BIOINFORMATICS

Vol. 16 no. 1 2000
Pages 48–64

Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times!

Amos Bairoch

Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland; E-mail: bairoch@cmu.unige.ch

Introduction

This is a personal recollection of the events that led me to develop software tools and databases in the context of what has recently been termed proteomatics (bioinformatics in the context of proteomics). As will be manifest from this article, the creations of PC/Gene, SWISS-PROT, PROSITE and ExPASy, were mostly serendipitous unplanned events. From the very beginning of my biochemistry studies in 1978 up to today, I was extremely lucky to be able to pursue my combined interests in proteins and computer analysis and to be able to follow new avenues when they opened up. I also feel privileged to have met and collaborated with many researchers whose work has been instrumental in the emergence of the field of bioinformatics. More significantly many of these people became much more than colleagues. They are friends scattered around the world, united by a common passion, uncovering the meaning of the genetic information. It is to these friends that I dedicate this article.

The early days

As a teenager, I was interested in space exploration and the search for extra terrestrial life. After graduation from high school in 1977 I thought that studying for a university degree in biochemistry was a good way to train to be an exobiologist! In high school I had also become acquainted with computers. We had access to a time-sharing Honeywell Bull mainframe system that we could program in FORTRAN using a teletype console. We could also use a Wang series 700 programmable calculator that had a capacity of about 1000 programming steps. It was a wonderful feeling to be able to program a machine that was not dependent on a large computing center and one that I could fully exploit without time constraints.

In 1978, during my first year at university, my father bought me a Radio Shack TRS-80 microcomputer. With it I wrote game programs using the first versions of the Microsoft Basic programming language. During the summer I worked in a Lausanne medical research laboratory that was developing radio-immunoassays. This led me to write programs to analyze the results of this type of experiment. When I returned to Geneva, I contacted the Clinical Biochemistry Institute (IBC) of the University of Geneva which, I thought, might be interested in these programs. I was very lucky! The day I visited the IBC I was introduced to a visiting professor from Oxford, Robin Offord. Robin Offord, a biochemist and previously a nuclear physicist, is an expert in the semi-synthesis of proteins. In addition Robin has a long lasting interest and knowledge of computer systems that started with Mercury Autocode in 1959. Robin convinced the head of the IBC, Albert Renold, to hire me part-time to write data analysis programs. He asked me what computer I thought the lab should buy. I indicated that a small Californian company had recently announced a new microcomputer that seemed to correspond to the needs of the lab. I got the green light to buy what was the first Apple II microcomputer in Geneva. From 1979 to the beginning of 1981 I pursued my studies to obtain my bachelor degree in biochemistry and wrote programs for the IBC during my spare time.

In 1980 Robin Offord moved from Oxford to Geneva to head the Department of Medical Biochemistry in the medical faculty. When I obtained my BSc in 1981, I started to look for a lab where I could do a Master's degree that would allow me to do both 'wet' (laboratory benchwork) and 'dry' (computer) work. Robin suggested that I do my thesis in his department under the supervision of Keith Rose. Keith is an expert in mass spectrometry and he had just received a brand new machine from Kratos. Unfortunately it was not working according to the specifications and a Kratos engineer was spending most of his time in Geneva troubleshooting the machine. As I waited to start the experimental part of my master's thesis, I decided to tackle the computer system that was running the mass spectrometer. For that time, and especially for a young student such as myself this was an impressive Digital Research NOVA-4 minicomputer with a 5 Mb removable hard disk cartridge. Furthermore it was hooked up to a HP graphic terminal. We hoped that when we did manage to get the mass spectrometer running, we would use it to characterize and even, in some cases, acquire bits of protein sequence data. On the premise of doing data

analysis with the NOVA-4, I started developing programs for protein sequence analysis. This software suite included implementations of the Needleman and Wunsch similarity search and the Garnier secondary structure prediction algorithms. Also, I wrote programs to identify proteins on the basis of its amino-acid composition and to simulate the cleavage of proteins by a series of enzymatic or chemical methods.

In the process of building up this set of software tools I also typed in more than 1000 protein sequences. Some of them were entered from literature reports but the large majority were printed in a wonderful series of books, the *Atlas of Protein Sequence and Structure*. The 'Atlas', as it was then known, had first been published in 1965 by Margaret Dayhoff from the National Biomedical Research Foundation (NBRF) in Washington DC. In 1981 the Atlas consisted of a large book (the fifth edition of 1978) and three supplements. In total it listed 1660 protein sequences. At the time the Atlas was not available on any computer media.

I never had a chance to use the mass spectrometer and from then on I left the 'wet' lab and became what is now called a 'bioinformatician'. Of course this term did not exist at that time and people like me were generally thought of as failed researchers playing around with computers!

In February 1982 I published my first paper; it was a letter to the *Biochemical Journal* suggesting that research groups publishing protein sequences should compute and print a simple checksum that would 'facilitate the detection of typographical and keyboard errors'. To the best of my knowledge this checksum system was never used in a publication (but was implemented for more than 10 years in the SWISS-PROT SQ line) and nobody ever noticed that the example peptide sequence in the paper spelled out 'Help I hate math'! I also learned a useful lesson from that paper: standards or nomenclature are only used when they can be enforced!

The Sirius-1

In late 1981 I saw an article in a US newspaper describing a new microcomputer based on the Intel 8088 processor. It was called the Sirius-1. Sirius, the company that built it, had been created by Chuck Peddle, the designer of the Motorola 6502 CPU chip that powered the Apple II and the Commodore Pet, one of the first affordable personal computers. The Sirius-1 was, for that time, an incredible machine with a monochrome graphic resolution of 800 × 400, 128 Kb of RAM extensible to about 1 Mb, a speech synthesizer, and many other advanced features. It was running under a brand new operating system, MS-DOS. There was no comparison between the Sirius and the then newly introduced IBM PC. I easily convinced Robin to allow me to buy this machine for the department and managed to obtain from the Swiss distributor the first exemplar to be imported into the country. The decision to buy such a microcomputer and use it for scientific applications at a time when this was the hallmark of centralized time-sharing systems and expensive minicomputers had far-reaching consequences.

The BIONET user group

The first consequence of the arrival of the Sirius-1 in our department was that it attracted a lot of interest from young scientists in neighboring departments. It prompted me to create a user club for life science users of microcomputers. I named it BIONET, short for Biology Network. By coincidence this name was later independently used in the USA to name an on-line forum of life science users. BIONET was meant to be a forum to share experience in the use of the Sirius-1. It was also supposed to sponsor the shared development and distribution of software tools for sequence and statistical analysis. BIONET did achieve these goals but it soon grew in unexpected directions. It expanded geographically by attracting users from Lausanne and Fribourg and it quickly attracted users outside the life science fields. It also diversified in terms of its hardware support. However, the tide of IBM PC compatible computers slowly but inexorably swamped the far superior Sirius-1.

BIONET also quickly became a forum for exchanging 'informally obtained' copies of commercial software packages. At the time the high cost of microcomputers and software made it impossible for academic laboratories to buy full price licenses for most if not all the necessary software tools. In fact most software companies at that time were acutely aware of this problem and turned a blind eye to these activities. Some of them used the BIONET group as a source of expertise at a time when such expertise in PC-based software and hardware was very scarce and made their products available to our group on an unofficial basis.

In 1985 the price of software and computers went down and software and operating systems became easier to use. The University of Geneva finally recognized that microcomputers were here to stay and their usefulness went beyond 'playing games'. With the installation of a local area network spanning all the university's buildings and a support team in charge of PC selection, network connection and maintenance, the usefulness of BIONET decreased and it ceased operation in October 1986. At that time it was supporting a heterogeneous community of 54 research units that possessed more than 170 microcomputers.

The legacy of BIONET has persisted up to the present time. Some of the participants of this users' group are

working in different groups of what is now the Swiss Institute of Bioinformatics and at GeneBio. It also taught me how to organize a professional system of user support, a skill that turned out to be very precious later.

From NAPDB to PC/Gene via COMPSEQ

The second consequence of the availability of the Sirius-1 and its high quality graphic capabilities was that I was attracted to the idea of using it as a platform to write a sequence analysis package in the context of a PhD thesis. I submitted to Robin and the faculty a project whose summary was:

> The development for the departments of the University of Geneva working in the field of biochemistry or molecular biology of a powerful sequence analysis software package running on a 16 bit microcomputer which will be: fully interactive and user-friendly; integrating nucleic acid and protein analysis and capable of managing complete sequence data banks.

The project was accepted and I started work on it in October 1983.

Version 1.00 was released in March 1984. The package was then known as the NAPDB (Nucleic-Acid and Protein Data Bank) system. There were, at that time, 15 programs. It was distributed with release two of the European Molecular Biology Laboratory (EMBL) nucleotide sequence database and an 'in-house' protein data bank of 1200 sequences. In the summer of 1984 I changed the name of the package to COMPSEQ. During that summer, I received a visit from Michel Gazeau who, with a colleague, had created a small company called GENOFIT SA. Their original business was to sell restriction enzymes to Swiss labs. They wanted to expand their activity and were considering becoming the representative of the US software company, IntelliGenetics Inc. IntelliGenetics was, at that time, selling a minicomputer and mainframe software sequence analysis system called the IG-Suite. Gazeau asked me if there was any interest in Geneva in acquiring the IG-Suite. My answer was that it was too expensive, not very easy to use and anyway I was developing my own package. After a short demonstration of COMPSEQ, GENOFIT dropped the idea of representing IntelliGenetics and negotiated with the Medical Biochemistry Department the exclusive worldwide rights to commercialize COMPSEQ.

From its onset COMPSEQ had been written to be user-friendly. It was menu driven (and it took advantage of a newly introduced device, the mouse!) and integrated a very extensive context-sensitive help function. Therefore, it was not too difficult to make the transition from an academic to a commercial software package. Version 2, the first commercially available version, was released in October 1984. There were 30 programs. The first customers of COMPSEQ included Charles Auffray from the French CNRS who later headed the Genethon human ESTs sequencing effort and Plant Genetic Systems in Belgium, one of the first European biotechnology companies.

In October 1984, GENOFIT rented a space for a commercial booth at the *Computers in Science* conference in Washington DC. The booth adjacent to that of GENOFIT was occupied by a company called International Biotechnologies Inc (IBI). They were displaying another PC-based sequence analysis system: the IBI-Pustell package. While I was demonstrating COMPSEQ, Jim Pustell was demonstrating his software package. We quickly started discussing and immediately found out that we shared many ideas and principles on what should be the most optimal sequence analysis software. It also became evident that COMPSEQ and IBI-Pustell were, at that time, quite complementary. COMPSEQ was rich in protein sequence analysis tools while the IBI-Pustell program shone in the field of nucleotide sequence analysis and similarity searches. After almost a full day and night of discussion we decided to do two things.

The first was to write an article for a new journal that had just been announced, *CABIOS* (*Computer Applications in the Biosciences*) and the predecessor of *Bioinformatics*. We wrote a paper on the need to standardize data and software tools for sequence analysis. This paper was rejected on the grounds that we had no experience in this field, with the underlying assumption that we were dreamers!

Our second goal was to try to get GENOFIT and IBI to agree that we join our efforts and merge our two packages. Both companies were interested in this idea and the next day I flew with Jim to Boston. We first visited the Harvard lab where he was working under the supervision of Fotis Kafatos. Little did I know at that time that my path would cross that of Fotis 10 years later at EMBL. After Harvard we went to IBI in New Haven and started to discuss the practicalities of a joint development. A few days later I flew back to Geneva, the commercial people from IBI met with those of GENOFIT and it became apparent that both companies had very different aspirations on revenue sharing! The discussions broke down to the dismay of Jim and myself. We kept e-mail contact for a few months before losing touch. That was until 1989 when I met Jim again at NCBI. We then found out that we had both got married at about at the same time in 1985 and that our first child, Alison Ostell and Alice Bairoch were both born in 1986. The two girls not only have similar names but are also very similar both physically and in character! The name 'Ostell' above is not a typo, Jim and his wife are

the only people I know that have carried out a non-genetic crossing over by creating, when they got married, a new last name from part of their respective family names.

In 1985 it became apparent that an IBM PC version of COMPSEQ was necessary. Thanks to the help of two friends and programming wizards, Dominique Garin and Daniel Cerutti, COMPSEQ became relatively device independent. This change was later useful to port the program to Japanese computers and to be able to take advantage of new peripherals such as laser printers, speech synthesizers and sequencing gel readers. The first PC-compatible version, 2.3, was released in July 1985. It contained 33 programs.

In 1986 COMPSEQ started to be known outside Europe, where it competed with older well-established sequence analysis packages such as MicroGenie developed by Laurence Korn and Cary Queen and marketed by Beckman or DNAStar developed by Fred Blattner. IntelliGenetics became interested in the idea of distributing a PC-based software. On the instigation of Doug Brutlag, one of the original founders of IntelliGenetics (IntelliGenetics and IntelliCorp were two companies founded by a group of Stanford biologists and computer scientists that also included Bob Abarbanel, Peter Friedland and Larry Kedes) they decided to negotiate with GENOFIT for the exclusive rights of COMPSEQ for the USA and Canada instead of developing their own software package.

In the summer of 1986 IntelliGenetics asked me to visit them in Mountain View. They first wanted me to stop over in Atlanta where a molecular biology meeting was taking place, and to demonstrate COMPSEQ in their booth. The day before I left I got an e-mail from Mike Kelly, the CEO of IntelliGenetics. He did not like the name COMPSEQ and had thought of a better name, PCGene. He asked me if it was possible to change the name in time for the exhibition. I did not want to recompile all the code at such short notice and end up with debugging nightmares. I asked him to consider calling it PC/Gene as this name had the same number of characters as COMPSEQ, which meant I could simply replace all occurrences inside the binary code!

In a rerun of what happened in Washington DC in 1984, I stumbled on a tiny booth where a young student, Manuel Glynias, was demonstrating MacGene, a Macintosh-based sequence analysis program he had written. At that time the Mac RAM memory was limited to 128 Kb and the software development tools were very primitive. Nevertheless, Manuel had managed to develop an impressive software package using Forth computer language. It had many functions, was easy to use and had an impressive graphic interface. We swapped ideas and quickly became good friends. As I was sharing a hotel room with Denis Smith, then Chief Scientific Officer of IntelliGenetics, I talked to him about Manuel and he was easily convinced to invite him to Mountain View to ask him to develop a Macintosh equivalent of PC/Gene. Manuel developed something much more ambitious than a Mac version of PC/Gene. Over the years he developed GeneWorks, the first object-oriented sequence analysis package, using object Pascal. While developing GeneWorks, Manuel stayed in his hometown of Cleveland, finished his PhD degree and had time to do some research on the evolution of introns with Walter Gilbert. The connection between Manuel and Wally was instrumental in the creation of NetGenics. As the CEO of NetGenics, Manuel is currently fulfilling his long-lasting dreams of developing an object-oriented platform for bioinformatics using a technology based on Java and CORBA.

PC/Gene continued to expand both in the number of programs that it contained and in the number of its users. It underwent many successive releases and each brought new functions. At the end of 1988 the growth of the sequence databases started to cause problems. At that time PC/Gene had to be shipped with 53 1.2-Mb floppy disks that contained the DNA and protein databases. We therefore decided to start to distribute the databases (and later also the program files) on CD-ROM. The first CD was made in January 1989. It was a time-consuming and expensive process, and was only the second CD-ROM with molecular biology data, the first one having been made by Hitachi a few months earlier for their DNAsis/PROsis software package. The main problem we encountered was that none of the PC/Gene users had a CD-ROM drive. This led GENOFIT and IntelliGenetics to act as hardware suppliers to sell the only two models that existed at that time.

In February 1989 a major new version (6.0) of PC/Gene was released. It contained 76 programs, two of which had been developed by IntelliGenetics programmers. As I was spending more and more time on SWISS-PROT and PROSITE and also writing my PhD thesis, I was becoming much less productive in terms of new developments. Therefore for PC/Gene, the next 3 years were not so eventful. One should note, however, that in 1991 a Japanese version of PC/Gene was released; Teijin Ltd distributed it. It ran on NEC systems that (at that time) were incompatible with PCs.

In 1991 GENOFIT registered for bankruptcy. Two years before it had started to develop an automatic DNA sequencing machine. Such a development was too ambitious for a small, under-funded company. It could not compete against the likes of ABI, Dupont or Pharmacia. IntelliGenetics then took over the European distribution of PC/Gene.

In late 1991 I obtained from IntelliGenetics a budget that allowed me to hire a programmer in Geneva to take over most of the developments of new programs. Among the persons that responded to a job advert posted on the

BIOSCI user group was a Canadian student, Dorothy Miyake. I interviewed her in the office of IntelliGenetics in Mountain View, hired her and she arrived in Geneva in February 1992. She had in the meantime met and married one of the IntelliGenetics programmers, John Lowry, thus the PC/Gene development team in Geneva immediately tripled in size!

Starting in 1992 I tried to convince IntelliGenetics that it would be useful to port PC/Gene to the new Microsoft Windows environment. It was one of the major failures of the company that they did not believe that this was the natural path to follow to develop a new generation of sequence analysis software packages. WinGene, as the project was then called, never came into being. This oversight was, as it later became apparent, fatal to IntelliGenetics. In 1994 the company started to have financial problems and it was acquired by Oxford Molecular. Meanwhile, thanks to the work of Dorothy and John, new programs were developed and in April 1995, a release 6.85 of PC/Gene came out. It contained 82 programs and was the last version to be released. Oxford Molecular wanted to concentrate their effort on the development of OMIGA, their own Windows-based package and I had no more time for software development. PC/Gene was officially discontinued in December 1996.

In its 12 years of existence PC/Gene became the most widely used PC-based sequence analysis software. It was used by more than 2000 labs in 45 different countries. At least 600 published papers quote its use. For many researchers it was a very useful tool at a time when no other user-friendly sequence analysis programs were available. It was also the first software package with an emphasis on protein sequence analysis rather than nucleotide sequence analysis, as was the case for all other existing programs.

The PC/Gene saga owes much to all its users and especially a number of beta testers in Switzerland and around the world. It is also important to remember that, in addition to all the people already listed above, there were many employees of GENOFIT and IntelliGenetics who played a major role in all aspects of the debugging, documentation, sales and support of the package. I am, therefore, very grateful to Saeid Akhtari, Nancy Bigham, Eddie Brayman, Tania Broveak, Dave Callender, Mike Chalup, Jean-Pierre Dautricourt, Alan Engelberg, Williams Ettouati, Mark Good, Jean-Pierre Huber, Larry Krone, Claude Matringes, Sunil Maulik, John Moore, Patrice Pasquier, Nina Robinson, Lisa Schaechter, Murray Summers and Ganesh Sundaram.

The birth of SWISS-PROT

In 1983, when I started to develop what was going to become COMPSEQ, I needed access to both a nucleotide and a protein sequence database. I asked for and obtained a computer tape of the then newly available EMBL Nucleotide Sequence Data Library. The first version I received was release 2; it contained 811 sequences with a total of more than 1 million pairs of bases. Just to put these numbers in perspective, this is less than the total amount of bases that is now deposited in an average 5-hour time span in the DNA sequence databases!

It may sound bizarre in a period when massive amounts of information can be transferred around the world that in 1983 it was a major problem to transfer data from a computer tape to a microcomputer. In Geneva I could only read the EMBL tapes at the University Hospital computer center which was then equipped with a CDC Cyber mainframe system. Of course that computer did not have any communication program that allowed it to 'talk' to a microcomputer. So, to make a long story short, the only way to transfer the EMBL database was to hook up a Sirius-1 to a 300 baud acoustic modem, transfer a screenfull of data, quickly check to make sure the text was not corrupted and then to press the 'Enter' key. I remember spending almost a whole night doing that!

The protein sequence database that I initially used was the sequence collection that I had typed in while doing my master's degree. In 1984 I received the first available computer tape copy of the *Atlas of Protein Sequence and Structure*. It quickly changed its name to the 'Protein Sequence Database of PIR' (PIR first meant Protein Identification Resource and later Protein Information Resource). I started to use and distribute PIR with COMPSEQ, but I was quickly confronted with a number of problems. The format in which PIR was stored made it quite difficult to parse out information such as those concerning post-translational modifications. There was no mechanism that allowed a link to be made between a protein sequence and its parent nucleotide sequence in the DNA database (lack of cross-references). More significantly, most of the new sequences lacked any annotations concerning function, subcellular location and other important characterization information. As I was not interested in building up databases I kept sending letters to PIR to ask them to remedy this situation. But since I never got any satisfying answer I began to feel that I should try to address some of the above issues myself.

I found the line-oriented format of the EMBL database with its two-letter code for each type of data items very elegant and first tackled the problem of converting the PIR database to a format similar to that of EMBL. I did this using a mixture of software tools and manual intervention. I also started to add various types of information to what I then called 'PIR+'. After a few months it became clear that users of COMPSEQ liked this new format and appreciated the inclusion of new information. In 1986, when IntelliGenetics begin to distribute PC/Gene

in the USA they asked me if the database could also be distributed on the US BIONET on-line resource, which they were maintaining and developing under a NIH contract. In the summer of 1986 I decided to distribute the protein database independently of PC/Gene so that it would be available free of charge to anyone who needed it; I called it SWISS-PROT. The first release was made available on 21 July 1986. It contained just less than 3900 sequences (the exact number is not known as I have yet to find a copy of the first floppy disks!).

As SWISS-PROT followed the format of the EMBL database very closely I contacted the Data Library group of EMBL to see whether a collaboration was feasible. Among other things I wanted the EMBL to distribute SWISS-PROT on computer tape as they did with the nucleotide database. In June 1986 I went to Heidelberg to meet with three people: Greg Hamm, then director of the data library group, Graham Cameron, who replaced Greg as director a few weeks later and Patricia Kahn, who was in charge of scientific issues and user support. They decided not only to distribute SWISS-PROT but, more significantly, to collaborate in the maintenance of the database. At that time I was hoping I would return to write PC/Gene full time and I accordingly drafted a scheme whereby I would gradually phase out my involvement in SWISS-PROT. I planned to stop working on it in mid-1987. This of course did not happen and SWISS-PROT gradually started to take over all my time and my thoughts.

Patricia Kahn was instrumental in selecting and hiring two people at EMBL to help me in the annotation process. The first person to work on SWISS-PROT outside of Geneva, Rolf Apweiler, stayed for a few months before starting his PhD thesis at Boehringer. Five years later he came back to the EMBL and now leads the SWISS-PROT group at the European Bioinformatics Institute. The second person was Brigitte Boeckmann. She has worked for SWISS-PROT ever since, first in Heidelberg and now in Geneva. It should also be noted that another significant contribution of Patricia Kahn to DNA and protein sequence databases was her crusade to persuade journals to only publish nucleotide sequences after they had been submitted to the database. Until this was achieved the most time consuming activity of the Data Library was to manually type in the DNA sequences printed (and generally very badly typeset!) in different journals. Such a development seems obvious nowadays, but it took a lot of diplomatic skills and perseverance to achieve this goal.

In February 1987 the EMBL and the NIH organized a joint workshop 'Future Databases for Molecular Biology' in Heidelberg. During this workshop, the decision was made to consolidate the collaboration between EMBL and GenBank, represented at that time by Jim Fickett and Christian Burks. The creation of what is currently known as the international scientific advisory board to the nucleotide sequence databases was also decided during this meeting. But for me this workshop was very important because it was there that I met Jean-Michel Claverie, then at the Pasteur Institute. Jean-Michel had developed two protein sequence databases: PseqIP, a non-redundant collection based mainly on PIR and Russ Doolittle's NEWAT and PGtrans, a computer-generated translation of GenBank. He had also designed a FORTRAN-based sequence analysis package, SASIP, that was used at the Pasteur Institute and at some other locations. We decided to write a short funding proposal to the European Union to develop 'EuroProt', a non-redundant protein sequence database. EuroProt would have been built using SWISS-PROT and PseqIP/PGtrans. Jean-Michel quickly got an unofficial answer from Brussels that nobody would fund a proposal coming from people without an established track record in protein databases and that the inclusion of a scientist from Switzerland, a country not part of the EU, would not be seen in the best light. It was my first indirect encounter with European science politics. Over the years I have kept contact with Jean-Michel. He accepted the invitation to be a member of the jury on my PhD defense and I visited him very often, first at Pasteur, later at NCBI and now in Marseille. It should also be noted that Jean-Michel and his co-worker at Pasteur, Lydie Bougueleret, who has since made her own bright path through bioinformatics, were the first to develop, in 1986, heuristic methods for the detection of coding regions in eukaryotic genomic sequences.

At the EMBL the data library group was positioned next to two research groups that made up what was then known as the biocomputing group. These groups were respectively lead by Pat Argos and Chris Sander. Over the years I had numerous contacts with them and with many members of their group. Both Pat and Chris have played a major role in the development of bioinformatics. They have pioneered many novel approaches in the analysis of protein sequence and structure. The two group leaders were also very successful in attracting many bright PhD and post-doctoral students. One needs only to remember that Thure Etzold, Toby Gibson, Peter Sibbald and Martin Vingron have worked in Pat's group and that Peer Bork, Georg Casari, Liisa Holm, Christos Ouzounis, Burkhardt Rost, Reinhard Schneider and Gert Vriend were members of Chris's group.

It is less widely known that both groups also played an important role in the development of biomolecular databases. For example in 1988, Chris was the person responsible for the idea of initiating EMBnet, a group of collaborating nodes throughout Europe that provides bioinformatic services to the molecular biology community. The early inclusion in SWISS-PROT of structural information (for example secondary structure features)

stems from the many discussions I had with Chris. The collaboration with Peer Bork, which, as will be later described, was important in the development of PROSITE, was initiated when Peer came to EMBL as a visitor for a few months after the dismantling of the Berlin wall. Ten years later he is still at EMBL and still officially a visitor from the Max Delbrück Center for molecular medicine in Berlin! Another important development was the creation of the Sequence Retrieval System (SRS) by Thure Etzold. SRS, which allows complex queries to be made to a variety of heterogeneous databases, was the first software tool to make use of the cross-references that were an early hallmark of SWISS-PROT. Finally, the GeneQuiz software used for the automatic annotation of complete genomes was developed in Chris's group. Two of the people behind this work, Georg Casari and Reinhard Schneider, later funded Lion Bioscience, one of the leading European bioinformatic companies.

Thanks to the royalties of PC/Gene, I could hire first one then two people, to help me develop SWISS-PROT. The first people to work with me on SWISS-PROT in Geneva were Serenella Ferro and Jean-Pierre Patthey. Both of them have now moved to South America. Serenella is in Bolivia, but she continues to work for SWISS-PROT. She regularly comes to Geneva to train on new annotation tools and to keep track of the fast-changing methods used to build-up SWISS-PROT, and then goes back to La Paz to annotate entries. Jean-Pierre is in Chile and has become a gentleman farmer. He oversees a farm where cacti are grown and then infected with female cochineal insects so as to produce carmine, a red pigment used to color food. It is also interesting to note that for almost 10 years Jean-Pierre was the only male annotator in the Geneva SWISS-PROT group. One of the key reasons that made and still makes the SWISS-PROT group attractive to women scientists with children is that it is possible to work part time, with a flexible schedule and that part of the work can be done from home. All of which is not possible with practical laboratory work.

The birth of PROSITE

Like SWISS-PROT, the birth of PROSITE as a database was not a planned decision but rather a by-product of the development of PC/Gene. In 1986, Russ Doolittle published a small introductory book on sequence analysis: *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*. It provided a clear and impressive description of the then available methods that could be used to get the most from a protein sequence. This small green-covered book became my bible and, from what I later learned, this was also true for many other biologists striving to enter the world of what was then known as computer sequence analysis. Russ's book contained a chapter that stated: 'There are many short sequences that are often (but not always) diagnostics of certain binding properties or active sites. These can be set into a small subcollection and searched against your sequence'. This was followed by a figure showing some examples of such short sequences or 'patterns'. I thought it would be nice to have a program in PC/Gene that would scan a sequence with such types of patterns. I wrote such a program and called it 'PROSITE'.

I thought I would find many published examples of sequence patterns that would allow me to populate PROSITE. But, to my dismay, it turned out that this was not the case. Not only were there very few published sequence patterns but, to make matters worse, most of these patterns were not specific enough. They either identified too many proteins which were not members of the family under consideration or they failed to detect some bona fide members of a protein family. I decided to develop patterns myself as well as to document them. Each pattern was accompanied by an abstract that described the corresponding protein family or domain. PROSITE was thus becoming a hybrid beast, half program, and half database. When it was first made available, in PC/Gene release 5.16 in March 1988, it contained 58 entries.

In January 1988, while developing the first PROSITE patterns, I published with Jean-Michel Claverie a short note in Nature entitled 'Sequence patterns in protein kinases'. It provided a means for the quick identification of new members of this then emerging superfamily. Almost exactly a year later, Victor Jongeneel, a molecular biologist from the Ludwig Cancer Research Institute in Lausanne, contacted me. Victor was, at the time, a BIONET member and a very effective beta-tester of PC/Gene. He has since left the bench for the computer and played a crucial role in the development of bioinformatic activities in Lausanne (see below). Victor contacted me with a specific problem in mind. He and his colleagues were studying neprylisin (endopeptidase 24.11), a zinc-containing protease. There was no overall similarity between the sequence of that protein and other known sequences with the exception of a small region that contained two histidines. Such a region was reminiscent of those found in better-characterized zinc-proteases such as thermolysin and collagenase. We developed a PROSITE pattern from that conserved region. It only detected known zinc proteases with the exception of three proteins, one of which was the tetanus toxin. We published our observations, 'A unique signature identifies a family of zinc-dependent metallopeptidases' in FEBS Letters in January 1989. Since it was published, research groups that have discovered new families of zinc proteases have extensively cited this paper. But the most rewarding payoff of this work is that it provided the first indication that tetanus toxin could be a metallopeptidase. This was

experimentally shown to be true by Montecucco in 1992, thus opening up interesting pharmaceutical leads on how to inhibit the potent lethal effect of tetanus and the related botulinum neurotoxins.

The situation that had arisen with SWISS-PROT was quickly mirrored with PROSITE. Many people asked me if they could have access to the database independently of the PC/Gene. This prompted me to make PROSITE available to everyone, much to the dismay of IntelliGenetics who would have preferred that it remained in the exclusive realm of PC/Gene and other IG software products. In October 1989, I officially announced the availability of PROSITE in a talk I gave at the EMBO conference, *Patterns in Protein Sequence and Structure*, at the EMBL. In November 1989, I released a new version of PROSITE (4.0 with 202 entries). To make the database more widely known Chris Sander had it printed as one of a series of EMBL biocomputing documents.

In 1990 Victor convinced Bernhard Hirt, the director of the Swiss Institute for Cancer Research (ISREC) in Lausanne, that the creation of a biocomputing unit at ISREC was desirable. Bernhard asked me if I was willing to direct such a unit. I declined the offer but promised to help find the people that would be part of the research staff. As I was in e-mail contact with two Swiss post-doctorates, who were at the time doing bioinformatics research in Californian labs, I decided to combine one of my regular visits to IntelliGenetics with side trips to their labs. The first person I visited was Philipp Bucher. After a PhD in Zürich, he had done a first post-doctorate with Ed Trifonov at the Weizmann Institute and was then working with Sam Karlin at Stanford. He was an expert in the mathematical and statistical analysis of DNA sequences and had developed a database of eukaryotic promoters (EPD). The second person was Roland Lüthy who was working with David Eisenberg at UCLA. Roland had developed a new method for the prediction of a three-dimensional structure using sequence profiles (inverse threading). Both accepted an invitation to Lausanne. Roland left ISREC in 1993 to go back to the USA to work for Amgen. But Philipp stayed on and, among other activities, has played an important role in the development of PROSITE.

One of the weak points in using sequence patterns for the classification of proteins is that they are very sensitive to any sequence 'exception', whether due to a bona fide divergence or to a sequencing error. This is not the case for weight matrices (or profiles) built from sequence alignments. Philipp embarked on a research project to develop new methods for the development and validation of sequence profiles, which first made their appearance in release 12 of PROSITE in June 1994. Meanwhile, Kay Hofmann had joined Philipp's group. Both of them have applied profiles to the discovery of many new intracellular protein domains important for signaling. Kay left ISREC in 1998 to go to Memorec, but Philipp continues to work on profiles.

As mentioned above, I have collaborated for many years with Peer Bork. This collaboration took many forms, but has been extremely fruitful in one specific aspect, that of the description of extracellular domains in SWISS-PROT and PROSITE. Peer is the discoverer of many of these extremely versatile and modular domains. Together with Peer, we have created and published a nomenclature scheme for these domains.

Over the years many people have written programs that make use of PROSITE. Probably the most well known program was that developed by Rainer Fuchs (then at EMBL, now at ARIAD) and which was called MacPattern. Until the World Wide Web became popular, it provided the most user-friendly way to access PROSITE.

In 1991 I met Terri Attwood, at that time in Leeds, who was developing a database called PRINTS. Although PRINTS was based on a completely different algorithm than those used in PROSITE, it shared the same philosophy as to how protein domains and families should be documented. We often discussed that it would be useful to unify the formats of PROSITE and PRINTS. Meanwhile, PROSITE and PRINTS had been joined by BLOCKS (Fred Henikoff in Seattle), PRODOM (Daniel Kahn in Toulouse) and Pfam (Sean Eddy and Richard Durbin at the Sanger Centre). All these databases use different yet complementary algorithms to detect and classify protein domains and families. It was, therefore, logical to join forces. We applied for and obtained a EU grant to develop a project called InterPro. InterPro is a joint effort to create a unified, yet methodologically diverse, system for protein families/domains identification. It will provide a single set of ≪documents≫ linked to the various methods.

Currently PROSITE contains almost 1400 patterns and profiles. It is used routinely by a large user community. I hope it will continue to be useful in making sense of the wealth of sequence data which is accumulating.

The Trieste bioinformatic courses

In 1989 Doug Brutlag was asked to organize a one-week practical course on 'Computer Methods in Molecular Biology' for the International Center for Genetic Engineering and Biotechnology (ICGEB). An international organization, ICGEB belongs to the United Nations Industrial Development Organization (UNIDO) and whose role is to transfer biotechnology know-how to the developing world. There are two labs, one in Trieste (Italy), the other in New Delhi. The sequence analysis course was organized in Trieste and was one of the first of its kind. It provided students with a complete hands-on overview of all aspects of sequence analysis. Theoretical morning sessions were fol-

lowed in the afternoon by practical sessions and exercises. Such a model is now used by many EMBnet nodes all over Europe. Doug Brutlag asked me to be one of the instructors and to teach the use of PC/Gene, SWISS-PROT and PROSITE. Each student had access to a PC, and a Sun server was used to store databases and to teach the use of the IG-Suite.

The course was a big success and has taken place every year. Since 1993 Sandor Pongor, who heads the Protein Structure and Function group of ICGEB, has organized it. During the years the software, databases and hardware have evolved, but not the conception of the course. With pleasure I have returned to Trieste every year until 1997. Thanks to these courses, over the years, I have met many students from all over the world. Many of them are now active in the field of bioinformatics in their respective countries. Many anecdotes are associated with the Trieste workshops! For example:

- In 1989, the majority of the students were from Eastern European countries. The focus of evening discussion in the pizzerias of Trieste was how it would be possible to provide access to software and hardware across the iron curtain. Little did we know that this curtain was going to come up less than 4 months later.
- One morning in 1990, there was an electrical failure. The room where the course took place and where the computers were installed was plunged into darkness. A few minutes later the power came back. The system engineer entered the room and shouted, 'The sun is on fire!'. My first thought was that there had been a massive solar flare and that it had disrupted both electrical and electronic equipment. I rushed to a window to see whether something was visible before realizing that it was the Sun computer server that was burning! The machine was totally destroyed.
- In 1991 we were woken up very early one day by jet fighter planes passing over our hotel at a very low altitude. Slovenia had just seceded from Yugoslavia and the Italian air force was patrolling the Slovenian border, which was only 200 meters from our hotel.
- Martin Bishop, from the UK Human Genome Mapping Project (HMGP), has been an instructor at most, if not all, the courses. He generally comes with his family and they camp on the Karsic plateau above Trieste. There is an almost perfect correlation between the arrival of the Bishop family and that of a major thunderstorm that usually inundates their tent.

GCG

One of the earliest sequence analysis packages was developed by John Devereux from 1981 onward at the University of Wisconsin. Originally called UWGCG (University of Wisconsin Genetics Computer Group), it later became known as the GCG package. The original philosophy of the GCG package was to offer powerful software tools with a reduced user interface. They could be combined together in different ways so as to answer specific queries. Until a few years ago, GCG only ran on Digital VAX computers. It now runs on most UNIX platforms. For a long period of time, there was no interaction between GCG and PC/Gene. They co-existed on different hardware and operating systems. My first real contact with the GCG group took place in 1988 when they started to distribute SWISS-PROT with their software. These contacts intensified when PROSITE was first released. Outside of PC/Gene, GCG was the first package to implement a program (MOTIF) to scan a sequence with PROSITE patterns. In addition to John Devereux, the GCG core team also included Maggie Smith and Irving Edelman. These three individuals are models of generosity and dedication toward the goal of providing the best possible tools to life scientists.

It is not widely known and quite ironical that it is IntelliGenetics that forced John and his group to leave Wisconsin University and start their own company. IntelliGenetics threatened to sue the University for unfair competition on the basis of the fact that the GCG group was developing and selling commercial software while benefiting from the University's infrastructure. Thus, GCG became an independent company in 1990. It was acquired by Oxford Molecular in May 1997. Maggie and Irving left the company in late 1998, but the original spirit is still alive and the GCG package is constantly evolving.

NCBI

The GenBank nucleotide sequence database was originally developed at the Department of Energy (DOE) Los Alamos National Laboratory. Part of the distribution and maintenance effort was first contracted to a company called Bolt, Beranek and Newman (BBN) and, starting in 1987, to IntelliGenetics. In 1988 the US Congress supported the creation of a National Center for Biotechnology Information (NCBI). The goals of the NCBI, which is part of the National Library of Medicine, are to perform basic research in the field of computational molecular biology as well as build and distribute molecular biology databases. The NCBI was given the mandate to develop and distribute GenBank and gradually took over this task. Since its creation the NCBI has been directed by David Lipman. David has an extraordinary broad view of the challenges to be met in bioinformatics and it is thanks to his visionary approaches that NCBI has been so successful in the last 10 years.

The achievements of NCBI are well known and include, among others, the continuous development of the BLAST

family of similarity search software, the Entrez browser, the PubMed MEDLINE retrieval engine or the Taxonomy database and browser. But the biggest achievement of NCBI is that it attracted the most incredible team of researchers in the field of bioinformatics. Whether it is in the realm of software and database development (Jim Ostell, Dennis Benson, Greg Schuler, etc.), in that of sequence analysis (Mark Boguski, Eugene Koonin, David Landsman, John Wooton, etc.) or in that of algorithmic developments (Steve Altschul, Steve Bryant, John Spouge, etc.).

Since 1989, I have been a regular visitor to the NCBI and have interacted with many of its members. This led to a number of publications, but also to new developments in SWISS-PROT. For example, it is thanks to the NCBI that SWISS-PROT was the first database outside of NCBI/NLM to include cross-references to MEDLINE. The NCBI databases are modeled in Abstract Syntax Notation 1 (ASN.1), a protocol designed for the purpose of exchanging structured data between software applications. The first database external to those developed by the NCBI to be available in ASN.1 was ENZYME, a database of enzymatic nomenclature, which I have been developing since 1990.

On the way to ExPASy

In May 1990 I defended my PhD thesis. Its title was 'PC/Gene: a protein and nucleic acid sequence analysis microcomputer package, PROSITE: a dictionary of sites and patterns in proteins, and SWISS-PROT: a protein sequence data bank'. In my conclusions I wrote:

> Ideally one would like to present a protein sequence to a protein analysis system and obtain from this system some hints regarding the function of that protein, its similarities to other known proteins, if possible a tertiary structure model, and finally propositions for experiments that would prove or disprove some of the conclusions obtained by the system. Such a system should also be capable of explaining the reasons that led to reach a specific conclusion.

and later 'we have called this system EXPASY: for EXpert Protein Analysis SYstem'.

Around that time I met Denis Hochstrasser. Denis was head of the Digital Imaging Unit of the Geneva University Hospital. A medical doctor, Denis pioneered developments in two-dimensional PAGE electrophoresis techniques and their use for diagnostic purpose. Since 1983 he has teamed up with Ron Appel who did his PhD in computer sciences under his supervision. The subject of Ron's thesis was Melanie (Medical ELectrophoresis ANalysis Interactive Expert), a comprehensive two-dimensional gel analysis software package which is now in its third generation. After a few meetings it became apparent that we could and should collaborate to develop software tools and databases for the studies of proteins.

Denis proposed that together we develop the concept of ExPASy. He found a French biochemistry student, Eric Langevin, fresh out of a 1-year course on computing who was interested in doing a Master's degree using artificial intelligence techniques. Eric used the IntelliCorp KEE environment to develop an expert system that was capable of classifying protein sequences into a limited set of families based on the results of sequence analysis tools. The success of the prototype led to a full PhD project to develop this approach on a larger scale using C++, intelligent agents and an object-oriented database system. Unfortunately, after a few months, Eric decided to leave science and start a new career as a social educator. The ExPASy project was left in limbo. Ron then proposed that we implement an on-line system for molecular biology users of Geneva and Lausanne using Eric's former Sun machine. On that computer, named ExPASy, we installed a number of software packages such as GCG, the IG-Suite and FASTA as well as a comprehensive ftp archive of all the major life science databases then available. After a few months, more than 250 local users were routinely accessing the databases and software tools.

In 1991 I met Manuel Peitsch, who was at that time working in the Biochemistry Institute in Lausanne. Manuel was already an expert in protein three-dimensional structure. He had completed a 2-year post-doctoral stint with Jacob Maizel at the Laboratory of Mathematical Biology at the NCI in Fredericks where he had started to develop ProMod, a software program for three-dimensional structure homology modeling. Manuel was very active in helping Ron and myself build the software environment on ExPASy. In January 1994 he joined GIMB, the Glaxo research lab in Geneva where he started a career that, in less than 5 years led him from the position of a three-dimensional structure analyst to that of worldwide director of scientific computing for Glaxo Wellcome.

In 1993 thanks to the work of many bioinformaticians—especially David Kristofferson at IntelliGenetics, Don Gilbert at Indiana University and Reinhard Doelz at the Basel Biozentrum—life scientists were able to interrogate biomolecular databases across the Internet using two different network retrieval systems, WAIS and Gopher. Both systems offered menu-driven interfaces that allowed navigating across distributed resources; plus, WAIS offered a powerful indexing engine. In the framework of EMBnet Reinhard was promoting a new protocol to access databases across the network: HASSLE (Hierarchical Access System for Sequence Libraries in Europe). In Geneva, we were wondering what we should do to

make SWISS-PROT available on the Internet until one day Ron came back from a meeting where he had seen a demonstration of a program running on a new network access protocol. The software was Mosaic and the protocol was the World Wide Web protocol. The fact that the Web also supported Gopher and WAIS and that the graphic interface of Mosaic was far superior to that available before, convinced Ron and myself that we should set up a small experimental Web server around SWISS-PROT. Ron immediately started working on that concept.

The ExPASy server (www.expasy.ch) was born on 1 August 1993. Then there were less than 150 Web servers worldwide. To the best of our knowledge it was the first on the Web for the life science community. We were very pleased to see that it was accessed 7295 times during its first month of activity. We never imagined that a few years later it would be accessed at a rate of more than 2 million per month. The first year of ExPASy was a very exciting time period. I vividly remember installing Mosaic and demonstrating the concept of the Web using ExPASy in both Russ Doolittle's and Milton Saier's labs at UCSD in 1994. The excitement generated by the discovery of what was suddenly possible was exhilarating. A trip to the Weizmann Institute in May 1994 led Leon Esterman, the head of the Israeli node of EMBnet, to mirror part of ExPASy a few weeks later and Joel Sussman, then head of the Protein Data Bank (PDB), decided on the spot to start a Web server for the PDB. Some of the first-time reactions to the extent of the information available on the Web had nothing to do with science. I still have in mind the image of Keith Tipton's pipe dropping from his mouth when he realized that there was a picture of his favorite Dublin pub on a Web page!

A number of individuals played a key role in the early days of the Web. Keith Robinson, then finishing his PhD in George Church's lab at Harvard, created the first 'portal' for life sciences, the *WWW Virtual Libraries: Biosciences* was linked to all emerging resources for the life sciences on the Web. It also included a searchable index. Peter Murray Rust, then at Glaxo in the UK, pioneered the concept of 'clickable biology'. He started the first 'cyberspace' course in protein structure analysis. Students from all around the world would convene at a certain preset time to chat and ask questions to the various tutors in different labs. The course material and the results of student exercises were posted on the Web.

In May 1994 the first Web conference was hosted at the CERN. A tiny room was reserved for people wishing to discuss applications to biology. The small group of people that drifted in that room included Peter Murray Rust, Peter Stoehr from the EMBL data library, Ron and myself. The meeting was useful in setting up some standards for biology-specific mime-types like the one that enables web browsers to automatically load three-dimensional structure visualization tools such as RASMOL or the SWISS-PdbViewer.

Over the years, many people have participated in the development of ExPASy. But one person must be singled out, Elisabeth Gasteiger. Elisabeth, who has a degree in mathematics, arrived in Ron's group in early 1994 as a participant in the European community Comett student visitor program. She was supposed to stay in Geneva 6 months. Ron asked her to develop a number of software tools on ExPASy using Perl. She was so skilled that we did our utmost to convince her to stay in Geneva. She joined the SWISS-PROT group where she now heads software development. She has made major contributions to all aspects of ExPASy. The large community that uses ExPASy owe much to her dedication to the task of building a comprehensive proteomics Web server.

In February 1993, Ron and Denis decided to incorporate two-dimensional gel reference images (maps), into a database with information on the proteins identified on these maps. We toyed with the idea of calling it SWISS-SPOT before deciding to call it SWISS-2DPAGE. Thanks to the Web it was made available with an intuitive interface where information on a protein can be obtained by clicking on the corresponding spot on the gel image. As SWISS-2DPAGE shares many features with SWISS-PROT including primary accession numbers, in 1995 Ron proposed a system to federate two-dimensional gel databases. By abiding to a series of simple rules, developers of two-dimensional gel databases made it possible for users to seamlessly navigate through different databases so as to access information on a specific protein.

In 1992, Manuel Peitsch developed SWISS-MODEL, the first fully automated software for three-dimensional structure modelling. It was first ran as an e-mail server, and in April 1994 it was implemented on the ExPASy Web server where it became an instant success. In 1995, Manuel hired Nicolas Guex, a plant molecular biologist who had just finished his PhD degree. Nicolas, who had been programming since he was a child, in his spare time had developed a Macintosh program for three-dimensional structure visualization. As soon as he joined Manuel, he embarked on the development of a new generation workbench for protein three-dimensional structure. The first version of SWISS-PdbViewer was released in autumn 1995. It is now in its third generation and runs on three major operating systems (Windows, Macintosh and Unix).

ExPASy has constantly evolved during its 6 years of existence. It has now been accessed 60 million times by a total of more than 820 000 computer hosts from 151 countries. Three mirror sites have been established in Australia, Canada and Taiwan, and new sites are due to open in Bolivia, Brazil, China, India, Israel, South Africa and the UK.

The European Bioinformatics Institute (EBI)

In the early 1990s the EMBL Data Library group in Heidelberg was going through a very rough time. It was severely under-funded and could not deal efficiently with the many missions that it had gradually acquired in its 10 years of existence. Many people at EMBL, but also those in most European scientific circles, were aware that it was time to create an efficient European infrastructure for bioinformatics and especially for the maintenance of the nucleotide database.

In 1992 the European Union had commissioned a study to develop plans for the establishment of a European Nucleotide Science Center (ENSC). The ENSC was envisaged to be a

> reliable and professional DNA database service to the existing and future European user community which will enable Europe to maintain a leading role in computational activities relating to DNA sequences, and further enhance the European position in molecular biology research, development and applications technology.

In parallel, the EMBL presented a report that concluded that an efficient and quick solution to the creation of a bioinformatics infrastructure with a critical mass was to create a new EMBL facility, an outstation in the manner akin to those which already existed in Grenoble (synchrotron) and Hamburg (DESY). EMBL outstations are geographically independent units with a specific research focus.

The idea quickly gained ground and in December 1992, it was decided by the EMBL Governing Council to create the European Bioinformatics Institute (EBI). The next step was to decide where the EBI would be located. A call for proposals, with a rather short deadline (February 1993), was then published. Many countries were interested to host the EBI, but only two were able to meet the deadline. Germany was proposing to install the EBI in a new building next to those already built for the EMBL in Heidelberg. The UK proposed to host the EBI in a new genome campus in Hinxton near Cambridge. The UK bid was spearheaded by Michael Ashburner, professor of Genetics at Cambridge University, where he combined his life-long expertise (read passion!) on *Drosophila* with his interest in bioinformatics. Michael developed FlyBase, the most comprehensive genomic database and the only one to contain bibliographical references dating back to 1684! In his opinion *Drosophila* is the most (if not the only) interesting organism as it is much more complex than mere mammals such as humans. After all, we do not have wings, we lack antennas and have only one pair of legs! Michael's sense of humor and understatement make any discussion or meeting with him most memorable.

The attractiveness of the UK bid was manifold. One crucial reason was that the Wellcome Trust, one of the largest charity organizations, was willing to foot the bill for building, in the grounds of Hinxton Hall, a complete infrastructure for genomics. In addition to the EBI, the campus would also host the newly established Sanger Centre (now involved in human genome sequencing) and the HGMP. The UK bid was accepted in March 1993 by the EMBL Council. It was then time to start the process of defining more precisely the EBI mission and structure. Lennart Philipson who had directed EMBL for 12 years was leaving and the newly appointed director was Fotis Kafatos. So, one of his first tasks was to set up the EBI.

Fotis created an Advisory Committee for the European Bioinformatics Institute (ACEBI). I was asked to be a member of that committee. The ACEBI had to deal with many issues, one of which was to nominate candidates for the position of director of the new outstation. When I told Denis that the EBI was looking for a director, he asked me if the job profile necessitated an extensive knowledge of bioinformatics. While it was something we were originally looking for, it was more and more obvious what we were looking for was someone capable of leading a relatively large infrastructure and with knowledge of big computing projects. Denis mentioned that he might know someone who fitted that profile. That person was Paolo Zanella. Paolo was at CERN in Geneva for a number of years. He built up and directed the CERN Data Handling division (300 persons). Paolo modestly said that one of his achievements at CERN was that he was in the position to 'kill' the Web project when it was in its infancy but that he did not do so! Instead he set up a policy that allowed people working in his group to use part of their time to start new projects. It was in such spare time that Tim Berners-Lee created the World Wide Web. In 1994 Paolo was a professor of computer sciences at Geneva University and director of the CRS4 computing center in Cagliari, Sardinia. I met Paolo and spent an afternoon with him discussing bioinformatics. The outcome of this discussion was that he applied for the position and that he was selected to be the first director of the EBI.

At the end of 1994, the Data Library group moved out of Heidelberg into temporary barracks on the Hinxton campus. In September 1995 the building of the EBI was completed and the EBI became fully operational. It is organized under three programs: service (whose main activities are the EMBL nucleotide database and SWISS-PROT), research and industry. There are now more than 120 persons working at EBI.

Paolo stayed at the EBI for 3 years and left in late 1997 to become the chairman of Synomics, a newly created bioinformatics company founded by Tom Flores (formerly from EBI) and Steve Gardner (previously at Astra). The EBI is now co-directed by Michael Ashburner

and Graham Cameron who both provide, since the birth of EBI, the scientific and organizational impetus necessary for the development of this highly crucial research and service center. The EBI plays a key role in European bioinformatics. It differs from the NCBI in that its international nature makes it more open to external collaborations. In many ways it acts more as a federator of bioinformatic activities rather than a centralizator.

The birth of the concept of proteome

In the summer of 1993 Denis went to Australia and established a scientific collaboration between his group and that of Keith Williams, then at Macquarie University in Sidney. Both groups were heavily involved in the development of two-dimensional gel separation and protein identification methods. In October 1995 I became the recipient of the Helmut Horten Foundation award. This award provided me with funding for a 5-year period for salaries and equipment. The award was directed toward research into the characterization of proteins and more specifically the study of post-translational modifications. Denis proposed to use part of the money to hire Marc Wilkins, a post-doctoral fellow, who was at the time working in Keith Williams's group. Marc stayed for 2 years in Geneva and his extreme scientific productivity is reflected by the many papers he published during that period and, with Elisabeth Gasteiger, the development of a complete suite of software tools for protein identification on ExPASy. But there is one thing that Marc did which had a much more profound impact on the life science community than anyone could envisage at the time. In 1995 Marc created a new word, *proteome*. It is defined as 'the total protein complement of a genome'. The success of this new word is exemplified by the number of articles, conferences, books and companies that already mention either the word proteome or its derivative, proteomic.

It should be noted from an historical perspective that while the first published use of the word proteome is in a publication from Marc in the August 1995 issue of *Electrophoresis*, the first time that most people became aware of the concept is when they read a news article in the 20 October issue of *Science*. It was entitled 'From genome to proteome: looking at a cell's proteins' and it was written by Patricia Kahn who, since her days at EMBL, had become a science journalist.

In 1998 Marc went back to Australia where he and Keith Williams funded Proteome Systems Ltd. (PSL), a proteomic biotechnology company.

Bacterial genomes

One of the people I met during my visits to NCBI is Ken Rudd. Ken was working for GenBank but his primary interest was and still is *Escherichia coli*. For years Ken has been tracking down all possible sources of information on this very versatile bacteria. He has done an incredible amount of detective work in track down sequencing errors, to detect as yet undetected protein-coding genes and to identify the function of many genes. In the last 6 years, we have exchanged hundreds of e-mail messages and have worked together to annotate *E.coli* protein entries in SWISS-PROT as completely as possible. While at NCBI, Ken started to develop EcoGene, a genomic database. Now at the University of Miami School of Medicine, he continues to maintain it and to study various aspects of the *E.coli* genome and proteome.

I caught the 'virus' for bacterial genomics from Ken. I am constantly in awe when confronted with the adaptability and complexity of the microbial world. My interest for bacterial genomes has also allowed me to meet very interesting people whose diversity probably reflects the scope of the domain.

I remember that Ken once introduced me to someone who was spending a lot of time photocopying scientific papers in the corridor outside Ken's office at the NCBI. This person was Bobby Baum, well known to the yeast and *E.coli* communities. He was very regularly distributing newsletters packed full with highly interesting discoveries he had made while studying newly published DNA or protein sequences. What made Bobby's achievements very special is that he had never used a computer. All his observations were done solely on the basis of looking at printed sequence data. He could literally run similarity searches in his head! Furthermore his newsletters were all typed in, multiple alignments and all, using a typewriter.

In 1994, at the third international *E.coli* genome meeting in Woods Hole, I met for the first time a number of people which whom I have been interacting since. Among those were Antoine Danchin, Monica Riley, whose classification of *E.coli* gene products has been used by many genome sequencing groups and Terri Gaasterland, the author of Magpie, a powerful automatic genome analysis software. Antoine Danchin was the then coordinator of the international project to sequence the genome of *Bacillus subtilis*. My meeting with Antoine led to a long-term collaboration to annotate *B.subtilis* proteins. Ivan Moszer, a former PhD student of Antoine, plays a key role in this ongoing project. He is the developer of the SubtiList database, the most comprehensive depository of molecular information on *B.subtilis*.

In Geneva one of the earliest and most active user of PC/Gene was a plant biology department, the Laboratoire de Biologie Moléculaire des Plantes Supérieures (LBMPS), headed by Bill Broughton. The LBMPS is particularly interested in studying the interaction between leguminous plants and the symbiotic bacteria that provide such plants with nitrogenous compounds. They embarked on the molecular characterization of a large plasmid

from NGR234, a *Rhizobium* strain. This plasmid harbors the genes responsible for the process of nodulation and nitrogen fixation. As the lab was not equipped to do a massive amount of sequencing it concentrated its effort on sequencing regions that seemed to contain key genes. All this was to change in 1996. Xavier Perret, a long-time member of the lab, had met André Rosenthal while doing a post-doctoral year with Sidney Brenner in the UK. André was establishing a large sequencing lab in Iena (Germany) that was intended to take part in the sequencing of the human genome. Xavier managed to convince André that sequencing the total genome of the symbiotic plasmid was a good side project to test his new equipment and sequencing strategy. André delegated this project to one of his PhD students, Christoph Freiberg. In the summer of 1996 the sequence—536, 165 base pairs—was finished and Xavier asked me to take part in the analysis of its proteins. We worked on it very hard for a few months and some of the results of this analysis were published in *Nature* in May 1997, as 'Molecular basis of symbiosis between *Rhizobium* and legumes'.

The development of SWISS-PROT up to the funding crisis in May 1996

In the 1990s SWISS-PROT continued to grow in size along with the extent of the annotations that accompany each entry. The amount of work necessary to maintain SWISS-PROT also paralleled this growth. It was time to find some 'real money' for an activity that had, up to that time, never really been officially funded! I applied and obtained a Swiss National Science Foundation (SNSF) grant in April 1993 for an initial period of 2 years. This grant covered four new positions as well as computer equipment. In July 1993, I applied for and received a 3-year grant from Glaxo for an additional annotator position. At the end of the 2 years, the SNSF extended my grant for an additional year. They indicated that, due to the international nature of the SWISS-PROT database, it ought not to be funded by money reserved for national projects, but rather from funds intended for projects at the European or International level and to which Switzerland participated. With the EBI, we therefore applied in December 1995 for a EU infrastructure grant (Framework 4). Thus began a rather unfortunate story.

In April 1996 we were advised that the proposal had been evaluated favorably by the scientific experts of the EU, but was not accepted at a higher level. In retrospect it seems that the main reason for this rejection was that according to the EU regulation these infrastructure grants were only provided to complement existing local funding. But as Switzerland was not going to fund SWISS-PROT using financial resources other than those that would have been used to pay for their part of the grant, the proposal could not be supported. This last sentence may seem obscure and illogical, but believe me a full description of the intricacies of the situation would require a couple of pages. In any case, the proposal was rejected while proposals for projects that depended on the existence of SWISS-PROT for the outcome were accepted.

Having learned the extent of the problem, the EU seemed genuinely concerned but did not have the means to reverse the decision. We were asked to resubmit the proposal. Such a process would have taken almost a year and we only had 2 months left for salaries. In Switzerland, money for SWISS-PROT was available, but could not be assigned to such a purpose without a positive backing of the EU towards the grant. We were in a classical 'catch 22' situation. Everyone agreed that there was a problem, that it ought to be solved, but were unable to do anything because of procedural reasons!

We had no choice but to make the SWISS-PROT users community aware of our plight. We drafted an appeal for help that was posted on ExPASy on 10 May. In this appeal we informed users that SWISS-PROT, the associated database and ExPASy would disappear, due to lack of funding, on 30 June 1996. The results of this appeal were incredible. The first messages of support started to come in only a few minutes after it had been posted. In total more than 2500 e-mails, letters and even petitions signed by whole departments or institutes arrived in the next 2 months from all over the world. I was very emotionally moved by many of these messages of support. When, as today, I reread some of these letters, I can only feel grateful for what people wrote to us and how strongly they reacted to the possibility that SWISS-PROT could disappear.

The reaction to the appeal started with what was later called: an 'Internet storm of protest'. It quickly also reached the press. *Nature* and *Science* published news articles about the plight of SWISS-PROT. Swiss newspapers and magazines also reported extensively on the problem. They published extracts of some of the messages of support. One of these messages, originating from India, proposed collecting financial donations from scientists in Indian biological research centers, exemplifying (more than others) the dysfunction that could allow a major industrial nation to be unable to fund a service of worldwide importance.

Two weeks after we sent the appeal, the Geneva State Counselor, Guy-Olivier Segond, asked Ron and myself to come and see him. At the outcome of that meeting he had committed financial support for both SWISS-PROT and ExPASy until the end of 1996. Meanwhile, discussions were taking place both in Brussels and in Bern. Ruth Dreyfus, the Swiss minister of health, responding to a question in parliament, emphasized the importance of SWISS-PROT for life scientists and stated that a long-term

solution to the funding problem of Swiss bioinformatic service activities was necessary.

The birth of the Swiss Institute of Bioinformatics

The outcome of the funding crisis of 1996 was a recommendation from the Swiss scientific funding agencies that a stable long-term funding mechanism be sought for both SWISS-PROT and the Swiss EMBnet node (which also had major funding difficulties). The Swiss science budget is established on a 4-year basis and the next budget cycle started on 1 January 2000. Temporary, limited funding was allocated for the interim period of 1997–1999. In 1997 the leaders of the five groups working on various aspects of bioinformatics in Geneva and Lausanne, namely Ron Appel, Philipp Bucher, Victor Jongeneel, Manuel Peitsch and myself decided to create an institutional framework around the long-standing collaboration that had evolved over the years. We wrote a 'white paper' that stated:

It has been emphasised by Swiss scientific authorities that it is now essential and urgent to promote the creation of 'centres of excellence' in interdisciplinary domains that are economically important and crucial for tomorrow's society. Therefore we propose the creation of a Swiss Institute of Bioinformatics (SIB). The goals of this institute are:

- To promote the development of software tools and databases in the field of bioinformatics;
- To sustain a high-quality research program in bioinformatics;
- To provide, in collaboration with academic partners, a curriculum of courses and seminars for the formation of research scientists in the field of bioinformatics;
- To offer services to the Swiss scientific user community through the Swiss-EMBnet node.

It would take too long to describe the complex obstacle course that we had to go through to achieve these goals. It is never easy to create a new institution, especially if that institution has many academic and funding partners as is the case with the SIB. To summarize; on 30 March 1998, the SIB was created as a non-profit foundation. It then successfully applied to the Swiss Federal government for the funding of parts of its activities within the legal framework of an article of law that allows the Federal government to fund research institutions of national or international importance. It is important to note that this law does not allow the government to fund more than 50% of the budget of such an institution. It is expected that these institutions seek funding from other sources; the preferred solution being that they generate revenues through the commercial exploitation of their research activities.

The SIB is organized in a manner akin to that of Switzerland, a decentralized federal state. Each of the five groups in SIB has its own budget for research and service activities. Each group contributes to a common budget that is mainly spent on computer infrastructure, administration and teaching activities. The institute is overseen by a foundation council where all institutional partners are represented. An international scientific advisory board reviews the scientific activities of the SIB. The group leaders make up the executive board of the institute and elects one of its members as the director. We have just re-elected Victor Jongeneel for a second 1-year mandate as director of the SIB. Thanks to the way that the SIB is organized it can organically expand by either 'budding out' a new group from one of the existing groups or by co-opting bioinformatics groups in Swiss academic institutions which are not yet partners of SIB.

As I write this article in November 1999, there are more than 75 researchers working for the five SIB groups and we expect to reach the 100 persons mark in the year 2000. We have just started a full curriculum in bioinformatics; it is a Master's degree recognized by and co-organized with the Universities of Geneva and Lausanne. It consists of several courses as well as hands-on exercise sessions. Fifteen students have started the course this year. When they finish their degree in about a year, we expect that they will be the first representatives of what we hope to be the new generation of Swiss bioinformaticians.

The birth of GeneBio

The process that started at the onset of the funding crisis of 1996 and that led to the creation of the SIB gradually brought us to confront a major dilemma. It was clear that the increased data flow had created a requirement for resources that could not be addressed in full by public funding. This had caused SWISS-PROT to fall behind the research. We needed to find a way to address this substantial resource shortfall but, on the other hand, it was of utmost importance that the database remained freely and easily accessible to the academic community. The solution was to ask commercial users to pay a license fee. First, we assessed two models: one in which the SIB and the EBI would directly deal with the licensing process, the other one in which an existing bioinformatics company would administer it. Both solutions were rejected, the first because it implied grafting a commercial infrastructure into academic research institutions, the second because no company was willing to channel a major proportion of the license fees back to SIB and EBI. That left a third solution, the creation of a new company. As scientists with no experience of the process of building up a start-up company, we were quite circumspect in taking such a decision. But after a few months Ron, Denis and

myself decided to act as the scientific founders of Geneva Bioinformatics SA (GeneBio).

Thus, GeneBio was founded in November 1997, and in February 1998, the first employees were hired. In July 1998 it became the exclusive commercial representative of the SIB. It completed its financial set-up in August and started its commercial operation on 1 September 1998. In March 1999 it moved into its current location, in a building next door to the University Medical Center where Geneva's part of the SIB is located.

The contract between GeneBio and the SIB stipulates that GeneBio returns up to 75% of the income generated by the sales of annual licenses of databases and software products developed by the SIB. This is already providing a major revenue boost for the SWISS-PROT groups at SIB and EBI. Another important issue was that we wanted to ensure that nothing would change in the methods by which academic or commercial users would be able to access SWISS-PROT. Thus the licensing system used by GeneBio is atypical in that it is based on trust. Companies and academic users continue to have full and unencumbered access to the databases from a wide variety of Web and ftp sites. Commercial users are asked to contact GeneBio and to pay for a yearly license. The price of the license is a function of the number of users in an organization. So far, we have no reasons to regret this decision. There are currently 120 major companies that have subscribed to SWISS-PROT and more companies are in the process of signing up for their license.

Today GeneBio has 20 employees. In addition to its activity as commercial representative of the SIB it is developing its own products. These include added value specialized databases for the pharmaceutical industry.

SWISS-PROT from 1996 to today

In 1996 it was already clear that the increased data flow from genome projects was going to be a major challenge for SWISS-PROT. Maintaining the high quality of the database requires careful sequence analysis and detailed annotation of every entry. This was, and still is, a major rate-limiting step. We did not wish to relax the editorial standards of SWISS-PROT and there was a limit to how much the annotation procedures could be accelerated. Yet it was vital to make new sequences available as quickly as possible. To address this concern, in 1996 we introduced TrEMBL (Translation of EMBL nucleotide sequence database). TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences in the EMBL database, except for those already included in SWISS-PROT. TrEMBL is, therefore, a complement to SWISS-PROT and sequence entries only move out from TrEMBL and enter SWISS-PROT after having been curated by one of the annotators in Geneva or Hinxton. It should be noted that the name TrEMBL was coined by Thure Etzold. He had written a program to generate a conceptual translation of all coding sequences from EMBL and had called the resulting data set, TrEMBL.

In November 1996, a scientific meeting took place in Jerusalem (24th Aharon Katzir-Katchalsky conference; BioInformatics—Structure) which was co-organized by Joel Sussman, then director of the PDB, and by myself. It was set-up to celebrate the 25th anniversary of the PDB and the 10th anniversary of SWISS-PROT. It was one example, among others, of the collaboration that has been existing for a long time between SWISS-PROT and the PDB.

From 1996 to 1999, SWISS-PROT grew by 25 000 sequences to reach a total of 80 000 entries. This growth was supplemented by many enhancements and by a significant increase in the number of databases to which SWISS-PROT is cross-referenced. In this period of time, TrEMBL grew from the 86 000 entries in its first release to a little more than 200 000 entries. Thanks to the work of the EBI group, lead by Rolf Apweiler, and in particular to a wonderful team of programmers, TrEMBL has been significantly enhanced by computer-generated high-quality (yes these are generally quite contradictory concepts!) annotations.

We have recently started a major overhaul of SWISS-PROT. This has taken place at various levels. We are making improvements to the format (for example, the conversion from all upper case to mixed case) and adding new types of information (for example, data on the use of specific proteins as pharmaceutical drugs). Also we recently launched two major and long-term initiatives. The first one, the Human Proteomics Initiative (HPI), is a major project to annotate all known human sequences according to the quality standards of SWISS-PROT. The second project is to speed up the annotation process of proteins originating from complete microbial genomes by an approach combining human expertise and automatic annotation processes.

There are now more than 60 people working in the SWISS-PROT groups of Geneva and Hinxton. We have reached a critical size, which can either allow the efficient development of the database or may lead to a bureaucratic, inefficient organization. I feel that my main mission is to make sure that we take the best path and that we do not betray the trust that thousands of users have vested in us.

Conclusion

Lists of names are quite boring, yet there are many people not yet cited with whom I have had fruitful and very enjoyable collaborations. So please allow me to also thank Enrique Abola, Alex Bateman, Tim Clark, Stuart Clarkson, Jean-Jacques Codani, Jos Cox, Philippe Dessen, Laurent Duret, Takashi Gojobori, Shigeaki Harayama,

Bob Harper, Bernard Henrissat, Florence Horn, Bernard Jacq, Peter Karp, Toni Kazic, Frank Kolakowski, Jack Leunissen, Patrick Linder, Rodrigo Lopez, Hanah Margalit, Francis Ouellette, Guy Perriére, Carl Price, Neil Rawlings, Alex Reisner, Pierre Rouzé, Aiala and Jonathan Reizer, Christoph Sensen, and Michael Zuker.

It is also obvious that my foremost thoughts are for all my co-workers in the SWISS-PROT groups in Geneva and Hinxton. Thousands of users depend on their work. They should, therefore, not stay anonymous. So a big thank you to: Rolf Apweiler, Kristian Axelsen, Kirsty Bates, Pierre-Alain Binz, Margaret Biswas, Marie-Claude Blatter Garin, Brigitte Boeckmann, Silvia Braconi, Sergio Contrino, Danielle Coral, Livia Famiglietti, Nathalie Farriol, Stephanie Federico, Serenella Ferro, Wolfgang Fleischmann, Gill Fraser, Elisabeth Gasteiger, Alain Gateau, Cathy Gedman, Arnaud Gos, Nadine Gruaz-Gumowski, Henning Hermjakob, Chantal Hulo, Nicolas Hulo, Ivan Ivanyi, Janet James, Eva Jung, Vivien Junker, Alexander Kanapin, Youla Karavidopoulou, Corinne Lachaize, Fiona Lang, Minna Levhvaslaiho, Michele Magrane, Maria-Jesus Martin, Karine Michoud, Nicoletta Mitaritonna, Virginie Mittard, Madelaine Moinat, Steffen Möller, Nicola Mulder, Julia Williams Nef, Claire O'Donovan, Isabelle Phan, Sandrine Pilbout, Bernd Röchert, Lucia Rodriguez-Monge, Claudia Sapsezian, Margaret Shore-Nye, Christian Sigrist, Shyamala Sundaram, Arno Velds, Anne-Lise Veuthey, Eleanor Whitfield, Nadine Zangger, Evgueni Zdobnov, and Angella Zutta.

My last words are to thank all users of the software tools and database mentioned in this article. You are the participants in one of the most exciting endeavors of the 20th and 21st centuries. Understanding the molecular basis of life is a task that has and will continue to have far reaching consequences for many areas of life sciences and for society in general. Each life sciences researcher brings his contribution to this massive pool of data and the bioinformatician mission is to act both as librarian and analyst of this information.