

Report on

In-Silico Analysis of Proteins

Celebrating the 20th anniversary of Swiss-Prot

Christophe Dessimoz (cdessimoz @ inf.ethz.ch)

Zurich, August 2006

We celebrated the 20th Anniversary of the Swiss-Prot database in the context of a conference in Fortaleza, Brazil. During the 5 days, we heard presentations from about 50 speakers, all somehow related to the history of Swiss-Prot database, and all of very impressive background for a young scientist such as myself. In the course of its 20 years of existence, Swiss-Prot has not only established itself as the "golden standard" of protein databases, but has indeed also collaborated with much of today's bioinformatics establishment.

I have organized the present report in two main topics that were addressed throughout the meeting from different speakers and from different perspectives.

1. The Data Overflow

The steep increase in biological data that results from the emergence of high-throughput methods was obviously a very general concern. Strategies to cope with the high volume were outlined in several talks, in particular the ones of Amos Bairoch (Swiss-Prot), Barend Mons (Knewco Inc.), Rolf Apweiler (EBI), Manuel Peitsch (Novartis) and Lionel Binns (HP).

Ideally, everybody seemed to agree that the data should flow into databases directly from the source, and not go through scientific journals, which rarely capture data in consistent and structured formats. Furthermore, the data should

be "pushed" by the researchers rather than "pulled" by database curators, the former having both the advantages of scalability and distributed costs. Such practice could be enforced through funding agencies requirements.

Large amounts of data commonly have to be processed and analyzed by automated procedures and algorithms, which greatly benefit from structured data. The usage of controlled vocabularies and ontologies gain momentum: the Gene Ontology consortium is now involved in the annotation of about 20 genomes, Swiss-Prot is now using controlled vocabularies in all areas of annotation. In parallel, some scientific publications, such as the BMC journals, now offer content in XML format to ease text mining and other automated procedures on manuscripts.

The astonishing success of Wikipedia in the last few years has obviously inspired more than one data repository to consider community annotation as a possible remedy against the explosion of data. For instance, Swiss-Prot has an initiative called "adopt a gene", the EcoGene database accepts user contributions, and Barend Mons is working on a joint wiki project with the Wikimedia foundation.

Quite often, large amounts of data are crucial for the success of an analysis. But Binns came up with a word of caution to this respect: faster computers and more data can be difficult to manage, and are no substitute for careful thinking or adequate algorithms.

2. Evolution of proteins: primary sequence

Understanding of the evolution of proteins even at the primary sequence remains a challenge. Des Higgins and Cédric Notredame reported the latest developments on the front of multiple sequence alignment. The current software packages are improving rapidly. Higgins pointed out the fact that his package

Clustal W still is very successful despite the fact that other programs have superseded it for some years already. That shows that the success of bioinformatics tool do not solely depend on their performances. On the other hand, Notredame showed how his new tool M-Coffee integrates the results of different approaches and software to improve alignments.

Correct sequence alignment can be even a challenge for pairs of proteins. At low level of similarity, it can be difficult to detect homology. To overcome this problem, alignments are sometimes performed using structure information. William Pearson (Fasta) questioned the accuracy of the statistics of such procedures. In particular, an argument was the apparent frequent occurrence of convergence at the level of structure, a phenomenon that almost never happens at the sequence level.

The two talks of Antoine Danchin clearly demonstrated that we have yet much to understand about the primary sequence of proteins. In one study, he identified several biases in bacterial genomes strong enough to classify bacteria in termo-, meso- and psychophile just on the basis of amino-acid composition. More generally, Danchin stressed the effect of physical and environmental constraints on genes.

The present report only covers a modest part of the whole conference (in particular, it cannot convey the strong emotional dimension of the meeting), but it will hopefully provide the reader a survey of some current subjects of interest in bioinformatics.