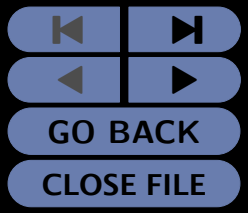
BIOINFORMATICS
Electronic edition http://www.bioinformatics.oupjournals.org

VOLUME 16
NUMBER 5
MAY 2000
PAGES 411–411

From data to knowledge

Francois Rechenmann

INRIA Rhône-Alpes, Grenoble, France



Editorial

It has become an overused commonplace statement that the volume of data in molecular biology is increasing at an exponential rate. And the same recurrent arguments are invoked to support this assertion: the tremendous progress in sequencing technology, illustrated by the doubling time of GenBank; the development of DNA chips which announces an overwhelming flow of expression data; and so on.

However, a quick glance at several other scientific areas reveals that it is not so much the volume of data which characterizes biology, as their diversity and heterogeneity. They are indeed related to various entities, concern different organisms, and originate from multiple experimental sources.

While local computational methods have been developed for handling the different classes of data, the overall process, through which these data will be converted to more synthetic knowledge, is expected to be highly exploratory, progressive, and recursive since it makes use of previously acquired knowledge. Life sciences are not physical sciences and it is very unlikely that these data could ever be synthesized into a set of mathematical equations through a straightforward inductive process. Then, how can biological knowledge be expressed as it is incrementally extracted from these data?

At the present time, natural language, in its written form, is the usual support for expressing knowledge, as attested by the ever-increasing volume of scientific publications. While texts are clearly adequate for human communication—as long as their flow does not overly exceed the reading capacity—they suffer from

several drawbacks regarding computer processing, even when they are stored in their electronic form and efficiently organized in document databases. Several research projects are thus attempting to develop computer programs able to automatically extract data and knowledge from texts, relying on statistical or linguistic techniques. These efforts are motivated by the conviction that there is more information encrypted in the scientific literature than available from databases.

The situation is even more critical when knowledge is expressed as drawings or pictures. How many networks of molecular interactions, for example, have been described solely as drawings of directed graphs, the semantics of which has not always been made explicit?

The natural question is then: why not express biological knowledge directly in a form which could be as efficiently processed by a computer as easily understood by humans?

Knowledge representation is a very active and very mature area of computer science, which has produced expressive knowledge models together with powerful consistency checking and inference mechanisms. Most of these models are object oriented, but the most expressive ones associate objects and relationships. Several fruitful experiences of knowledge base building have shown that these models are adequate for representing the various categories of biological knowledge: the entities and their relationships (such as genes and their components, their products and their interactions, and also the different types of maps, etc.), the dynamic processes in which they are involved, and

even the description of the investigation and computing methods which have been used to characterize them.

The technical advantages of knowledge modeling are obvious. Knowledge bases can be automatically checked for consistency; they support inference mechanisms which derive data which have not been explicitly stored; they also offer extensive request and navigation facilities. However, the most immediate benefit of knowledge base design lies in the modeling process itself, through the effort of explicitation, organization and structuration of the knowledge it requires.

Less than fifteen years ago, an important mutation took place in the way sequence data were published: the sequences were to be deposited directly in databases, while the description of the experimental processes and of the results were submitted for classical publication. A similar, but undoubtedly more drastic, mutation could now be advocated: a simultaneous submission of a formal description of knowledge, together with the necessary explanations and comments in a textual form. Indeed, every piece of knowledge does not need to, or cannot, be formally expressed. Knowledge bases should therefore be linked with (hyper-)texts and then be used as a very structured and powerful index of these texts.

There are clearly several obstacles to the direct encoding of biological knowledge, the first of which is certainly the lack of formal knowledge literacy and the absence of a widely accepted standard for knowledge modeling. Expressive models, implementing an easy-to-understand, although formal,

semantics, will have thus to be appropriately designed. These models will be used through powerful person–machine interfaces which should take advantage of adequate graphical metaphors. This could constitute the essence of a challenging common research project for the bioinformatics community.

The crucial role of computer science is now unanimously recognized for the management and the analysis of biological data; through knowledge modeling, it could be brought to play a role in life sciences similar to the role mathematics plays in physical sciences.

François Rechenmann
INRIA Rhône-Alpes
Grenoble, France