I

take the liberty of mentioning here that starting at the end of the last millennium, Germany has made an especially strong national push of developing integrated bioinformatics curricula. Today there are quite a few places in Germany that offer well developed interdisciplinary curricula in this field, for instance in Berlin, Bielefeld, Munich, Leipzig, Saarbrücken, and Tübingen. Germany has been commended world-wide for the success of this initiative and the quality of the people that have graduated in the field and have spread across the globe.

Suggested Readings

Lengauer, T., A. Altmann, A. Thielen, and Kaiser, R. Chasing the AIDS virus. *Commun. ACM* **53**, 3 (2010), 66-74.

Lengauer, T., M. Albrecht, and Domingues, F. S. Computational Biology. In *Systems Biology*. Ed. R. A. Meyers. Wiley-VCH**,** Heidelberg, Germany, 2012, 277–348.

Lesk, A. Introduction to Bioinformatics, 4th ed. Oxford University Press, Oxford, 2013.

About the Author

Walter Tichy has been professor of Computer Science at Karlsruhe Institute of Technology (formerly University Karlsruhe), Germany, since 1986. His major interests are software engineering and parallel computing. You can read more about him at www.ipd.uka.de/Tichy.

DOI: 10.1145/2892034