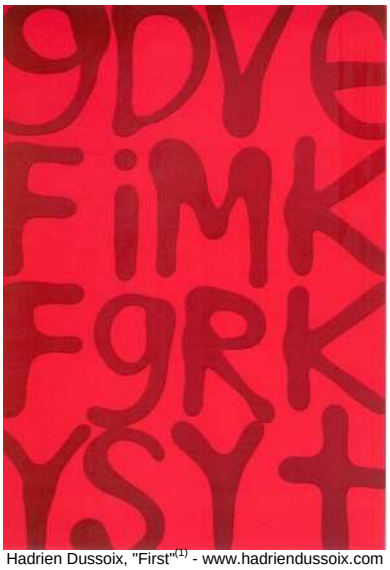
prolune
PROTEINES À LA >UNE<

Issue 18, August 2006

The beginnings of a database

An interview with Prof. Amos Bairoch



Hadrien Dussoix, "First"[1] - www.hadriendussoix.com

The Swiss-Prot database is celebrating its 20th anniversary this year. This electronic encyclopedia on proteins – which is now acknowledged throughout the world – saw the light of day in July 1986. The story of Swiss-Prot is inseparable from its founder's – Professor Amos Bairoch – who is regarded as one of the pioneers in bioinformatics. Here is the extraordinary background of a database that – over the years – has gained great respect, and a man who has dedicated his life to its development.

Amos Bairoch, what is the purpose of a database such as Swiss-Prot, and what is the meaning of "bioinformatics"?

Swiss-Prot is intended for the use of researchers in the life sciences. The database sums up all the available information we can find on a given protein in a given species, i.e. where it is found in a cell and how it is modified in time for instance, the most important information being of course the protein's role in an organism. If we don't know what a protein is for, we can get an idea by comparing it with one which is similar. Swiss-Prot is therefore a tool that helps to characterize newly identified proteins, for which the order of amino acids has been determined, i.e. its sequence. But this condensed information does not replace scientific papers, in the same way as information found in an encyclopedia does not replace the original texts.

The creation and updating of databases is what bioinformatics is all about. Bioinformatics is simply the analysis of biological information using computers. As protein and DNA sequences began to accumulate, the question of how to stock them and compare them arose. Computers were the answer. And that is how bioinformatics began, 15 years after the first protein – insulin – had been sequenced in

1953. The first tools to be developed were analysis programs that could compare proteins from different species, thereby enabling the study of evolution.

Fig.1 *Prof. Amos Bairoch*

But in those days, bioinformatics was not called…bioinformatics; the word was coined barely 15 years ago. You spoke of sequence analysis. Consider a protein as a succession of characters that can be analyzed much as a text can be, i.e. searching for letters that are repeated or, as in the case of a protein, characteristic regions associated with certain functions. Also recognizable are zones that act like so many signposts directing the protein to a precise location in the cell.

Bioinformatics then turned to other types of analysis such as the shape of a protein, or in other words its three dimensional structure.

How did you become a bioinformatician?

As an adolescent, I was interested in science fiction, astronomy and extraterrestrial life. I had read a lot about how to detect life, its emergence and evolution, and the making of the very first amino acids. So naturally enough, I trained to become a biochemist at the University of Geneva, because if extraterrestrial life is one day discovered, biochemistry is undoubtedly the discipline that will enable us to study it.

I was also a computer enthusiast. The man who introduced me to the world of bioinformatics – although indirectly – was Robin Offord. Robin is an all-time great in biochemistry, who worked with Sanger on the first protein sequencing. I met him whilst on a summer job at an institute for diabetics in Geneva, and I had suggested programming software that would enable the analysis of the laboratory's experimental results. Then, as I was starting my second year of biochemistry, I offered to create a program that could assemble the different fragments which resulted from the sequencing of a protein, a thing he had been doing manually until then.

I carried on with this project in Robin's own laboratory while preparing for my Master's degree in biochemistry. He had suggested I use a mass spectrometer for analyzing proteins. This apparatus analyzes fragments of a protein once it has been sliced up in a test tube. But this particular spectrometer remained out of use throughout most of my degree! And, as a consequence, I created various programs myself.

In the end, it was my keen interest in both computers and proteins that led me into the world of bioinformatics. Proteins are fascinating in that they act on a cell or the organism, whereas DNA is merely a support for genetic information.

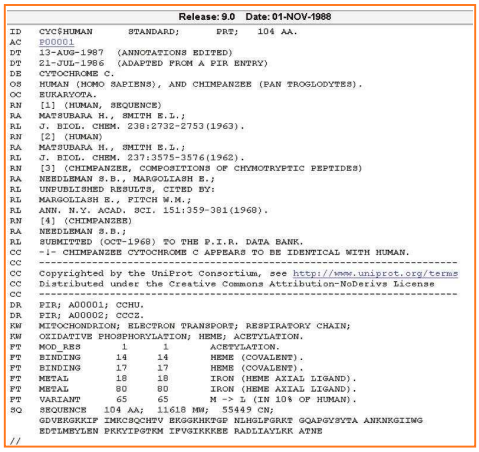
Where did you get the idea to create a database?

Before the advent of computers, protein sequences were listed in a book written by Margaret Dayhoff, known as 'The Atlas of Protein Sequence and Structure'. Its first edition dates back to 1965 in the United States. Towards the end of the 70s, this repository became the first computerized database - NBRF/PIR or Protein Identification Resource – and its successive versions were available on tape.

While developing PC/Gene – a collection of protein sequence analysis programs – I needed to use the NBRF/PIR bank. But it was not well suited for use on a computer, so I undertook to convert it into a more convenient format, and modeled it on a DNA database, produced and maintained by EMBL – the European Molecular Biology Laboratory in Germany. We then increased the initial number of proteins found in the NBRF/PIR database by manual capture of sequences taken from scientific papers.

Fig.2. *Prof. Bairoch's first interests: astronomy and extraterrestrial life. (Nebula of Orion).*

Whilst reformatting NBRF/PIR, I noted a certain number of errors, omissions and bibliographic disparities that I pointed out to those who were responsible for the database. But I never got a response. I even had the opportunity to meet them at a conference in the United States in 1984, where my remarks left them unmoved.



```
                    Release: 9.0   Date: 01-NOV-1988
ID   CYCSHUMAN       STANDARD;       PRT;   104 AA.
AC   P00001
DT   13-AUG-1987  (ANNOTATIONS EDITED)
DT   21-JUL-1986  (ADAPTED FROM A PIR ENTRY)
DE   CYTOCHROME C.
OS   HUMAN (HOMO SAPIENS), AND CHIMPANZEE (PAN TROGLODYTES).
OC   EUKARYOTA.
RN   [1] (HUMAN, SEQUENCE)
RA   MATSUBARA H., SMITH E.L.;
RL   J. BIOL. CHEM. 238:2732-2753(1963).
RN   [2] (HUMAN)
RA   MATSUBARA H., SMITH E.L.;
RL   J. BIOL. CHEM. 237:3575-3576(1962).
RN   [3] (CHIMPANZEE, COMPOSITIONS OF CHYMOTRYPTIC PEPTIDES)
RA   NEEDLEMAN S.B., MARGOLIASH E.;
RL   UNPUBLISHED RESULTS, CITED BY:
RL   MARGOLIASH E., FITCH W.M.;
RL   ANN. N.Y. ACAD. SCI. 151:359-381(1968).
RN   [4] (CHIMPANZEE)
RA   NEEDLEMAN S.B.;
RL   SUBMITTED (OCT-1968) TO THE P.I.R. DATA BANK.
CC   -!- CHIMPANZEE CYTOCHROME C APPEARS TO BE IDENTICAL WITH HUMAN.
CC   ---------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ---------------------------------------------------------------------
DR   PIR; A00001; CCHU.
DR   PIR; A00002; CCCZ.
KW   MITOCHONDRION; ELECTRON TRANSPORT; RESPIRATORY CHAIN;
KW   OXIDATIVE PHOSPHORYLATION; HEME; ACETYLATION.
FT   MOD_RES      1      1       ACETYLATION.
FT   BINDING      14     14      HEME (COVALENT).
FT   BINDING      17     17      HEME (COVALENT).
FT   METAL        18     18      IRON (HEME AXIAL LIGAND).
FT   METAL        80     80      IRON (HEME AXIAL LIGAND).
FT   VARIANT      65     65      M -> L (IN 10% OF HUMAN).
SQ   SEQUENCE   104 AA;  11618 MW;  55449 CN;
     GDVEKGKKIF IMKCSQCHTV EKGGKHKTGP NLHGLFGRKT GQAPGYSYTA ANKNKGIIWG
     EDTLMEYLEN PKKYIPGTKM IFVGIKKKEE RADLIAYLKK ATNE
//
```

Fig.3 *Swiss Prot format at its beginnings*

So I began to make corrections. When we sold the PC/Gene software, we also supplied the corrected version of NBRF/PIR. Since it was adapted to computers, the new version appealed to users very quickly. And – in order to satisfy an obvious demand – I decided to create a database. And called it Swiss-Prot – the first version of which appeared in 1986. As I wanted to complete my thesis and continue developing PC/Gene, I made an agreement with EMBL that, from the end of 1986 on, they were to be solely in charge of the production and updating of Swiss Prot.

But the future decided otherwise! Protein data never ceased to pour in. New staff had to be selected and trained. Despite the help of Rolf Apweiler – who is currently the director of the European Bioinformatics Institute (EBI) in Cambridge – Brigitte Boeckmann from the EMBL, Serenella Ferro and Jean-Pierre Patthey in Geneva, the size of staff was still too small to manage the volume of incoming data. So in the end, I decided to continue to apply myself to the evolution of Swiss-Prot!

Why did you wish to have free access to Swiss-Prot on the Internet?

When Swiss-Prot started, there was no Internet. The various versions were distributed on magnetic tapes for which a delivery charge was made. With the advent of the Internet, its long distance exchanges, and then the web, the database became totally free of charge so that those interested worldwide could benefit.

Who are the personalities who have marked you most in your 'bioinformatics career'?

Robin Offord, whom I met while preparing my degree at University and who took me under his wing for my masters and then my thesis. He left me free to choose and buy the computer I needed. And then he let me commercialize the software packages of PC/Gene and exploit the licenses so that more staff could be appointed to develop PC/Gene and Swiss-Prot.

But, above all, Robin relentlessly fought the computer scientists who were then in favor of computers and terminals the size of large cupboards. All were against microcomputers, which they thought fit only for computer games such as "space invaders", and believed they would never be of use to science… From 1982 to 1984, the battle was fierce! Although the large computers were very powerful, they were not adapted to our purposes; the results, for instance, could not be seen directly on the screen. But since the price of microcomputers was inordinately high at the time, the committees vetoed any order to buy… It was a great victory when Robin obtained the purchase of the first Apple II microcomputer in Switzerland – and it was for me…

Fig.4 *Robin Offord and Jean-Michel Claverie*

The second person to whom I owe a lot is Jean-Michel Claverie. I met him at an EMBL meeting, and we very quickly became friends. He was working at the Institut Pasteur in Paris with a young researcher, Lydie Bougueleret, who has since become my "right hand" at Swiss-Prot! Jean-Michel was at the head of the Scientific Unit of Informatics at the Institut Pasteur and, together, we planned to set up a European database. Unfortunately, we were unable to gather the necessary funds to do so. Later, I was to ask him to be my tutor for my thesis in bioinformatics, while

Robin was my tutor in biochemistry. Jean-Michel accepted and I took the train up to Paris regularly to see him.

What kind of difficulties did you come up against when setting up Swiss-Prot?

As I mentioned earlier, the main obstacle was the unwillingness of the computer scientists of the time to use microcomputers.

The other great difficulty was finding private funds. In those days, no one knew how to finance databases. As an illustration, the project that I had submitted with Jean-Michel Claverie rested on the appointment of one person only. And we were refused European funds on the basis that one person was one too many! As for the Swiss national fund, it supported research projects and not infrastructures. And, in effect, a database resembles more a long-term infrastructure. We were faced with a kind of inertia and it was only after the 1996 crisis, that things started to move.

What happened in 1996?

The Swiss national fund had underwritten the financing of 2 to 3 positions for a period of two years. They suggested we apply to European funds for further sponsoring. If our project were to be accepted, they undertook to release more funds.

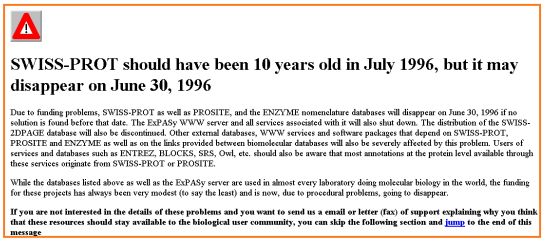
So we submitted a dossier, together with EMBL who was already giving us financial support. But our application was refused. The European Commission recognized the scientific value of our project but was opposed to the idea of supplementing the financial aid given by Switzerland and the EMBL. In fact, the EMBL and the Swiss national fund were both counting on the Commission's support to renew their contract with us. A case of the dog chasing its tail!

Graham Cameron, the then director of the EMBL bank, and I disputed the decision taken in Brussels. Unfortunately, it was impossible to retract and we were invited to renew our application the following year. The problem was we could not afford to run Swiss Prot and its staff for so long! We only had two short months to find a solution or to shut the shop!

How was Swiss-Prot saved?

In May 1996, we sent out an appeal on ExPASy, Swiss-Prot's web server. The same day, there was an influx of e-mails. In all, we received over 1'000 e-mails and letters of support, which unleashed a chain reaction not only in the local press, but also the international scientific press – Nature and Science – and the European Union.

In Switzerland, huge pressure was put upon the Federal Parliament. Ruth Dreyfus, who was then Minister of Science, had to give her pledge to take the required steps to save Swiss-Prot as soon as possible. Concomitantly, Guy-Olivier Segond, the Minister of Health in Geneva, released funds in order to pay Swiss-Prot's staff to the end of the year hoping there would be a Federal solution by then. And towards the end of 1996, the Swiss national fund paid an emergency sum that covered two years in the hope that a long-term solution would be found.



SWISS-PROT should have been 10 years old in July 1996, but it may disappear on June 30, 1996

Due to funding problems, SWISS-PROT as well as PROSITE, and the ENZYME nomenclature databases will disappear on June 30, 1996 if no solution is found before that date. The ExPASy WWW server and all services associated with it will also shut down. The distribution of the SWISS-2DPAGE database will also be discontinued. Other external databases, WWW services and software packages that depend on SWISS-PROT, PROSITE and ENZYME as well as on the links provided between biomolecular databases will also be severely affected by this problem. Users of services and databases such as ENTREZ, BLOCKS, SRS, Owl, etc. should also be aware that most annotations at the protein level available through these services originate from SWISS-PROT or PROSITE.

While the databases listed above as well as the ExPASy server are used in almost every laboratory doing molecular biology in the world, the funding for these projects has always been very modest (to say the least) and is now, due to procedural problems, going to disappear.

If you are not interested in the details of these problems and you want to send us a email or letter (fax) of support explaining why you think that these resources should stay available to the biological user community, you can skip the following section and jump to the end of this message

Fig.5 *Swiss-Prot's existence is threatened: appeal sent out on ExPASy in 1996.*

The long-term solution involved the creation of the Swiss Institute of Bioinformatics. In a little less than two years, the legal framework was determined, appeals to raise funds were sent out, the statutes of the Institute set down and a founding committee formed. The Institute was the beneficiary of article 16 of the Federal Constitution that authorizes the Confederation to finance research insofar as it is non profit -making and of national interest. The Institute's birth certificate was signed in April 1998.

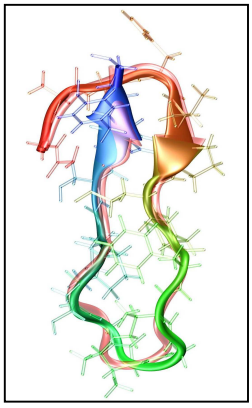
Fig.6 *Amos Bairoch sporting the logos of the Swiss Institute of Bioinformatics and Swiss-Prot.*

The Swiss Institute of Bioinformatics is a confederation of research groups "without walls". These groups are based in several towns: first in Geneva, followed by Lausanne in 1998, and later Basel and Zurich. The Institute thrives on a mixed system of financing, where both the Confederation and local universities contribute.

How is the database updated?

Annotators update Swiss-Prot. In the same way as a text is annotated by adding comments, annotators introduce explanatory and informative notes on the structure and function of a protein. They do this by extracting information from scientific papers, but the starting point of their work is the retrieval of protein sequences.

20 years ago, researchers studied the function of a protein before sequencing it. Today, things are done the other way round. Groups all over the world are sequencing complete genomes, i.e. all the genetic information in a species' DNA. DNA carries the genes that can be described as recipes for the synthesis of proteins. Genomes carry the information for the production of hundreds to thousands of proteins. And all these virtual proteins – whose existence has not yet been proved experimentally – are poured continuously into the databases. As a consequence, there are scores of protein sequences which need to be analyzed and no experimental information on a protein's structure or function. An annotators' job is therefore that much more important today.



Fabrice David, ISB Genève

Fig.7 *Besides databases, the field of bioinformatics also helps to study protein structure (here: microcin J25).*

A direct consequence of this massive influx of new sequences was finding a way to stock them. In 1996, a new database – TrEMBL – was created and run by EBI, the European Bioinformatics Institute in Hinxton. The DNA sequences that reached us were directly converted into protein sequences, and then deleted. Currently however, the DNA sequences are kept in the TrEMBL database before analysis by a Swiss-Prot annotator.

Besides accumulating protein sequences, researchers also accumulate experimental results. So an annotator's work also involves following the scientific literature very closely and updating continually a given protein. A database is not static: as new discoveries are made, the information must follow.

What is Swiss-Prot's main asset?

Quality! At Swiss-Prot, the quality of information supersedes the quantity of proteins entered into the database. The information given on each and every protein must be precise. To avoid all possible errors – from typos to scientific imprecision – we read and verify all material at several levels. Our aim is to be "as precise as a Swiss clock"!…

Is any of the information relative to proteins of medical interest?

For human proteins, yes. Swiss-Prot props medical research in the study of genetic diseases. Many genes are liable to change, this is called a mutation. Mutations arise when an amino acid is modified in a protein's sequence, for instance, and such modifications can be the cause of a disease. We list these mutations and sum up the consequences they may have in an organism.

Besides human proteins, bacterial proteins can be of great medical interest since the information we stock in Swiss-Prot can help to develop novel antibiotics.

What great changes has Swiss-Prot known?

Swiss-Prot has witnessed two major changes.

No doubt the most important is free access to scientific publications on the Internet. Only five or six years ago, we were still photocopying numerous articles in the library! This has obviously made life much easier for the annotators.

The second change concerns the specialization of annotators. In the beginning, an annotator was a generalist and worked on proteins from any given species – bacteria to humans. Over the years, the

amount of information we were receiving was not only massive but getting more and more complex, and the best answer to that was to specialize annotation. Currently, for example, a virologist would be appointed to work on viral proteins.

Within the Swiss-Prot database today, is there a species for which all the proteins are represented?

There are a few. Viruses are an example, simply because they have few proteins. The AIDS virus contains only 10 different proteins. The proteins of the best known bacteria – *Escherichia coli* – are also all represented in the database, and there are over 4'000! By the end of 2006, we should have managed to list all the proteins of Baker's yeast – *Saccharomyces cerevisiae* – 6'000 proteins in all.

And what of human proteins?

The sequencing of the human genome has been completed. The number of human genes is estimated to be between 20 to 25'000 and the number of proteins around 500'000. This estimation does not include antibodies, which are proteins synthesized by the immune system for a specific purpose and are no doubt produced by the millions.

When will Swiss-Prot be able to account for all the human proteins? We hope to have the total sum listed – at least in summary fashion – in two years' time…

What is Swiss-Prot's scope now?

In 2002, the National Health Institutes of the United States - NIH - invited bids to develop a unified database of proteins. The Swiss Institute of Bioinformatics, the European Bioinformatics Institute and the American database PIR decided to combine in a common project. Together we created a consortium called UniProt whose aim will be to continue the development of databases by maintaining the collaboration of these three Institutes. The project was accepted at the end of 2002 for a period of 3 years and was renewed a second time.

What future do you foresee for Swiss-Prot?

It is difficult to foresee where Swiss-Prot is going. Something unexpected has always turned up. At all events, there will always be more and more information, more and more sequences and more and more work. In the space of twenty years, the number of proteins entered in the databank has risen from 4'000 to 230'000, and the staff from one to over 70. Some annotators have been with us for 15 years and contributed their knowledge and competence. However, despite the quality of everyone's work here, we will certainly be needing more staff in the future to ensure the continuation of Swiss-Prot.

Perhaps other countries will join us later. One group in Brazil is already helping us and a Japanese group will soon be joining UniProt. It is one way to strengthen the finances and enlarge the scope of annotators…surely a combination which should help to guarantee the future of Swiss-Prot.

Interview by Séverine Altairac*

*Translation: Geneviève Baillie

(1) "First", by Hadrien Dussoix. The painting represents a small portion of the very first protein sequence that was entered into Swiss-Prot: human cytochrome c.

For further information

On the Internet:

- Insulin : protein of the 20th century
  http://www.expasy.org/spotlight/back_issues/sptlt009.shtml
- La Bioinformatique : une enquête à l'échelle du Vivant
  http://www.expasy.org/prolune/annexes/prolunea04_bioinformatique.shtml
- Bioinformatique : chronique d'une révolution annoncée
  http://www.expasy.org/prolune/annexes/prolunea05_bio-NZZPDF.shtml

A little more advanced:

- Bairoch A., "Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times!", Bioinformatics 16:48-64(2000) PMID: 10812477

Date of publication: August 9, 2006
Date of translation: September 13, 2006
