International Cooperation on Structural Genomics

a key issue in post-genomic research

di Antonio Rosato

Complete genome sequences have recently become available for several organisms. Structural Genomics is a new field of research aimed at the determination of the 3D structures of all proteins encoded by the genomes. This effort will contribute to understanding the molecular basis of Life, and will be crucial for many applications. However, it also raises important science policy issues. These have been recently debated at a meeting promoted by the OECD.

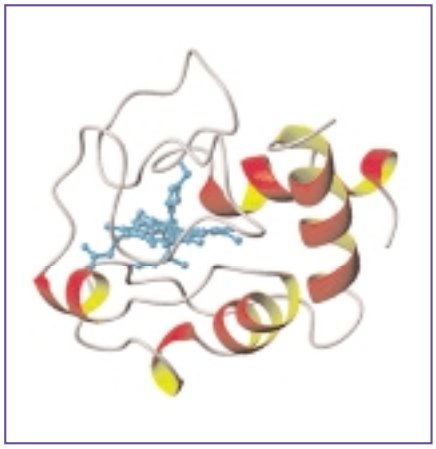

Figure 1 - Solution structure of horse heart cytochrome c, solved by NMR (L. Banci, I. Bertini et al., Biochemistry, 1997, **36,** *9867)*

Since 1995, when for the first time the complete genome sequence of a living organism (the bacterium *Haemophilus influenzae*) was determined and made publicly available, genome sequencing projects have afforded an outstanding wealth of genetic information. Just recently, the complete human genome sequence has been determined. All genome sequences currently available can be browsed at http://www.ncbi.nlm.nih.gov. At present, the complete sequences for 9 *archaea*, 36 *bacteria*, 4 *eukaryota* are accessible, together with a number of partial sequences. Knowing the genome sequence means knowing the primary sequences of all the proteins that an organism can produce in his life. This is a very precious resource in its own. However, it tells us very little about the function of each protein and does not tell us anything about how this function is carried out. Determining the function of each protein encoded by the genome and understanding how it performs its function is crucial to be able to fully exploit genomic information to improve the quality of life for mankind (e.g. by developing new drugs). These two problems constitute the very heart of post-genomic research (also called post-genomics). A particular field of research within post-genomics is Structural Genomics.

The availability of complete genome sequences has prompted the proposition that the corresponding structural information should be obtained for very large, complete sets of proteins. This bold concept has been termed "Structural Genomics". Protein structure determination consists of a series of complex and difficult steps, each one of which requires specialized expertise and the availability of the appropriate equipment and resources (Figure 1). It typically takes *from several weeks to a few months* to determine the structure of a protein of medium size, starting from the simple knowledge of the corresponding gene sequence (provided by genome sequencing projects). The work is usually done by a Ph.D.-level scientist, using sophisticated experimental schemes and instrumentation, high-performance computers, and specialized software. These facts can give an idea of the challenge posed by Structural Genomics, given that the genomes of even simple organisms encode thousands of distinct proteins (for humans, this number is of the order of 30,000). Thus, any Structural Genomics project should be operated in a high-throughput mode. In addition, completeness of coverage relative to a base of genomic information is also required, in order to optimize the utility and size of any Structural Genomics project. There are many ways of defining the "completeness" criterion; this can be illustrated by listing some sample proposals and ideas for projects that have been brought forward in the United States:

- tuberculosis drug targets;
- cancer-related proteins;
- signal transduction proteins;
- eukariotic model organisms;
- yeast proteins;
- new protein families.

For any strategy, the number of pro-

Antonio Rosato, Centro di Risonanze Magnetiche - Università di Firenze - Via L. Sacconi 6 - 50019 Sesto Fiorentino - rosato@cerm.unifi.it.

Ivano Bertini (Magnetic Resonance Center, University of Florence), organizer of the OECD Workshop on "International Cooperation in Structural Genomics" (Florence, June 8-9, 2000)

Group picture of the delegates to the Workshop

tein structures that actually must be determined (via X-ray crystallography or NMR spectrometry) depends on the desired accuracy of the output structures since computer modelling can, to some extent, substitute for physical measurements. An important problem is that a large fraction of proteins inevitably prove to be particularly difficult to analyze, e.g. cell membrane proteins, which may constitute approximately one-third of all proteins in a typical genome. Besides its scientific and technical challenges, Structural Genomics involves unique infrastructure, funding, organizational, legal, and international issues which deserve the special attention of science policymakers. For this reason the Global Science Forum (GSF) of the Organization for Economic Cooperation and Development (OECD) agreed in its second meeting, which was held in February 2000, to the proposal of the Delegation of Italy of organizing a Workshop on "International Cooperation on Structural Genomics".
The Workshop took place at the University of Florence, Italy on 7-9 June, 2000 organized by Ivano Bertini (Magnetic Resonance Center, University of Florence). The meeting was attended by 34 delegates representing thirteen member countries, the European Commission, invited observers from the European Science Foundation and the Wellcome Trust and members of the Oecd Secretariat (Figure 2). Three experts made presentations: John Moult (University of Maryland), John Markley (University of Wisconsin) and Joseph Straus (Max Planck Institute for Foreign & International Patent Copyright & Competition Law, Munich). The Workshop opened with a welcome by Undersecretary Vicenzo Sica on behalf of the Italian government.
The meeting started by reviewing the various governmental programs relative to Structural Genomics. It appeared clear that, in principle, this topic was of interest to all represented countries, although with varying (and sometimes unclear) priority. At the time however, only two countries (Japan and the United States) and one charitable foundation (the UK Wellcome Trust) were pursuing Structural Genomics in an active way, with dedicated programs and significant expenditures. The situation has evolved significantly meanwhile and now projects have been started in different European countries; in addition, the European Commission has issued a call for expressions of interest in the field of "Genomics and human health". Structural Genomics will be further addressed by the European Commission in its 6th Framework Program.
One key issue of Structural Genomics is the availability of infrastructures with the necessary equipment and expertise for protein structure determination. Any effort to systematically determine the structures of thousands of proteins will not be feasible until the various steps can be speeded up and automated. Economies of scale would certainly have to be exploited; thus, dedicated laboratories would need to be established to focus on specific tasks such as expression of proteins from genes, or acquisition of experimental data for structure determination. As mentioned, the two leading techniques for structure determination are X-ray crystallography and NMR spectroscopy. X-ray beamlines are associated with electron synchrotrons whose costs range from 100M dollars to over 1B dollars. Access to these facilities depends on a variety of factors (often involving a national financial contribution to the construction and/or operating costs of the facilities). NMR facilities are also established world-wide and mostly financed at the national level. Thus, national and regional programs that support researchers seeking access to synchrotrons and NMR facilities are particularly valuable (e.g. the European Commission's "Training and Mobility of Researchers" program). In this context, exchange of access among advanced research infrastruc-

Figure 2 - 800 MHz AVANCE NMR spectrometer, installed at the Magnetic Resonance Center of the University of Florence

tures is key to promoting integrated methodological and scientific development.
Structural Biology and Structural Genomics projects will generate large amounts of data that must be made available to researchers world-wide. This raises three issues:

a. permanent, stable funding for databases and the development of bioinformatics tools is essential, but has not always been adequately provided. This need is shared by many areas in the life sciences and it deserves global attention (since databases are increasingly being implemented on an international basis).
b. Structural and functional information needs to be linked and incorporated into existing protein databases, which were developed and optimised for receiving the results of genome sequencing projects. Protocols for cloning, expression, crystallization (where applicable) and structure determination should also be made available. There are a number of technical problems in this area, linked to the different characters of the datasets. These issues are being addressed in the scientific community, but the efforts (and the implementation of their outcomes) need the support of funding agencies.
c. There is a need for greater information-sharing about the structural work that is being done world-wide. Inevitably, there is some duplication of effort, especially for proteins that may be targets for new drugs, but there is some indication that some unnecessary duplication may be occurring as well. As in case b) above, this issue is being addressed by scientists.

Finally, it is important to mention the problem of Intellectual Property Rights. Genomic sequencing projects have raised several world-wide controversial discussions on issues such as gene patentability, etc. Post-genomic research is likely to pose similar problems in the future, which will make the courts busy for some time before a consistent set of rules emerges. A key requirement for obtaining a patent is the demonstrated utility of the invention. This is clearly relevant to Structural Biology whose ultimate goal is to reveal the connections between genes, protein structure, and biological function. To an extent, IPR represents an obstacle to the advance of Structural Genomics, since many researchers may be (understandably) reluctant to put potentially lucrative information into the public domain. Thus, it is not clear whether the standards that were agreed-to by scientists and institutions that participated in the Human Genome Project (which put great emphasis on the rapid release of raw data) would be transferable to Structural Genomics. The differences between patent regulations in Europe, the United States and Japan, e.g., with regard to the “grace period” that applies between releasing results and applying for patent protection, complicate matters still further.
A consensus report on the meeting has been approved by the Oecd bureau and is available from the Oecd web site (http://www.oecd.org/dsti/sti/s_t/ms/prod/cont-e.htm).