

Processors Piece Together the Book of Life

The book of life is 44 pages long – at least, the chapter on Stephan Schuster's desk is: 44 single-spaced pages, really just a pile of paper. Nonetheless: "I got goose pimples when I held it in my hands for the first time", says Schuster. "These pages allow you to read an organism just as you would read an open book".

Mind you, what is written on the pieces of paper could hardly look less impressive. The content of the pages looks rather like the inventory list of a chemicals factory. Indeed, this initial impression is not so far from the truth: the pages contain a list of all the proteins, enzymes and membrane proteins offered up by the bacterium *Wolinella succinogenes* through its different life phases – derived solely from a full analysis of the organism's genetic material. These pages indicate just how far biology has come in its understanding of the living world. Schuster: "Just a few decades back, biologists had to infer the inner workings of a living creature – its genotype – from its external appearance – the phenotype. Now we are on the brink of being able to do it the other way around: soon we will be able to infer the characteristics of an organism from its genetic blueprint."

Who is Stephan Schuster? Born in 1962, the young scientist first studied chemistry at the Technical University in Munich and later in Konstanz, cultivating a strong interest from the very start in the work of his

colleagues from the biology department. After completing his dissertation on an aspect of heterocyclic chemistry, he decided to put his knowledge of chemistry at the service of a biochemical study group.

THE LABORATORY AS A SOURCE OF DATA

His switch to the life sciences was soon complete: his doctoral thesis ("Studies on the flagellar apparatus of *Wolinella succinogenes*") was written under the supervision of Edmund Bäuerlein in Martinsried and, after a period of post-doc work at Caltech, he worked on the genome of a halobacterium alongside Dieter Oesterhelt in the Department for Membrane Biochemistry at the Max Planck Institute of Biochemistry in Martinsried. For two years now Schuster has headed up his own study group ("Genomics and Signal Transduction") at the Max Planck Institute for Developmental Biology in Tübingen, and when he takes visitors on a tour of his colleagues' laboratories, the enthusiasm this dynamic young researcher feels for his work with bacterial cultures and sequencing quickly becomes apparent.

"Ultimately, though, the data we're generating in the laboratory is only something we need for our real research", says Schuster. On a computer screen in his office a window appears, the left half of which is taken up with a bewildering block of letters: an excerpt from the genetic code of a bacterium currently being



Stephan Schuster, leader of the Genomics and Signal Transduction study group at the Max Planck Institute for Developmental Biology in Tübingen: "In five to ten years we will be able for the first time to find out the physiology of the first organisms almost entirely from knowledge of their genomes alone. But these predictions will still need to be verified by biochemical experimentation for a long time to come."

analysed in the laboratory next door. The book of life, chapter one. Stephan Schuster is an expert in bioinformatics.

"There is an astonishing number of parallels between the genetic and digital codes that encrypt the information contained in living beings and computers respectively", says

Without high-performance computers it would be impossible to sequence the complete genome of even the simplest organisms. Supercomputers compare short snippets of DNA from the sequencer and assemble the information thus obtained into complete genomes. But when it comes to correctly ascertaining the function of the gene that has been discovered, an expert's experience beats the computational power of parallel computers – as work carried out by **DR. STEPHAN SCHUSTER** at the **MAX PLANCK INSTITUTE FOR DEVELOPMENTAL BIOLOGY** in Tübingen shows.

the scientist. "A base pair in a DNA molecule – in other words, a letter of the genetic code – takes up one byte of space on a computer. That means that the genome of a simple bacterium – around 1,600 genes to about 1,000 base pairs – takes up about 1.6 megabytes on our databases."

Biological computers in the Petri

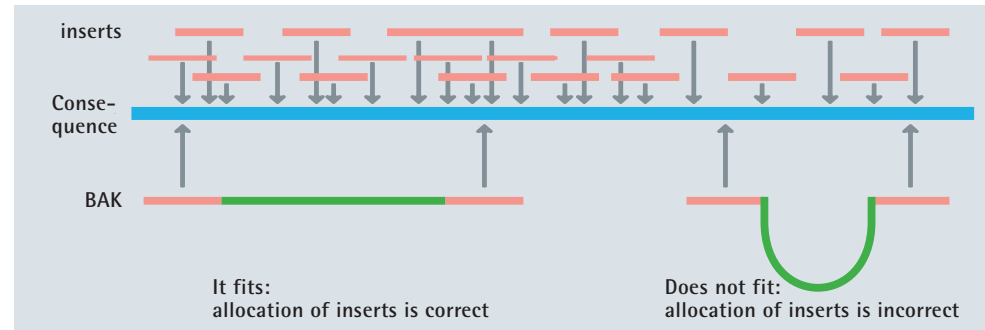
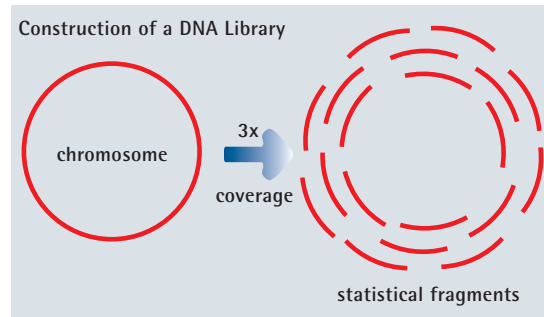
dish – now we've reached the heart of the matter. For if there were no supercomputers working solidly in parallel, like *E. coli*, Schuster's job would be a lot harder, and even impossible in the future.

This is because it takes an incredible amount of effort to be able to read the genome of an organism like

the open book lying on Schuster's desk. Piecing together the 2.1 million base pairs located on the chromosome of even simple organisms like *Wolinella succinogenes* is not exactly a stroll in the park even now, after the dazzling success of the Human Genome Project. The genes are simply too large to be decoded base by base, letter by letter. So they must first be cut up into smaller segments that can be analysed more easily. The problem comes afterwards in correctly piecing together the information coded on these snippets of DNA. And that cannot be done without a computer.

Schuster uses the "whole genome shotgun" method in his work. His current research is focused once again on *Wolinella succinogenes* – "Woli" in lab jargon. Woli can be thought of as the nice relative of the bacterium *Helicobacter pylori*, which causes stomach ulcers: the same family, but not pathogenic. In order to shed light on the *Wolinella* genome, Schuster and his colleagues break down the bacterium's chromosome into countless fragments, known as "inserts", in a centrifuge and smuggle these into *E. coli* bacteria, which make copies of the *Wolinella* genome fragments. As soon as the *Coli* bacteria have produced enough copies of these incomer genetic fragments in the incubator, the inserts that have been copied are "harvested" and copied again by means of a procedure known as cycle sequencing, which is

The sequencing of a complete genome using the "whole genome shotgun" method involves breaking a chromosome down into tens of thousands of fragments. These fragments are copied and analysed – fast computers are needed to assemble the information thus generated.



BACs solve difficulties that arise in correctly ordering the snippets of DNA. If the computer can fit them in, so that about 200,000 genetic letters are located in the consensus strand between their ends, this indicates that the way the computer arranged the inserts located in between was correct.

very similar to polymerase chain reaction. This is done using special building blocks that enable a further sequencer to determine the order of the bases of the original, tiny fragment of the *Wolinella* genome. The procedure may appear less elegant than the "primer walking" method used previously, in which the genome is likewise cut into pieces but the pieces are examined one at a time in the right order. However, it is incredibly efficient: it thrives on statistics and a high throughput.

Every day about 1,000 gene fragments are generated, cleaned and "sequenced" in Tübingen. Schuster and his team produce so many inserts in total that they would fit over the top of the complete genome eight times over. This lowers the potential for reading errors. If enough of these short fragments are generated, so the theory goes, then at some point every part of the gene will have gone through the analysers at least once. What you then have is a mountain of readable jig-

saw pieces, which – once they are put together correctly – produce the *Wolinella* genome.

A JOB FOR FAST SUPERCOMPUTERS

The researchers need a computer to piece together this mass of genetic material. "Of course, the random disintegration of the genome at the beginning generates fragments that overlap partially with one another in terms of their information content. It is then down to the computer to recognize inserts that have some of the same base sequences and to use these overlapping portions to assemble them into larger sequences, so-called contigs", says Stephan Schuster.

This job is tailor-made for fast supercomputers working in parallel, whose processors share the job of scrutinizing the digitalised genetic fragments. The computers have a sufficiently large working memory to store the base sequences of all sequenced inserts and contigs in their memory – and they also have sufficiently fast data links to enable them to deal quickly enough with the extremely high volume of data travelling between processor and memory. Even though personal computers now have very fast processors, they would not be up to the job, especial-

ly in these disciplines. This is why there is a parallel computer named DARWIN currently working on the *Wolinella* genome: it has eight processors, a 32-gigabyte main memory, and 800-gigabyte hard disk capacity. DARWIN was created partly by Stephan Schuster and his head of computer science Günter Raddatz, and partly by an entire researcher consortium, consisting of the Garching supercomputer specialist Hermann Lederer and his colleague Andreas Schott as well as Dieter Oesterhelt and his colleague Friedhelm Pfeiffer from the Department for Membrane Biochemistry at the Max Planck Institute of Biochemistry in Martinsried, and Folker Meyer from the University of Bielefeld. This team compiled the necessary programme packages without which efficient analysis and classification of genomes would not be possible – and undertook the laborious task of ensuring that the programmes worked together properly. The *Wolinella* genome project represents the first time DARWIN has been put to the test in a serious way.

But even though a computer has taken on the hard graft of piecing together the large Tübingen genome puzzle, human activity is still required – such as when it comes to filling in the gaps in the genome. As in all organisms, there are long segments in the *Wolinella* code that consist merely of repetitions of the very same base sequence, so-called repeats. How can we find out which of the inserts with sequences ending in a repeat are located next to one another? The danger here is that the computer sticks these pieces together incorrectly – with fatal consequences: if the sequence of the genes on the genome is muddled up, important information gets lost because genes that are linked to one another functionally are often also

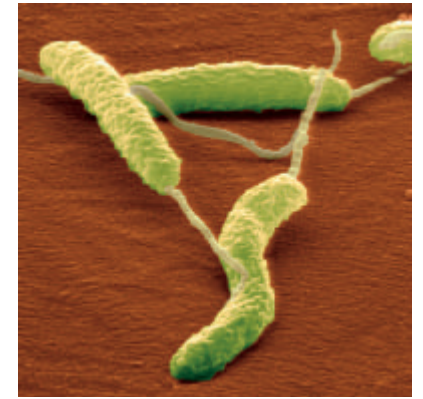
situated close together on the genome – in so-called operons.

In order to solve this problem, Schuster and his team resort to a trick. They cut up the *Wolinella* genome yet again, only this time not randomly but in a particular way, into segments roughly 200,000 base pairs long, which even the best sequencers are not capable of analysing completely. These so-called BACs (bacterial artificial chromosomes) are also copied using *E. coli*. Schuster's colleagues then identify just the base sequences at the ends of these long chains, as few as 1,000 base pairs or so in each case – in other words, as many as possible.

ARTIFICIAL CHROMOSOMES VERSUS JUNK DNA

The punch line is still to come: Schuster has the computer fit the sequences found on the end pieces into the probable genome. If both end pieces are indeed located about 200,000 base pairs apart in the computer's genome model, the "consensus sequence", then it can be assumed that the genes the computer suspects are located in between are indeed in the correct place, in spite of the repeats contained in them. If the long BAC does not fit, then something is not right – and the genome has to be re-assembled at this point.

However, in spite of the apparent dazzling simplicity of this procedure, it is by no means trivial: one nanogram of a BAC library contains 100 times fewer molecules than a typical insert library; Schuster's knowledge of chemistry really comes into its own in the fiddly analysis of these impossibly small amounts of material. Thus the fact that his trusty computer deals with a large amount of work on the puzzle does not mean that humans are superfluous in Schuster's laboratory – on the con-



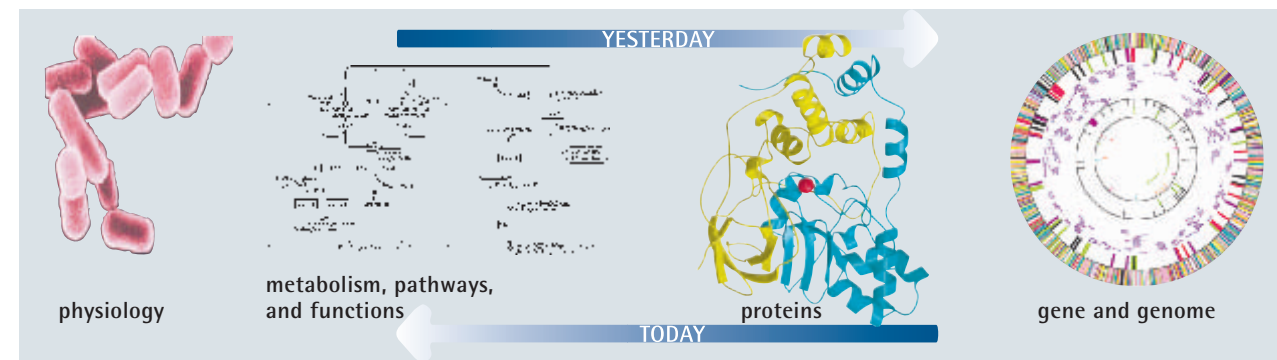
Wolinella succinogenes is a kind of "tame" relative of the bacterium *Helicobacter pylori* which causes stomach ulcers. The *Helicobacter* genome has already been sequenced, while the analysis of the *Wolinella* genome is just being completed in Tübingen. Direct comparison of both genomes may provide valuable insights as to the cause of the pathogenic qualities of *Helicobacter pylori*.



Without a supercomputer Schuster's work would be inconceivable. Nonetheless: "We need gifted colleagues more than ever now who can help make sense of the results produced by the computer."

trary, in fact: "We need gifted colleagues more than ever now who can help make sense of the results produced by the computer", says Schuster. "Even after the success of the Human Genome Project genome sequencing at the touch of a button is still a distant dream."

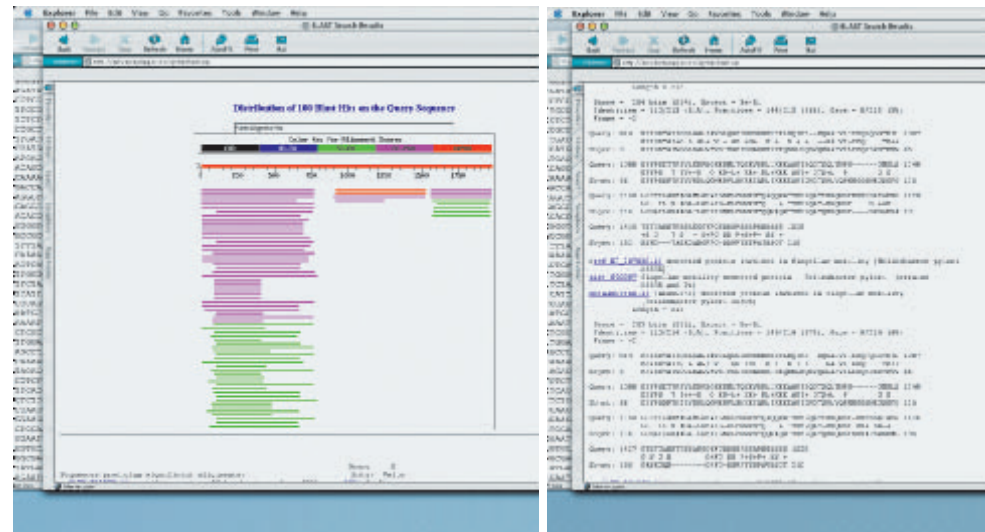
The *Wolinella* genome has now been almost fully sequenced. However, the really hard work only got underway a short while ago, when Schuster and his colleagues prepared to move their "information processing from the "wet lab" into the "dry lab", in other words, to analyse the raw data supplied by the sequencers. This, too, is now done at computer terminals, but they have to be oper-



Just a few years ago it was necessary to begin from the phenotype, that is, the external appearance of an organism, in order to infer its internal genetic make-up; today, bioinformatics are gradually enabling the opposite to happen: in the case of bacteria, for example, genetic maps can provide information leading to completely new antibiotics.

PHOTOS: MPI FOR DEVELOPMENTAL BIOLOGY (1) / WOLFGANG FISLER (1)

FIGURES: ROHRER ACCORDING TO DOCUMENTS PROVIDED BY STEPHAN SCHUSTER, MPI FOR DEVELOPMENTAL BIOLOGY



Biochemists nowadays make use of parallel computers for matching up genes to genetic products such as enzymes. In order to find known proteins for which an already sequenced gene might code, databases containing up to 800,000 entries have to be searched. Evaluating the best fit is the preserve of the expert, however.

ated by people who are able to make sense of the almost endless sequence of letters in the computer's memory.

For example, even the apparently simple question of where a gene starts and where it finishes cannot be settled in every instance by a computer alone. True, nature does use universal starting codes, for example the base sequence ATG, which acts like a sign: new gene starts here. But the gene fragments from the "field" can in principle be read off in both directions. So is it ATG or GTA? Another problem is, how can the computer know whether it is ATGTG, or ATGTG, or ATGTG? Thus, for every snippet of DNA there are six different possible ways of reading it. Only an expert can decide which is the right one – even if they are helped by efficient software.

The matter becomes even more complex when it comes to matching up sections of the genome assembled by the computer to gene products, that is, to particular tasks. Over the past few decades an unbelievable amount of information about proteins has been amassed, the blueprints to which are ultimately contained in DNA. This immense pool of

knowledge is at once a blessing and a curse for Schuster and his colleagues.

REAL DETECTIVE WORK AT THE MONITOR

This becomes clear when Stephan Schuster sits down in front of his computer screen, on which is displayed a section of the endless sequence of letters that make up the base sequence of the *Wolinella* genome. Schuster marks out a portion about 2,000 ACGT letters long. "What we're going to do now is key in a query to the database", he says. The series of letters goes via a fast internet link to a database that is likewise installed on the DARWIN computer in Garching. It contains the blueprints for about 800,000 enzymes, membrane proteins and other kinds of proteins which biochemists have compiled laboriously over the past few decades: 386 megabytes of amino acid sequences. The corresponding DNA database is even larger (3.6 GB), since one codon for an amino acid consists of three nucleotides at the level of DNA and furthermore contains a larger number of non-coding segments.

Within seconds the names of a series of enzymes from other organisms appear on the screen, whose corresponding DNA base sequence shows similarities to the one stored there before. "According to the data-

base, the sequence I have marked out here is highly likely to be the gene for methyltransferase", says Schuster. "But then again it might be that this protein actually fulfils a completely different task. Only an expert can tell. It's real detective work."

Nonetheless, with the help of all the snippets of DNA shuffled together by the computer as well as virtually inexhaustible databases, Schuster's team has pieced together its "*Wolinella inventory*". If this list is compared with the corresponding one for the pathogenic *Helicobacter pylori* related to Woli, it reveals some astonishing insights into the most intimate aspects of these two organisms' relatedness.

But even this 1-to-1 comparison of two living organisms turns out to be predictably mundane: here again, nothing but a list with the *Helicobacter* enzymes on the left and the corresponding *Wolinella* gene products on the right. But the entries are a tough proposition. Schuster again: "*Helicobacter* has a whole series of genes that we have not been able to match up to a known enzyme so far – an entire operon right next to genes whose pathogenic characteristics are already known. Although closely related, *Wolinella* does not have these genes. It is very likely that this gene complex is in part responsible for the pathogenic characteristics of *Helicobacter pylori*." A hot tip for pharmacologists? Much more! "With the decoded blueprint in our hands we can basically characterize a bacterium without ever having seen it. We can tell what it feeds on and probably even how it can be combated. Just imagine: you could develop pharmaceutical targets for completely new antibiotics just from knowing the genetic code!"

And then? Could this information be used at some point to design new organisms from scratch? "Maybe.

After the genetics era comes the era of chemistry. The task facing us now is to gain a detailed understanding of physiological processes and to model them. In five to ten years we will be able for the first time to find out the physiology of the first organisms almost entirely from knowledge of their genomes alone. But these predictions will still need to be verified by biochemical experimentation for a long time to come."

FROM BACTERIA TO ZEBRAFISH

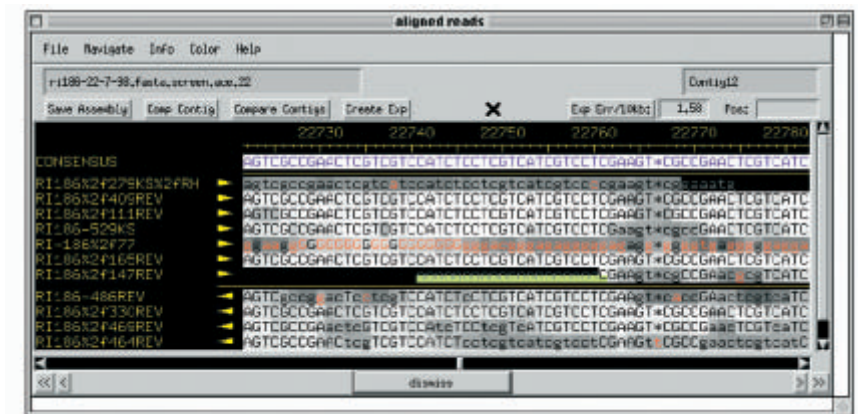
The first "vitreous" organisms are certain to be bacteria, which, even given their own complexity, are still woven of relatively simple cloth. Unlike in the case of bacteria, the detailed unravelling of the extremely complex biochemical processes that determine the growth and appearance of higher creatures is still in its infancy. In order to understand them better, it is necessary first to examine living things that are simple enough to study, and complex enough so that conclusions may be drawn from what has been discovered – about hereditary diseases in humans, for example.

One creature especially suited for this kind of research is the zebrafish. Alongside the fruit fly *Drosophila melanogaster*, it is to the developmental biologists working alongside Christiane Nüsslein-Volhard what *Escherichia coli* is to molecular geneticists and what *Arabidopsis* is to plant geneticists: literally a universal "working animal" that has provided a lot of information about the development of vertebrates from the egg through to the fully grown individual. For example, for every form of congenital heart defect in humans it has been possible to generate a corresponding zebrafish mutation. Still, many of the genes that are partly responsible for these de-

fects are not known, and the search for them is difficult and time-consuming.

It will be easier to reveal them once the zebrafish genome has been fully sequenced. An international team of researchers led by the British Sanger Centre has therefore set itself this ambitious goal; the Max Planck Society and the German Research Society are also participating in this project. The method involving the use of BAC libraries, cultivated by Stephan Schuster, is almost certain to come into its own here, as the problems hinted at in the sequencing of bacterial genomes are multiplied several times over in the case of the much more complex vertebrates – a lot of repeats, redundant junk DNA. "In total we will be analysing about 200,000 BAC end sequences", says Schuster, "and this will ensure that the inserts on the zebrafish genome analysed by the Sanger Centre are a good fit".

But that is not all. Once the BACs are available, some of them will be broken down and sequenced again individually using the shotgun method – this means that the zebrafish researchers are involved in a total of around 22,000 separate sequencing projects. The project will generate an enormous data set that would cover the zebrafish genome about eight times over.



The computational power needed to deal with this mountain of data is huge. The Garching computer centre already has another supercomputer for the project up its sleeve: a digital bolide with 6 x 32 processors and a 6 x 96 gigabyte main memory – and that is just the first phase of expansion.

So we can look forward with anticipation to the moment when the Tübingen researchers are able to put the "book of life" of the zebrafish on their desks. But even if it takes a little longer yet: the first successes have already kicked in. The BACs generated by Schuster and his colleagues can also be used as markers for "conventional" genetic experiments with the zebrafish. Using one of these BACs, a group working with Teresa Nicolsen in Tübingen has recently been able to localise the precise position of a genetic defect that leads to a disorder of the sense of balance in zebrafish. In humans and mice, a mutation in the same gene leads to muteness. So even the "proof" of the book of life of the zebrafish contains some promising surprises.

STEFAN ALBUS

Computer programmes help bioinformatics specialists to piece together the base sequences of fragments obtained by breaking down a complete genome, known as "inserts", into a correct gene sequence, the "Consensus", once they have been analysed. "Definite" bases are printed in large letters, "non-definite" ones in small letters.

PHOTOS: MPI FOR DEVELOPMENTAL BIOLOGY