

OECD Global Science Forum

Workshop on Structural Genomics

Florence, June 8/9, 2000

Final Report to the Global Science Forum

1. Introduction

This is a report on the OECD Global Science Forum Workshop on Structural Genomics that was held in Florence on June 8-9, 2000. It has been reviewed, revised, and cleared by the national delegates to the workshop. The convening of the workshop was authorised by delegates to the 2nd meeting of the Global Science Forum, based on a proposal from the Delegation of Italy.

The workshop was attended by 34 delegates representing thirteen OECD Member countries¹, the European Commission, observers from the European Science Foundation and the Wellcome Trust, three invited distinguished experts², and members of the OECD Secretariat.

The workshop agenda can be found at the end of this report, along with a list of participants. A significant portion of the meeting was devoted to a “tour de table” in which delegates and observers presented information about plans, priorities, programmes, levels of government support, decision-making and funding, foresight exercises, trends in governmental deliberations, and international collaborative programmes. Structural biology (and structural genomics, as defined below) are not yet characterised by an established network of government officials who meet frequently and are well acquainted. Therefore an immediate positive outcome was achieved through the establishment of personal contacts among the delegates.

2. “Structural Biology” and “Structural Genomics”

Much of modern biological science aims for an understanding of living organisms through the study of the properties and interactions of biomolecules such as proteins, lipids, nucleic acids (DNA and RNA), carbohydrates, and others. Knowledge of the three-dimensional physical structures of these molecules is a crucial part of the pursuit, since structure is closely linked to biological function. To cite just two examples: the control of cellular processes is often achieved through “lock-and-key” inter-actuation of signalling and receptor molecules; and the transport of chemicals within cells is often achieved by special-purpose molecules whose shape and physical-chemical properties are adapted to carry only certain specific types of “cargo” substances. Research on the links between structure and dynamics of biomolecules and their function has been pursued for many years, and is called “structural biology”.

The composition and configuration of one of the most important class of biomolecules - proteins - are uniquely determined by coded segments of DNA (“genes”). In particular, the recently-achieved availability of the full coded DNA sequences of certain organisms (“genomes”) inspired some scientists to propose that the corresponding structural information should be obtained for very large, complete sets of proteins, leading to a potential quantum leap in the understanding of the totality of the integrated functions

¹ Australia, Belgium, Canada, Denmark, Finland, France, Italy, Japan, Poland, Sweden, Switzerland, the United Kingdom and the United States.

² Prof. John Moulton of the University of Maryland, Prof. John Markley from the University of Wisconsin, and Prof. Joseph Straus from the Max Planck Institute for Foreign & International Patent Copyright & Competition Law in Munich

of the organisms. This bold concept (which is qualified below) has been termed “structural genomics”³. Because it is so new, and because its implementation would probably require a large, dedicated, co-ordinated international effort, structural genomics was the principal focus of the discussions at the Florence workshop.

Unlike structural genomics, structural biology is an established field that has been pursued vigorously by scientists, who, typically, need structural information in the course of targeted basic and applied research in very diverse areas: for example, the study of specific metabolic pathways, or investigations into particular diseases (with the concomitant search for new drugs and therapies). Two principal experimental methods have been used: X-ray crystallography (using high-intensity beams from electron synchrotrons) and nuclear magnetic resonance (NMR) (using large superconducting spectrometers with very high magnetic fields of 15 Tesla and above). Important information (especially regarding highly specialised molecular properties) can also be derived from mass spectrometry, electron microscopy and neutron scattering.

X-ray crystallography is based on the diffraction of electromagnetic radiation by periodically arranged (crystallised) atoms and molecules. Their separation is on the order of Ångstroms⁴, and thus the diffracted wavelength must be in the spectral range of x-rays. From the diffraction pattern it is possible to reconstruct the three dimensional structure of a protein molecule, by determining the relative positions of all its scattering atoms. To do this, the phase of the diffracted radiation must be measured. Several experimental approaches have been developed to tackle this problem. Synchrotrons constitute the most important X-ray sources because: *i*) they have very high brightness, which ultimately means very high sensitivity; *ii*) they allow the use of different X-ray wavelengths, which is the best strategy to date to solve the phase problem.

Structure determination by NMR is based on the study of the interactions of nuclei with a magnetic field, and among nuclei themselves. By irradiating nuclei in a high magnetic field with radiation in the spectral range of radio waves, transitions between nuclear energy states are induced. The exact frequency of transition of a particular nucleus depends on its chemical environment within the molecule. Thus, individual transition frequencies in NMR spectra can be assigned to individual nuclei in the molecule. Multidimensional NMR allows the detection of interactions between pairs of nuclei in macromolecules, as well as the reciprocal orientation of inter-nuclear vectors. From the quantitative or semi-quantitative measurement of a few thousands such interactions, and by using chemical information on bond lengths and angles, a three-dimensional structure can be calculated. Very high field magnets are needed because: *i*) the time needed to perform an experiment with a given signal-to-noise ratio decreases with the cube of the magnetic field; *ii*) higher magnetic fields generate a larger separation of resonance frequencies of individual nuclei.

The two methods described above are complementary. X-ray diffraction requires that a protein be crystallised (which is not always achievable), whereas NMR can be applied to proteins in solution. The maximum size of the protein molecules that can be analysed by means of NMR (currently on the order of 40,000 times the mass of a hydrogen atom) is considerably lower than that of X-ray diffraction.

Structural determination consists of a series of complex and difficult steps⁵, each one of which requires specialised expertise and the availability of the appropriate apparatus and resources. 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. The work must be done by a Ph.D.-level scientist,

³ Precise definitions of *structural biology* and *structural genomics*, as well as related terms such as *proteomics* and *functional genomics*, have not yet been universally adopted by the scientific community. The definitions herein are simply offered for the purposes of the workshop and of this report.

⁴ Ångstrom (Å) = one ten-billionth of a metre

⁵ The steps are: gene identification, gene expression, protein purification and characterization, protein crystallisation (for X-ray studies only), data collection, structure solution, refinement and analysis.

using sophisticated experimental schemes and instrumentation, high-performance computers, and specialised software. In this light, the challenge posed by the vision of structural genomics can be appreciated, given that the genomes of even simple organisms encode information for thousands of distinct proteins (for humans, this number probably exceeds 100,000). Any fully-fledged structural genomics project would have to be operated in a high-throughput mode, somewhat analogously to gene sequencing projects, where researchers concentrate on accumulating data as quickly as possible without (in most cases) pausing to analyse the significance and possible applications of that data. Besides these scientific and technical challenges, structural genomics involves unique infrastructure, funding, organisational, legal, and international issues which deserve the special attention of science policymakers, as described below.

3. Findings Regarding Structural Genomics

National and Regional Priority-setting

The “tour de table” revealed that while structural biology is a high priority field for all participating delegations, only two countries (Japan and the United States) are pursuing high-throughput structural genomics through dedicated programmes and significant expenditures. All delegates agreed that structural genomics and structural biology are not mutually exclusive, and that there will be a continuing need to provide support for projects that seek structural information as part of targeted biological research. However, policymakers in all countries have to consider whether, when, and to what extent they want to become involved in large-scale high-throughput structural genomics projects during the coming years. At the present time, the relative priority of structural genomics is not an issue on which all parties agree, and discussions will have to continue on national and regional levels.

Since 1998, the Science and Technology Agency of Japan has been providing significant funding for structural genomics (the “Protein Folds” and the “Structurome” projects) at RIKEN (<http://www.riken.go.jp>), while the Ministry of International Trade and Industry has begun funding structural genomics projects focussing on membrane proteins. The RIKEN research is closely linked to human, mouse and plant genome sequencing, as well as full-length cDNA, and functional genomics projects.

In the US, structural biology research has been supported by several agencies, including the National Institutes of Health (NIH), the Department of Energy (DOE), and the National Science Foundation (NSF). The success of these research programs has led to the development of several structural genomics projects by the DOE and the NIH. As the interest in this field has expanded, the National Institute of General Medical Sciences⁶ (NIGMS) has announced its “Protein Structure Initiative” to organise a co-operative, large-scale effort in this emergent field. This national program will fund research centres, each incorporating all of the experimental and computational tasks of structural genomics. These research centres will test strategies for high-throughput operations and serve as pilots for large-scale research networks of the future. The Institute plans to support approximately six such research centres, each costing about \$4 million annually. A separate grants program is designed to encourage methodology and technology development. Background information and the program details can be found on the NIGMS Web site at: <http://www.nih.gov/nigms/funding/psi.html>. The DOE “Microbial Cell Initiative” seeks a comprehensive understanding of the complete workings of a microbial cell, including the production and determination of the structure of all the proteins whose assembly instructions are contained in the model microbe's genes. In addition, DOE support of synchrotron facilities in the U.S. is crucial to structural genomics efforts.

⁶ NIGMS is one of the twenty-five institutes and centres of the National Institutes of Health (NIH).

The Wellcome Trust and NIGMS have initiated consultations on the scientific aspects of a potential international structural genomics initiative. They convened a meeting on April 4-6, 2000, bringing together prominent representatives of the structural genomics community. The principal organisers of this conference were present in Florence, and the consensus document from the gathering (<http://www.nigms.nih.gov/news/meetings/hinxton.html#agreed>) was distributed to the workshop participants. Policymakers in OECD countries may wish to follow closely the evolution of this effort, which is currently working through targeted task forces in preparation for a second meeting in the Spring of 2001.

The European Commission funded 63 structural biology projects⁷ in the Biotechnology programme of the 4th Framework Programme (1994-1998), and a pan-European co-ordination initiative⁸ was implemented in structural biology in order to improve synergy between the national research efforts. Industry took the initiative to set up SBIP - Structural Biology Industrial Platform⁹. In the 5th Framework Programme (1998-2002), structural genomics is included in the “*Cell Factory*” key action, and functional genomics is part of the Generic Activities of the “*Quality of Life and Management of Living resources*” programme¹⁰. Planning for the 6th Framework Programme is now under way.

The European Science Foundation is co-ordinating workshops on Protein Structure & Function with the aim of identifying areas where there is a need for co-operation between member states.

At the European level there has been little co-ordination and/or co-operation in structural genomics (as defined in Section 2 above). Given that large countries have a natural initial advantage in organising and implementing innovative large-scale projects, it may be desirable for European policymakers to strengthen their consultation processes. The current debate over the future of scientific research policy in Europe, (being conducted under the auspices of the European Commission, the European Science Foundation, the Council of Europe, Euroscience, and other bodies) provides a good opportunity for examining the future of structural genomics in Europe. The recent Communication from the European Commission “Towards a European Research Area”¹¹ provides an important focal point and context for the debate.

Target Selection (Scope of a Structural Genomics Project)

As it is being implemented in the US, structural genomics is a subset of structural biology, characterised by *high-throughput* operation, and *completeness* of coverage relative to a base of genomic information. The second characteristic is particularly important, since it can be used to optimise the utility and size of any potential 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
- Signal transduction proteins
- Yeast proteins
- Cancer-related proteins
- Eukaryotic model organisms
- New protein families

For any strategy, the number of proteins that actually have to be measured (via X-ray crystallography or NMR spectroscopy) depends on the desired accuracy of the output structures since, as described in below,

⁷ <http://www.cordis.lu/biotech/src/projects.htm>

⁸ <http://europa.eu.int/comm/research/biotech/biot-pg-pdf.html>

⁹ <http://www.sbiip.org/>

¹⁰ <http://www.cordis.lu/life/>

¹¹ Communication from the Commission to the Council, the European Parliament, the Economic and Social Committee and the Committee of the Regions, COM (2000)6, 18 January 2000
<http://europa.eu.int/comm/research/area.html>

computer modelling can, to some extent, substitute for physical measurements. In any case, the total effort needed is always hard to predict, since some proteins inevitably prove to be particularly difficult to analyse. The most notorious case are cell membrane proteins, which may constitute approximately one-third of all proteins in a typical genome. They are extremely difficult to crystallise - a prerequisite for successful X-ray analysis. Policymakers should note that the feasibility of any project that involves analysis of thousands of proteins requires the prior development of new techniques and instruments. The corresponding investments are being made at this time, most notably in the US and in Japan.

High-throughput Operating Mode and Infrastructure Needs

Any effort to systematically determine the structures of thousands of proteins will not be feasible until the component steps (enumerated in footnote 5) 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. Clearly, there would be a potential for specialisation under the terms of any internationally co-ordinated structural genomics project. Optimisation of existing or new X-ray beamlines would also be required, as well as strengthening of existing NMR facilities.

X-ray beamlines are associated with electron synchrotrons whose costs range from 100M\$ to over 1B\$. 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 programmes that support researchers seeking access to synchrotrons and NMR facilities are particularly valuable (the best-known of these is the European Commission's "Training and Mobility of Researchers" programme). In this context, exchange of access among advanced research infrastructures is key to promoting integrated methodological and scientific development.

Modelling and Computation

While DNA has a relatively simple physical structure (the familiar double helix), the proteins that it encodes exhibit a vast variety of complex shapes, each suited to a particular function in an organism. Proteins found in living matter have the astonishing characteristic that the structure of each one is completely and uniquely determined by its genetic identity; that is, a given sequence of DNA nucleotides (A,C,G or T) and the corresponding sequence of amino acids that make up the protein, can give rise to a molecule of only one characteristic physical configuration. This property, in turn, leads to the hope that, at some time in the future, protein structures will be determined computationally from genomic sequences alone, thus saving enormous amounts of time, money and experimental effort. Progress towards this goal has proven to be extremely difficult due to the very large numbers of atoms that make up large biomolecules, and the complex, non-linear forces among all of these atoms. Nonetheless, computation has already proven to be a valuable tool for deriving structures of proteins whose genetic sequences differ only partially from those of proteins with known (measured) structures. The closer the genetic sequences, the easier it is to accurately calculate derived structures. For any potential structural genomics project, this means that computation and modelling are a tool for reducing the size of the costly and time-consuming experimental programme. From a policy perspective, it seems prudent to pursue experimental *and* computational projects in parallel. The empirical and theoretical approaches complement and inform one another, as they do throughout the scientific enterprise¹².

¹² Structural Genomics is sometimes described as a direct and analogous follow-on to genetic sequencing efforts (such as the Human Genome Project). Participants of the Florence workshop agreed that this analogy can be misleading. The role of computation and modelling provides an illustration of this: for sequencing, there is no alternative to experimentally extracting the entire nucleotide sequence of DNA, whereas, for structural analysis, computation can partially substitute for measurements.

It should be noted that important functional information about some proteins can be obtained from less-than-perfect structural determination. For example, the binding sites and identities of potential ligands (molecules such as hormones and neurotransmitters that interact with proteins) can be established from structural analyses that only determine the rough overall shape of the protein molecule. On the other hand, accurate, high-resolution structures are needed for understanding the subtle ways in which proteins combine and interact with one another to form complex functional assemblies within cells.

Data Issues

All structural biology projects generate large amounts of data that needs to be made available to researchers world-wide. Three issues are of particular interest to policymakers:

1. Permanent, stable funding for databases (including appropriate human resources) 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 because databases are increasingly being implemented on an international basis.
2. Structural and functional information needs to be incorporated into the databases of appropriate publicly-accessible Internet-based data banks. Protocols for cloning, expression, crystallisation (where applicable) and structure determination should also be made available. There are a number of technical problems in this area, linked to the different characteristics 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.
3. 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. Some redundancy in structure determination is desirable to ensure the quality of structures, but researchers should be aware of the status of work being carried out by others around the world. The scientific community is addressing this issue at the present time.

Intellectual Property Rights and Release of Data

Questions surrounding the patentability of the results of genomic research are complex, and the courts will be 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 easily transferable to a structural genomics project. 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.

The scientific experts who participated in the Wellcome Trust/NIGMS meeting in April (see above) agreed to explore IPR issues in preparation for their next meeting, with special emphasis on developing recommendation for possible rules and procedures that would govern the release of various types of data during future large-scale structural genomics projects. Policymakers may wish to follow these discussions, and to further study genomics-related IPR issues.

4. Overall Conclusions and Follow-on Actions

The discussions at the Florence workshop revealed very strong interest in structural genomics. Participants agreed that the field is at an early stage of development, and that OECD countries naturally differ in their assessments of the priority and support that it is assigned at the present time. Preliminary discussions in the scientific community are under way regarding the technical and scientific parameters of potential international collaborative projects. Any such project would be daunting on grounds of technical feasibility alone, even if there was consensus on the merit of dedicated high-throughput structural investigations relative to biological research in which structural determination is simply a component of some broader targeted enquiry. The number of proteins that would have to be analysed, and the time and resource needs for preparing, measuring and analysing the samples, all present major challenges at the present moment. This situation could change significantly, even dramatically, if breakthroughs can be achieved in the relevant technologies (this happened in the case of genome sequencing, with the most gratifying results). Therefore, governments are encouraged to review their policies in this area to ensure that they are well positioned to contribute to, and take advantage of, progress in this field. Workshop delegates agreed on the importance of closely monitoring the developments in the projects that are being initiated in Japan and the United States, as well as future projects undertaken elsewhere.

Workshop delegates agreed that further intergovernmental consultations, complementary to those taking place in the scientific community, might be appropriate, given that some of the key issues are in the purview of governments: for example, funding of new facilities, negotiating international agreements, protecting intellectual property rights. Delegates agreed that the GSF should authorise the appointment of a 3-4 person contact group to keep abreast of developments in structural genomics and to decide if and when a future meeting might be desirable. The contact group would also propose an agenda of any future meeting, keeping in mind that a Europe-wide consultation should be sought and that private companies might be involved. If appropriate, the contact group could propose to hold a meeting under the aegis of the Global Science Forum, possibly late in 2001.

**OECD Global Science Forum Workshop on Structural Genomics
Florence, June 8/9, 2000**

Agenda

1. Welcome and Introduction
2. Definition of Structural Genomics
3. Information Exchange about Governmental Priorities, Plans, and Programmes - *tour de table*
4. Priorities and Targets for Structural Analysis
5. Data Issues
6. Facilities for Structural Genomics
7. Intellectual Property
8. Mechanisms of International Co-ordination and Collaboration
9. Conclusions and Outline of a Report to the Global Science Forum

Workshop Participants

Chairman	Prof. Ivano Bertini
Australia	Prof. Mark von Itzstein
Belgium	Prof. Robert Herzog
Canada	Mr. Marc LePage
Denmark	Prof. Karen Brøndum-Nielsen, Prof. Søren Brunak, Mr. Hugo von Linstow
European Commission	Dr. Philippe de Taxis du Poët
Finland	Dr. Sipo Vanhanen, Prof. Eero Vuorio
France	Dr. J. Haiech
Italy	Prof. Vincenzo Sica, Dr. Andrea Califano, Prof. Riccardo Cortese, Prof. Claudio Luchinat
Japan	Dr. Seiki Kuramitsu, Dr. Masashi Miyano, Prof. Tomitake Tsukihara, Dr. Shigeyuki Yokoyama, Mr. Mitsuru Fujii, Dr. Shigeo Ihara, Dr. Takayuki Odahara, Mr. Masanori Yoshida
Poland	Prof. dr. Piotr Zielenkiewicz
Sweden	Prof. Torleif Härd, Prof. Susanne Holmgren
Switzerland	Dr. Jean-François Conscience, Dr. Manuel C. Peitsch
United Kingdom	Dr. Richard Henderson, Dr. Mark Palmer
United States	Dr. Marvin Cassman, Dr. Keith O. Hodgson, Dr. Roland Hirsch, Dr. John Norvell
Observers	Dr. Marianne Minkowski (ESF), Prof. Joël Vandekerckhove (ESF), Dr. Barbara Skene (Wellcome Trust)
OECD	Dr. Stefan Michalowski, Dr. Michael Osborne, Ms. Sachiko Ishizaka
Invited Speakers	Prof. John L. Markley, Prof. John Moulton, Dr. Joseph Straus
Welcome Committee	Prof. Alessandro Bettini, Prof. Giuseppe Biorci, Dr. Gioacchino Fonti
Organising Committee	Dr. Antonio Rosato (Scientific Secretary), Dr. Rebecca Del Conte