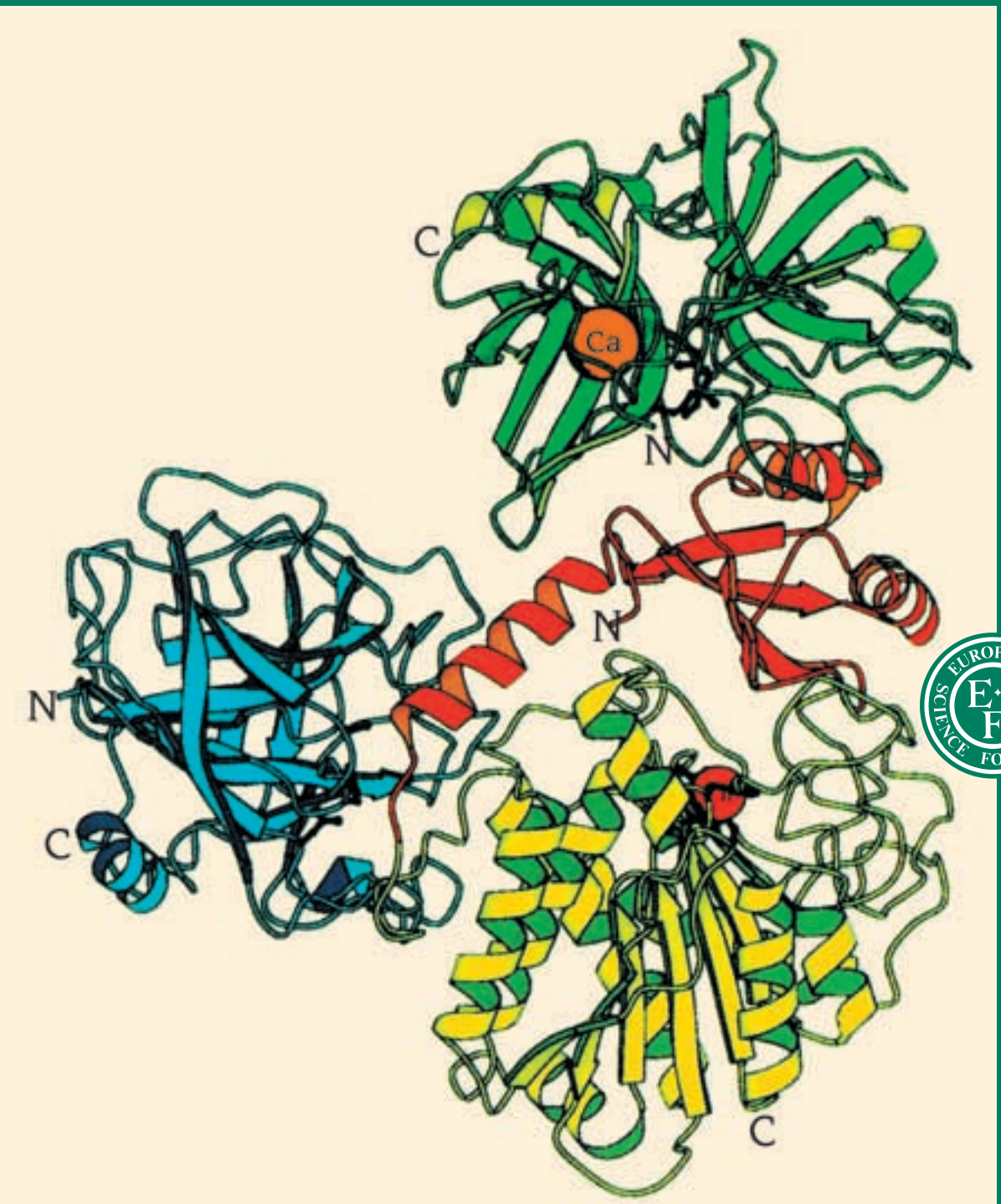
ESF Studies on Large Research Facilities in Europe

Protein Structure and Function in the Post Genomic Era



ESF Study Report
June 2001

The European Science Foundation (ESF) acts as a catalyst for the development of science by bringing together leading scientists and funding agencies to debate, plan and implement pan-European scientific and science policy initiatives.

ESF is the European association of 67 major national funding agencies devoted to basic scientific research in 24 countries. It represents all scientific disciplines: physical and engineering sciences, life and environmental sciences, medical sciences, humanities and social sciences. The Foundation assists its Member Organisations in two main ways: by bringing scientists together in its scientific programmes, EUROCORES, forward looks, networks, exploratory workshops and European research conferences, to work on topics of common concern; and through the joint study of issues of strategic importance in European science policy.

It maintains close relations with other scientific institutions within and outside Europe. By its activities, the ESF adds value by cooperation and coordination across national frontiers and endeavours, offers expert scientific advice on strategic issues, and provides the European forum for science.

An electronic version of the report is available at the ESF's web site www.esf.org

Cover picture:
A wide array of folds are found in protein domains and subunits. The figure shows the crystal structure of an oligomeric association of three digestive proenzymes. Gomis-Ruth et al. (1997), J. Mol.Biol. 269, 861-880. © Academic Press. The structure was solved at the Abteilung für Strukturforschung, MPI-Biochemie, München (Germany).

Protein Structure and Function in the Post Genomic Era

Contents

2

Foreword

In 1997, the European Science Foundation (ESF) was asked by the UK Medical Research Council to look into the prospects for synchrotron radiation in the Life Sciences in Europe. This prompted the ESF Board to set up a review of the needs for European synchrotron beam-lines for biological and biomedical research. The review on the *Needs for European Synchrotron and Related Beam-Lines for Biological and Biomedical Research*, was published as an ESF Study Report in November 1998. It proposed immediate and medium-term actions primarily related to increasing the efficiency of the use of existing national and European facilities, to investing in further development of beam-lines and detectors, and to implementing the existing plans for new synchrotron radiation sources.

As a follow-up to this ESF review, the Foundation invited its Member Organisations to a one-day meeting in London on 17 March 1999, to discuss the review and to agree on possible future steps to be taken in this area. This meeting ended with a general agreement that the priorities for shared attention fall into three broad areas: (1) protein structure and function; (2) proteins in cell function; and (3) informatics.

In the wake of this meeting, the ESF Board came to the conclusion that the ESF should play a leading role in the follow-up to the review and decided to set up a small steering group to consider possibilities for a broadly-based and multidisciplinary action in post-genome research. The steering group met in London on 23 June 1999 and identified several key points that must be strengthened if Europe is to remain competitive with the USA and Japan. In view of the new opportunities opened up by the ongoing developments in genomics and bioinformatics, the steering group proposed to focus on protein structure and function in the biological/biomedical context, and to set up a multidisciplinary working group in this area with a remit along the following lines:

- Defining the needs for different methods and technologies;
- Promoting the development of new methods and technologies;
- Identifying the bottlenecks in technologies and infrastructures as well as in human skills;
- Developing exchanges between scientists of different expertise;
- Promoting training in various traditional as well as new disciplines;
- Proposing actions to be taken by policy makers at national and European levels.

This proposition was discussed and agreed at the ESF Governing Council on 23 and 24 September 1999, and the multidisciplinary working group (*see p. 26*) met for the first time in London on 7 December 1999. Based on the general principles already identified, five thematic groups[1] were set up according to particular methodologies, each of them chaired by a member of the working group.

The multidisciplinary working group also listed a number of scientific issues for the thematic groups to reflect on, such as: protein identification, protein folding and unfolding, protein dynamics, protein-protein and protein-ligand interactions, protein identification, structure of supra-molecular complexes, membrane proteins, protein engineering, integrative biology, and manpower and training.

The reports of the five thematic groups were generated during the spring and summer of 2000. The chairs of the thematic groups met in Strasbourg on 18 and 19 September 2000, to work out a draft report based on their thematic reports (Appendices 1 to 5). The final report was written after thorough consultation among the members of the multidisciplinary working group and the five thematic groups.

[1] Group I: Protein production – Group II: Biophysical chemistry – Group III: 3D structure – Group IV: Proteomics – Group V: Bioinformatics

Executive summary

With the new opportunities for biology in the post-genomic era comes a requirement not only for expensive equipment, but also for building stronger European research units and to strengthen scientific collaboration between a range of experts, often beyond the reach of individual European states. Furthermore, the high-throughput approaches initially developed for genome sequencing are now generating an ever-rising tide of information about protein structure and function. It is important to exploit, and develop rapidly, such opportunities in concert with the more classical, problem-oriented research on proteins. In the USA and Japan major investments are already in hand in what is loosely called 'structural genomics'. It should be remembered that in these two countries there are no national barriers to the funding schemes used to support these activities, and in the USA a number of nation-wide consortia have recently been funded within the rapidly expanding budget of the National Institutes of Health (NIH). In Europe, there is now an urgent need to join national and European forces to reap the benefits on offer and to avoid wasteful duplication of effort. In so doing, it is important that we build on existing strengths in developing and exploiting the new research opportunities. We suggest that the newly established scheme of the European Science Foundation Collaborative Research Programmes (EUROCORES), could provide an effective and efficient need-driven mechanism for collaboration at a multinational level within Europe.

Based on these premises we suggest the following actions:

1. Establish networks of laboratories based on the best possible equipment and expertise, devoted to the various steps from genome analysis to the identification and understanding of protein function. In establishing these networks, it is important to ensure not only strong interactions between the use of complementary approaches and methodologies but also to develop and disseminate the use of new technologies.

2. Establish high-throughput research in the area of protein structure and function, based on existing national or European research groups and institutions, at the same time ensuring a proper balance between this novel approach and the different modes of more problem-orientated research on proteins. This will not only include access to expensive research facilities such as synchrotron radiation, Nuclear Magnetic Resonance (NMR) instrumentation and high-throughput mass spectrometry, but also dedicated efforts to solve progress-limiting steps related to the production and functional characterisation of proteins.

3. Establish both the networks and the high-throughput laboratories on a competitive basis, ensuring that they are efficiently managed and regularly evaluated, judged on performance and scientific needs.

4. Emphasise the need and take steps to educate and attract trained staff into the field, as exemplified by the pressing shortage of bioinformaticists.

5. Referring to the 1998 ESF report on the *Needs for European synchrotron and related beam-lines for the Life Sciences*, the working group notes major improvements in line with the suggestions of the ESF report. Access to beam-time does not at present appear to be a limiting factor for structural biology,

but the planned third-generation synchrotrons with beam-lines dedicated to structural biology remain vital investments for future European competitiveness in the field of protein structure and function.

6. Since most, if not all, Member Organisations of the ESF are currently taking steps to meet the new challenges in the post-genomic era of the Life Sciences, the working group proposes the establishment of a number of multinational collaborative research programmes in the field of structure and function of proteins, adopting the variable geometry of, for example, the EUROCORES. Education and training should be organised as part of such activities. Topics for programme collaboration are suggested under Recommendations (page 6).

7. Address the problems regarding the Intellectual Property Rights (IPR) that will arise from the results, before the EUROCORES or other projects begin.

8 . Additional funding is needed at national and European levels to reach the stated objectives in post-genomic research on protein structure and function, including the adoption and development of high-throughput methodologies where applicable.

Recommendations

Today, the rapid increase in genomic information offers a new stepping stone for the Life Sciences, often called 'functional genomics'. This research is aimed at defining the expression and function of each gene of a genome (in practice, the presence, abundance, location, temporal regulation and function of a protein), with the purpose ultimately of understanding the integrated function of the intact, living cell or organism.

Knowledge of protein structure and function remains central to an academic understanding of essentially all areas of the Life Sciences, and is crucial to the commercial sector over a whole range of applications from health care to food and agriculture. With the new opportunities for the Life Sciences in the post-genomic era comes a need not only for expensive equipment and the adoption of new methodological approaches, but also for education, training and scientific collaboration between various experts beyond the reach of most individual European states. There is an urgent need to link national and European capacities to reap the benefits on offer and to avoid unnecessary duplication of effort. In so doing, it is important that we build on existing strengths in developing and exploiting the new research opportunities.

For the European Life Sciences to be internationally competitive in elucidating the structure and function of proteins, the working group identifies a number of specific needs such as:

1. solving the progress-limiting steps in the production, purification and crystallisation of proteins;
2. determining high-resolution 3D structures of these proteins (structural genomics);
3. exploiting and developing new opportunities to determine the global protein patterns in cells and organisms (proteomics);
4. emphasising the need for in-depth studies of the relationship between the structure and function of proteins, including their interactions in supramolecular complexes;
5. strengthening the field of bioinformatics and ensuring state-of-the-art access for all biologists to bioinformatics and various databases;
6. establishing and developing high-throughput research technologies for the study of protein structure and function.

While these individual aspects of the field were discussed by separate subgroups, they clearly must be integrated in any real research programme in functional/structural genomics. The expression of soluble protein in a high-throughput manner is presently the rate-limiting step, and determined efforts must urgently be made to overcome this obstacle. However, attacking this in isolation is not appropriate: associated biophysical studies on the state of the proteins produced, bioinformatics assessment of the choice of targets and data-basing of the results of the expression procedures, must be closely linked in, as well as developing methods for NMR and crystallographic analysis. A holistic approach is vital: we do not want methods developed in isolation, but rather an integrated attack on the problems by laboratories with expertise in the relevant areas.

Furthemore, time is of the essence: the USA and Japan have had substantial programmes up and running for some time: the first round of results is already emerging from these teams. The USA through the NIH has funded 7 projects with an overall budget of 150 million dollars over five years for structural genomics alone. Too much delay in Europe with

protracted discussions of how we might proceed is not enough, we must act now if we are not to fall further behind the rest of the world. It is most important that the European States secure strong research programmes in these areas. Special emphasis should be placed on co-ordination and networking at the European level regarding (1) focused research efforts at the highest competitive level; (2) access to instrumentation; and (3) training of personnel.

The working group proposes the following recommendations as central to the success of European efforts:

Major programmes

Major programmes such as the EUROCORES should be undertaken by groups of networked national and European laboratories or research groups. These would not be limited to structure determination, but should include the development of novel scientific techniques for gene expression, protein production, crystallisation, biophysical characterisation, bioinformatics and studies of the various aspects of protein function in the living cell. Each programme should be composed of connected nodes following schemes of variable geometry adapted to fit the needs of particular programmes. Proper management of the programmes is vital, and a strong reviewing system will be essential for the selection of participating nodes and for the follow-up of research progress. The programmes should additionally provide the basis for exchange of staff, for education and training in key areas, such as gene expression and protein production, and for the development and wider dissemination of high-throughput techniques as these become cheaper and more efficient.

We suggest the establishment of programmes such as EUROCORES in the following areas (the need for education and training should be integrated in every programme, and workshops and seminars should be organised across the borders of interlinked programmes):

- Protein production and purification
- Structural genomics
- Proteomics
- Structural-functional relationships of proteins and supramolecular protein complexes
- Bioinformatics

Before launching a particular programme within the area of protein structure and function, the working group suggests a series of ESF exploratory workshops be held to develop the topics in detail. These workshops should be organised jointly by the ESF Standing Committees for Biomedical Sciences (EMRC), Life and Environmental Sciences (LESC) and Physical and Engineering Sciences (PESC) in order to gather the necessary expertise. The present study with its annexed reports on subject areas, together with the complementary efforts undertaken by the ESF Standing Committees, should give enough background information for a thorough planning of these workshops. The suggested programmes should aim at building on national and supranational complementary strengths, attracting the leading research groups in Europe.

To be able to establish these highly competitive research programmes between leading national or European laboratories or research groups, the working group sees an urgent need to strengthen, and sometimes even establish, key research nodes that can provide not only access to the best instrumentation, leading expertise and competitive

training, but also take the responsibility for developing and running large-scale and high-throughput research on protein structure and function.

We suggest that the establishment of **high-throughput research laboratories** be based on existing national or European research groups and institutions, but reinforced by the necessary instrumentation and staff, both scientific and technical. These nodes should pursue automated robotic-controlled approaches and be platforms for the development of new high-throughput technologies. Although the problems may be technical and engineering rather than scientific in the first instance, we suggest that the establishment of high-throughput operational approaches in association with existing target- or problem-oriented research on proteins will optimally create fruitful interactions between the two modes of research. To this end, we see an urgent need to develop high-throughput research in protein production, structural genomics and proteomics. The high-throughput mode must not be seen as an end in itself: its role is to improve our ability to address relevant biological problems.

Protein production laboratories

The establishment of laboratories in which new, highly efficient high-throughput production approaches are developed, and in which general and specific expertise and training in the production of proteins is provided, should help to remove one of the most evident bottlenecks in structural genomics. Such laboratories should ideally be located in active – already built and running – research centres, in order to decrease costs, facilitate a quick implementation, and provide complementary infrastructure. The laboratories should be financed with the requirement to provide central facilities and training, offer fellowships, organise workshops and practical courses, and to maintain a high quality of research on high-throughput methods for producing proteins. It is also important to organise and keep accessible systematic databases on protein production.

Three-dimensional (3D) structure laboratories

The equipment needed falls under three subheadings: Macromolecular X-Ray Crystallography, Nuclear Magnetic Resonance Spectroscopy and Cryogenic Electron Microscopy. All structural techniques rely heavily on automated computer suites for their most effective use. Only a small number of scientists contribute to the development of such software and a key role for the European agencies is to ensure these are strongly networked so as to avoid duplication of effort and provide user-friendly compatible systems. The provision of the necessary new instrumentation could be considered in the light of the location of the nodes and networks covered by the suggested programmes for structural genomics.

Macromolecular X-ray crystallography

requires availability of, and access to, synchrotron beam-lines dedicated to macromolecular crystallography, funded both nationally and at the European level (see page 24). Although the needs of X-ray crystallography are paramount, the more limited applications in spectroscopy and small-angle scattering should not be overlooked. Crystallography will remain the prime means of 3D structure determination for the next few years, with a typical third-generation synchrotron beam-line capable of producing 2000 or more data sets each year.

Nuclear Magnetic Resonance (NMR) is the second major structural technique. A small number of national and European centres with 800-900 MHz instruments already exist. Ensuring access to such facilities for all European scientists would require the establishment of a further set, perhaps 5 to 10, of such centres. Any one instrument will probably not produce more than 25 new structures per annum, but the provision of information about non-crystallisable proteins and protein dynamics and interaction are important, sometimes crucial advantages.

Cryogenic Electron Microscopy (CEM) is ideally used for looking at super-large structures and complexes, many of which are likely to resist crystallisation. Again there are already a small number of European centres with state-of-the-art instruments, but a doubling of this number would ensure access for many more scientists.

In addition, ***Neutron scattering*** methods also provide unique and important complementary information at different levels of protein structure and dynamics. Europe already has the premier neutron sources in the world; the challenge is to maintain and develop state-of-the-art instrumentation with access for biological studies.

Proteomics laboratories

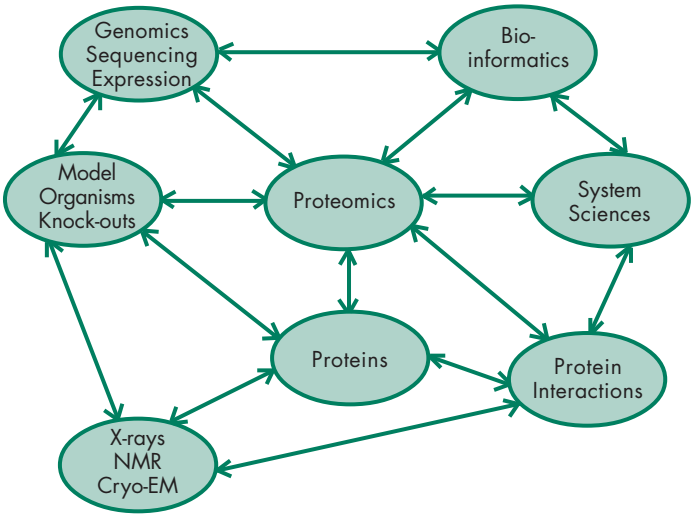
In addition to the need to determine the structures of proteins selected for high-throughput screening, there should be complementary efforts studying on a global scale the dynamics of gene activities in cells under specific growth conditions, when subjected to external stimuli or during certain stages of development. These studies are collectively referred to as ‘proteomics’.

While protein identification and the quantification of protein expression levels have so far been a limited approach, mainly due to technical limitations imposed by the use of 2D gel-electrophoresis, recently reported gel-free differential proteome display methods may start to revolutionise the field. They will stimulate the need for the creation of facilities in which large-scale mass spectrometry of proteins and peptides will become the heart of future proteome centres. The concentration of such instrumentation on this large scale was initiated by a number of specialised companies, and has only recently been recognised at national and international academic levels. It is most important for the European Life Sciences that strong proteome centres be established across Europe. These centres must be closely associated with structural biology laboratories for optimum impact (or vice-versa). Access to technologies for proteomics, such as mass spectrometry, must be made available throughout Europe, and common database standards for proteins must be worked out.

The challenges in the post-genomic era lie not only in identifying and solving the high-resolution structures of as many proteins as possible, but also in addressing questions to bridge the gaps between molecular and cell biology and between thermodynamics and structure.

The working group therefore also emphasises the need for the **broad support of small-size facilities** whose mission is mainly to identify proteins in protein-protein interactions or to study multi-protein complexes, and which normally operate at a modest cost. For such studies of the function of proteins and their supra-molecular complexes, appropriate instrumentation (for example various types of spectroscopy, microscopy, micro-calorimetry, ultracentrifugation) should be made available to a large number of institutions with the relevant expertise.

These new or renewed approaches will generate a vast amount of new information on extremely complex regulatory networks of protein expression, metabolic components and their interlocking activities. For the first time in history, biologists are in the challenging position of being able to see and to measure such events and they can start to discover and to unravel networks that are used by cellular systems. The complete integration of these data into global models of cellular function is possible only with the aid of a new type of bioinformatics, which can accept and store vast amounts of integrated cellular information, and do so in a format adapted to various means of data mining for maximal accessibility ('system sciences').



Interrelation scheme showing proteomics connected with different protein sciences

Bioinformatics

Bioinformatics not only provides the essential databases (DNA and protein sequence databases, 3D structure databases, gene and protein expression databases, metabolic pathway databases), it also supplies the computational tools required to:

- identify targets for high-throughput programmes in structural biology;
- predict protein folds and provide homology-based models of protein structures;
- simulate protein dynamics;
- predict protein function based both on theoretical analyses and experimental studies (DNA and protein microarrays);
- analyse the evolution of proteins;
- develop global models of cell function ('system sciences').

There is a pressing shortage of trained bioinformaticists in Europe. This shortage is felt in industry and academia alike and is damaging the successful dissemination of the results from genome projects. Unless Europe is able to vitalise very quickly its bioinformatics community, most of the fruits of the large genome sequencing and functional genomics/ proteomics efforts will go to the USA.

There is a great need to promote exchange and collaboration (networks between laboratories, workshops on specific topics, academic-industrial sector interactions), to promote training (PhD studentships, post-doctoral fellowships), and to create a professional website offering lectures, tutorials and practicals.

There is also a major need for the establishment, maintenance and provision of access to databases in the various key areas central to the study of protein structure and function. Gene sequences and primary structures of proteins are already covered by, among other things, the European Bioinformatics Institute (EBI), whereas there is a pressing need for databases covering: (1) expression systems for cloned genes and protein engineering approaches; (2) purification protocols for expressed proteins, including solution properties such as solubility, stability and aggregation; (3) protein folding;

(4) crystallisation protocols; and (5) database integration. In this context, it is very important to put the database activities of EBI on a secure footing.

Education and training

The need for skilled students, post-doctoral fellows and support scientists in all stages of the projects is acute. There is a shortage of suitable people trained to a high-level in practically all sectors, but especially in bioinformatics, protein chemistry and biophysics, protein production and purification, and the development of methods for structure determination. Europe must take the necessary action to ensure that these fields are attractive to young scientists in the future.

Intellectual Property Rights (IPR)

Appropriate means must be established to address the problems that may emerge regarding the IPR that will arise from the results obtained. Groups from the EU and USA have already met during 2000 to initiate discussions, and clear guidelines will need to be laid down at the start of the genome projects.

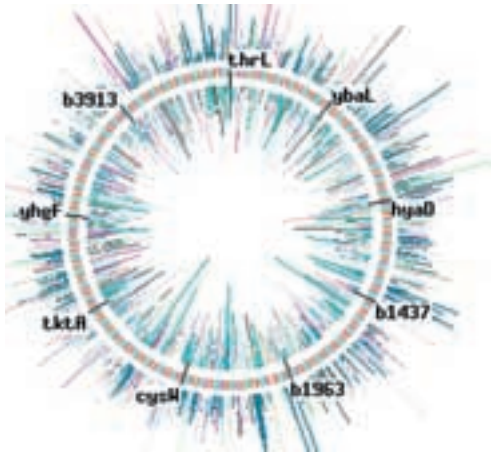
Background: New opportunities for biology

It has been said many times, but that does not make it any less true, that today we are witnessing the beginning of a revolutionary development in biology that is transforming not only our basic understanding of life but also opening up totally new opportunities for society in terms of food production, health care, and a variety of technical advances based on biological concepts. In this process, biology needs and will attract an increasing number of scientists and concepts from other disciplines, for example from the physical and mathematical sciences, to deepen our understanding of biological complexity. Furthermore, new disciplines, such as we see in the growing field of bioinformatics, will develop at the interfaces between old ones, and technology and biology will become more and more interlinked in the use of biology as the technology platform to provide goods and services to society. Research in biology today has an unprecedented level of interdisciplinarity.

Major scientific steps

The beginning of this transformation has been brought about by several major scientific advances. It is over fifty years since it was recognised that DNA is the carrier of the genetic information in the cell while the proteins are the functional components that carry out the host of functions which living organisms are required to perform. These include enzyme catalysis, redox reactions, ligand binding, an extensive range of specific macromolecular recognition processes in immunity and signal transduction, as well as key structural roles. Early studies of proteins were restricted to those that were produced naturally in large amounts and easy to extract. The development of DNA cloning methodology changed this forever: any protein can, in principle, be obtained in large amounts once its gene, or for eukaryotes its cDNA, has been cloned and inserted in an appropriate expression vector, and once the folding and solubility problems have been solved.

A second explosion came with the advent of techniques for the sequencing of whole genomes. At present about 40 microbial genomes have been sequenced and another 130 are on the way. The *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (worm), *Drosophila melanogaster* (fruitfly) and *Arabidopsis thaliana* (plant) sequences are complete, and the human one has just been completed.



Schema of the content of *E coli* genome

It was also realised some years ago that the knowledge of the three-dimensional (3D) structure of a protein is essential for a full description and understanding of its function. Protein crystallography and NMR spectroscopy are the techniques of choice for determining such atomic structures and both have advanced rapidly in recent years in their automation and throughput. Other approaches, like Cryogenic Electron Microscopy and proteomics, are also progressing quickly,

providing very interesting complementary information. However, our ability to determine 3D structures has fallen way behind the speed at which genome sequences at present are being determined. This mainly reflects an inability to express, purify and crystallise the proteins in an efficient and high-throughput manner rather than in intrinsic limitations of X-ray crystallography and NMR spectroscopy as techniques.

Today, the rapid increases in genomic information offer a new stepping stone for biology, often called 'functional genomics'. Such research is aimed at defining the expression and function of each gene in a genome (in practice, normally the presence and function of a protein), with the purpose ultimately of understanding the integrated function of the intact, living cell or whole organism. This development is driven not only by the growing accumulation of genomic information, but also by new technologies for gene sequencing and mapping, for studies of gene expression, and for exploring protein function and metabolic interactions. This is supported by new developments in bioinformatics that allow the management of vast experimental databases, integration of various types of information to explore biological complexity, and the creation of unforeseen opportunities to conduct computer-assisted experiments. However, such analyses of the structure and the function of proteins is a much more difficult and time-consuming process than the determination of the DNA sequence of the corresponding genes. Moreover, it requires not only the use of expensive research facilities and instrumentation such as access to synchrotron radiation and high field NMR spectrometers, but also very strong efforts in protein over-expression, purification and crystallisation, which, if not achieved, will become an increasingly limiting step in studies of protein structure and function.

New research modes and requirements

The genomic projects of the last decade have introduced a new type of research into biology, based on high-throughput and dedicated research facilities for the purpose of rapidly and accurately producing and storing DNA sequence information. The data are stored in data banks accessible to all researchers, including those more interested in target-oriented research of a basic or applied character aimed at solving particular biological or technological problems. This approach, combined with bioinformatics, has proved to be very effective, not least in developing new technologies and mathematical algorithms that are becoming accessible to individual researchers at affordable cost. The question now is how the very successful high-throughput research concept of gene sequencing and genomics can be harnessed for the elucidation of how the genome functions, in the selective expression of a multitude of proteins and other macromolecules participating in regulatory processes and in various metabolic pathways characteristic of living cells and organisms. High-throughput approaches are conceptually easy to envisage for determining the structure of small, single-domain proteins and such developments are already aggressively under way in the USA and Japan, but with a few efforts only thus far in Europe. More complicated proteins, their association into complexes, macromolecular interactions in biological processes, the level of protein expression both spatial and temporal, the identification of unknown protein functions, etc., will require more sophisticated solutions and the development of new techniques.

Progress in the new biology requires not only investment to develop these new and expensive technologies; it is also very

important to gather the necessary expertise and to train new generations of young scientists to exploit the existing opportunities and to develop as yet unforeseen new ones. Consequently, we now see a restructuring of biological research internationally into defined national centres (either as single units or as networks between units) where the necessary expertise and technology platforms are gathered. However, there is great concern that this development is too sluggish in Europe and that the political systems at national and European levels are responding too slowly to the needs of modern biology. European competitiveness in biology and, as a result, the exploitation of biological knowledge for the benefit of industry and society, is being jeopardised. Major investments are already in hand in the USA and Japan in what is loosely called 'structural genomics'. In these two countries it should be remembered that there are no national barriers to the funding schemes used to support these activities, and in the USA a number of nation-wide consortia have recently been funded by the National Institutes of Health (NIH).

The fragmentation of research effort and the lack of co-operation in Europe are inimical to the concentrated dedicated efforts needed in modern biology. European States must now join forces to achieve what is needed to secure the optimal development of competitive biology for the future. It is therefore high-time to define strategies, priorities and plans at national and European levels.

This would include:

- establishing more effective means of collaboration to secure access to the most appropriate methods and facilities for determining the structure, dynamics and function of proteins;
- meeting the needs for international training and mobility of young researchers;
- developing the networking across Europe that is required to exploit to the full the scientific and commercial opportunities that biology now offers.

Current status of protein research

Proteins are at the core of biology. Inevitably, perhaps, the massive advances in genomics research over the past few years have now led to a renewed focus on protein structure and function. The immense successes of high-throughput approaches in genomics (DNA sequencing, DNA microarrays) have inspired similar initiatives in protein science, with high-throughput programmes for 3D structure determination as one of the most tangible result so far. However, to capitalise on the genomics revolution, large areas of protein biochemistry, molecular biophysics, structural biology, and theoretical work (bioinformatics) on protein structure and function must rapidly be strengthened. In this section, we discuss the future of protein research from various perspectives.

To facilitate the discussion and eliminate any misunderstandings, the working group agreed the following working definitions based on discussions at various meetings over the last couple of years:

Biotechnology: the use of living organisms, or components of living organisms, to make new products or provide new methods of production. Relevant techniques include the use of recombinant DNA, cell cloning, cell fusion and bioprocessing.

Genomics: the development of knowledge about the identity, nature and function of the material contained in the genome. This knowledge can be used to identify and manipulate organisms for use in the development of biotechnological processes and products.

Structural genomics: the science of determining the structures of the proteins in an organism.

Functional genomics: the science of determining the function of a gene; in practice determining the function of a protein. Functional genomics implies a systematic 'whole organism approach' as opposed to genetics, in which a single specific gene is the focus.

Proteomics: the quantitative and qualitative analysis of the global protein pattern of a tissue, a cell, an organelle or a supramolecular protein organisation present or synthesised at a given moment.

Bioinformatics: the use of computational techniques to handle, analyse, and add value to the flood of data coming out of modern genomics and proteomics research.

Structural genomics

In the post-genomic era, we are now in the position of being able to complement studies of protein function, solution interactions and dynamics, with work to define their high-resolution structures (Appendix 3). This creates an enormous incentive for developing high-throughput methods of analysis, from the expression of the cloned DNA sequence to the full 3D structure. It is most important to emphasise that all structural studies will benefit from high-throughput techniques. Thus, while we have usually been restricted to looking at systems integrated on a relatively small scale such as operons, the world would now appear to be our oyster. At first sight it seems that all we have to do is express our target gene and off we go: if only things were so easy!

The first and obvious category of study – that of determining the 3D structures of all the proteins in a genome, generally referred to as structural genomics – is of course possible only when the sequence of the complete genome, or at least of a significant

component, has been established. However, especially with this target, there is a catch: only some of the proteins can be expressed in a soluble form and in a homogeneous 3D-fold suitable for crystallisation or NMR spectroscopy. Smallish, especially single-domain, proteins are clearly viable, whereas others, such as larger multidomain proteins, glycosylated proteins expressed in the "wrong" organism and most membrane proteins, are likely to prove problematical. Using current technologies, high-throughput approaches are proving easy to develop for the first category. Such studies are already aggressively underway in the USA and Japan, but there are only a few comparable efforts to date in Europe (see Appendix 3). This highlights one aspect of the work, turning current easy expression and crystallisation approaches into scaled-up, robotic-controlled systems requiring little user input, with the results being scanned using video devices, the output stored and analysed on disk, and a computer identifying potentially suitable crystals. To emphasise once more, this is the easy part of structural genomics: solving the structures where no problems are encountered, and with only technical scale-up needed. These proteins will provide the first 3D structures from any structural genomics programme, and will start to fill in at least some of the holes in fold space.

The second category of proteins, perhaps the majority at present, are those that prove resistant to this straightforward approach. Clear identification of this set may actually prove to be one of the great gains of structural genomics: it will rapidly become clear just how extensive this set is, and what type of sequences it reflects. This is where novel science starts again, as new techniques will be needed to over-express, solubilise and crystallise these proteins. Assuming crystallisation can be successfully achieved at a later date, these proteins can then be passed over to high-throughput with the appropriate factors exploited.

So much for the so-called 'structural genomics' projects themselves, let us turn to the structural study of proteins and complexes with targeted function. This is more typical of the type of project that the structural biologist has conventionally addressed: pick a protein with an interesting function, usually relevant in a biological, medical or commercial sense, and carry out extensive studies on its crystal or NMR structure and its function. Such studies are now greatly assisted by our knowledge of genome sequences. If the sequences are known of homologous proteins from one or more sequenced genomes, a whole set of constructs can be produced for high-throughput screening of expression, solubility and crystallisation, rather than the one or two normally used. As an example, the Berlin team is starting from a very rough estimate that only 1 in 25 constructs may lead to a successful 3D structure by crystallography or NMR spectroscopy. The known sequences can come from other species or other related proteins within the genome. In addition, comparison of the conservation and nature of the sequences may suggest appropriate constructs with single domains or omitting potentially flexible linker regions.

Thus the key point of high-resolution structural biology in the post-genomics era is NOT to restrict ourselves solely to structural genomics projects in the strictest sense, but rather to exploit high-throughput for a broad range of structural and targeted function projects which will complement one another to a large extent.

Although crystallography has been referred to almost exclusively as the method of structure determination in this section, the same conclusions essentially apply to NMR

spectroscopy. The latter has the advantage that crystals are not required. However, NMR spectroscopy is at present limited to solving structures of proteins with molecular masses of no more than 30 - 40 kDa. Although technical advances are continuously evolving, and NMR spectroscopy will have a valuable part to play in many structural analyses, it is unlikely that it will replace X-ray crystallography as the high-throughput method of choice.

Bioinformatics

The definition of the term 'bioinformatics' adopted in this report is "the use of computational techniques to handle, analyse, and add value to the flood of data coming out of modern genomics and proteomics research" (Appendix 5). From the point of view of research on protein structure and function in the post-genomic era, bioinformatics understood in this way not only provides the essential databases (DNA and protein sequence databases, 3D structure databases, gene and protein expression databases, metabolic pathway databases). It also supplies the computational tools required to:

- identify targets for high-throughput programs in structural biology;
- predict protein folds and provide homology-based models of protein structures;
- simulate protein dynamics;
- predict protein function based both on theoretical analyses and experimental studies (DNA and protein microarrays);
- analyse the evolution of proteins, and to develop global models of cell function ('system sciences').

Given the very rapid growth in the amount of sequences – and structure-related information – available to the community [*figures 1, 2*],

Growth of Protein Data Bank (PDB) (3D protein structures)

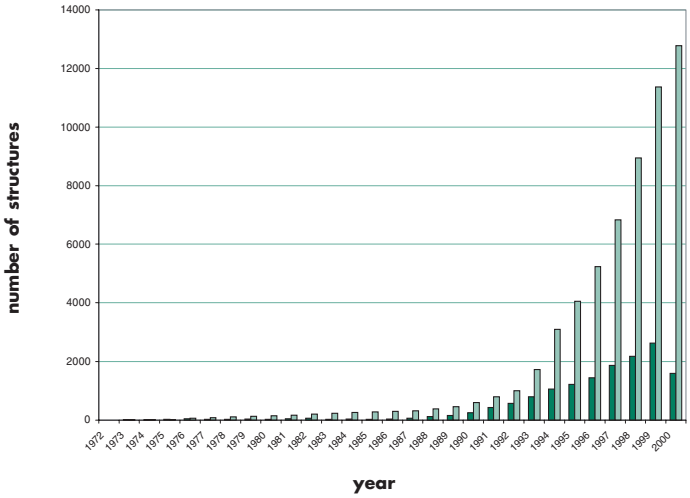


Figure 1: Bars indicate the total number of structures (light green) and the number of new structures (dark green) entered into PDB each year.

Growth of SwissProt (protein sequences)

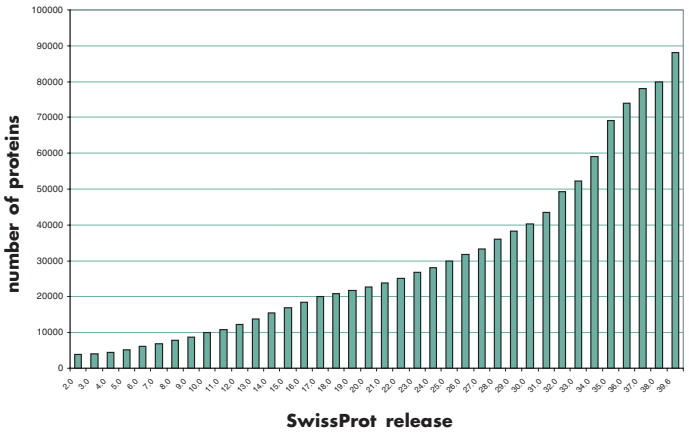


Figure 2: The total number of protein sequences in SwissProt is shown for each SwissProt release.

a strong bioinformatics infrastructure is absolutely critical for the many expanding areas of contemporary biology and biomedicine. In this context, we would like to point to the important work on the standardisation and dissemination of databases done by the Inter-Union Bioinformatics Group.

To cite but one example, as of today there are more than 40 publicly available, fully sequenced microbial genomes, and more than 130 additional ones in progress. The recent publication of the human genome sequence only further underlines the importance of bioinformatics as an enabling technology for future research in protein structure and function.

Generally speaking, Europe is still reasonably competitive in most of the key areas in bioinformatics, especially in the prediction of protein structure and function, protein interaction analysis and database annotations. However, given the current increase in the demand for bioinformatics expertise, both from the commercial and academic sectors, Europe is desperately short of skilled people in this field. The continued failure to secure a long-term funding solution for the centrally important database activities at the European Bioinformatics Institute (EBI) is a problem that Europe must solve immediately.

Proteomics

By 'proteomics' we understand a "quantitative and qualitative analysis of the global protein pattern of a tissue, a cell, a cell organelle or a supramolecular protein organisation present or synthesised at a given moment" (Appendix 4). This is a very rapidly developing field that promises global views of cellular functions as reflected in the dynamic evolution of the protein complement of a cell (the 'proteome') under various conditions. From a technological point of view, proteomics includes both *in vivo* analyses of protein expression, e.g. by various fluorescence microscopy techniques, as well as high-throughput identification and analysis of the proteome, for example, by 2D-gel electrophoresis and mass spectrometry. Proteomics will be central to the functional genomics efforts, in both the industrial and academic environments.

In vivo *analysis of the proteome*

Many of the most important aspects of cellular function and physiology can only be studied in the intact cell. Techniques now exist to localise proteins within a cell and to follow their movements between different cellular compartments in real time, for example, using fusion proteins incorporating a fluorescent reporter such as green fluorescent protein (GFP) and confocal microscopy. Proteins can also be visualised in fixed cells using immunofluorescence. In both cases, high-throughput approaches are being developed, such as genome-wide libraries of GFP fusions and large-scale generation of specific binding reagents such as antibodies. New nanotechnologies have recently emerged that allow the visualisation of individual protein molecules 'at work' in their 'natural habitat', the cell.

In vitro *analysis of the proteome*

Although an unambiguous identification of an isolated protein can be done by classical approaches, such as N-terminal sequence determination by Edman degradation complemented by methods that give information about the size of the protein (for example SDS-PAGE), in recent years approaches based on mass spectrometry have come to the forefront. In these approaches, the protein, in an isolated state or in a mixture, is submitted to high-resolution mass spectrometric analysis, in order to identify it on the basis of a very precise measurement of its overall molecular mass or the masses of its proteolytically or physically derived fragments. The sample requirements are extremely small (pico- to femtomoles). This type of analysis has been greatly facilitated by the improvements in isolation procedures based on 2D-gel electrophoretic methods, high-resolution/high-speed chromatographic

methods, and direct analysis of complex mixtures by means of mass spectrometry coupled to liquid chromatography.

The correct identification of a protein is a necessary initial step for all subsequent work, whether the protein is being obtained from natural sources or by recombinant DNA techniques. The information about the state of a mature protein after isolation or *in vivo* (in terms of post-translational modification, proteolytic processing, mutation, heterogeneity, etc) is crucial if we are to understand its properties, modulate them, or carry out further analysis (for example, the presence of bound carbohydrates makes X-ray crystallography more difficult). The quantification of each expressed protein in its various states is also essential and remains one of the most challenging problems of proteomics.

Like all disciplines in the field, proteomics is also expanding and evolving rapidly at the technological level with recent promising additions of an arsenal of novel techniques.

Quantitative proteome approaches are being developed that compare protein levels in two different samples, based on the measured intensities of two isotopically labelled peptides derived from the same protein labels in the two instances. The data output of such studies forms a differential proteome array directly comparable with the DNA microarray outputs.

A more direct version of the protein microarray technology is to use a protein affinity reagent (for example an antibody) which is spotted and covalently attached onto small surfaces of glass or other material. Protein chips may soon become a powerful alternative or complement to the conventional proteome techniques. Undoubtedly they will evolve to become high-throughput methods. Of course, this technology will be restricted to already identified proteins for which highly specific high-affinity antibodies are available.

Post-translational modifications

Notwithstanding the extreme chemical diversity of proteins due to the 20 different amino acid side-chains, it turns out that many different proteins undergo selective post-translational modifications that are essential to their proper biological functioning. Among these may be listed, to name but a few, the reversible chemical modifications, such as phosphorylation, that underlie signal transduction and metabolic control processes, N-terminal acylations and C-terminal amidations, the latter often associated with peptide hormones, histone acetylations in the control of gene expression; the attachment of swinging arms such as lipoic acid, biotin and phosphopantetheine to multienzyme systems; and the attachment of carbohydrates. The principal investigative methodologies involved are those of amino acid sequence analysis, most notably the use of mass spectrometry.

Important aspects of these modifications remain to be elucidated, not least the specificity and control of the enzymes that catalyse the post-translational modification reactions. A longer-term aim must be the ability to predict, from the amino acid sequence of a protein inferred from the DNA sequence of a genomic open reading frame, whether, where and in what way an unidentified protein will be modified. There is at present, and for the foreseeable future, a continuing need to establish not just whether a modification has taken place, but also to identify the nature of the modification itself, as the list is growing all the time and is obviously incomplete.

Protein production and engineering

Although for certain purposes (for example functional characterisation), it is not always necessary to have a protein in a pure state and in substantial quantities, this is not true for most structure-function analyses, including structural genomics. In this case, the protein is usually required at the milligram level, and quantities in the range 10-50 mg or higher of pure material need to be produced. Also, quite frequently, proteins are required in trimmed (for example individual domains), modified/engineered (for example to eliminate or introduce post-translational modifications, or increase solubility, or attach tags, etc), or labelled (deuterated, ^{15}N, ^{13}C, Sel-Met, etc) forms to facilitate structural and functional analysis. This is not always straightforward, particularly for large multidomain or membrane-bound proteins and complexes. These require particular and time-consuming approaches to their production, not easily adapted to high-throughput schemes. In fact, recent reports derived from large structural genomics projects seem to indicate that at best only 10-25% of proteins are adapted to such schemes (Appendix 1).

Clearly, new preparative approaches for wild-type and engineered proteins, and selective labelling of sites or regions within them (for example by intein-based strategies) are required, otherwise the databases derived from structural genomics projects might become biased by the acquisition of information from the more easily produced (and structurally solvable) proteins. This is a limitation that must be overcome in such projects. The problem is even more acute when very large membrane proteins, and protein complexes, are considered.

Besides providing tools necessary for the production and labelling of proteins for structural genomics projects, protein engineering is a logical extension of such projects at a more applied level: the elucidation of the three-dimensional structures of many proteins will allow structure-based rational design to be utilised in the production of variants with new specificities, altered stability, resistance to proteinases, altered cellular locations, minimised hybrid or chimeric forms better suited for basic science or industrial purposes, etc. Protein engineering will therefore greatly benefit from the output of structural genomics projects.

Protein folding

Much effort has been expended in the past decade to uncover the rules that govern the processes of folding and unfolding of proteins and, although the problem is still not solved, a reasonable precise view is emerging, at least for the simpler single-domain globular proteins. This goal is very important, not only in developing better procedures to predict three-dimensional structures, a primary goal of structural genomics projects, but also for understanding how proteins pass through membranes and take up their intra- and extracellular positions. Moreover, such knowledge facilitates the development of better strategies to produce, redesign or minimise proteins (creating the smallest protein with a given function is an important goal for some biotechnological or biomedical purposes).

Membrane proteins

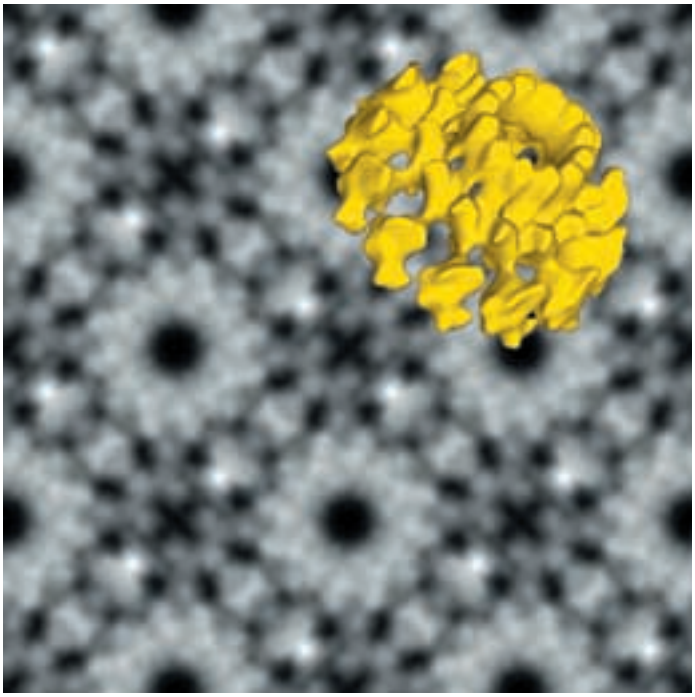
Membrane proteins present a special challenge. Although they represent 20-30% of all proteins and are of central biomedical and pharmaceutical importance, only a handful of high-resolution structures are available. Over-expression and purification of membrane

proteins, as well as their crystallisation, are much harder than for soluble proteins. Moreover, NMR spectroscopy cannot normally be applied to determine the structure of intact membrane proteins at high resolution, although specialised NMR techniques can produce very useful detailed information about, for example, the structure of bound ligands, and of the region of the protein with which they interact. Because of their hydrophobic nature, membrane proteins are also more difficult to study using standard proteomics tools such as 2D-gel electrophoresis and mass-spectrometry.

Europe has so far been able to match the US and Japan in this field. The major obstacles that need to be overcome in the next five years are mainly related to structure determination. In particular, high-throughput techniques, specific to membrane proteins, need to be developed for over-expression, purification and crystallisation.

Protein-protein interactions and multifunctional protein complexes

Protein-protein interactions and the assembly of multifunctional protein complexes, some permanent, some only transient, are increasingly recognised as essential to the normal functioning of the cell. In the case of a multi-enzyme complex, for example, it may be possible to determine its full three-dimensional structure by X-ray crystallography or cryo-electron microscopy, either as a single entity or (more likely) by piecing together the structures of its individual component proteins. With transient complexes, on the other hand, a proper understanding of their structure and mechanism will depend on knowledge of the structures of the individual parts and of the kinetics and thermodynamics of their interaction. In both instances, techniques of modelling the three-dimensional docking of protein structures are required. Solution techniques of studying protein-protein interaction are also essential (Appendix 2). New developments of traditional methodologies, such as small angle X-ray and neutron scattering, ultracentrifugation and microcalorimetry, have emerged, and new approaches, such as surface plasmon resonance, have been devised that enable real-time kinetic and thermodynamic data to be derived. New developments in neutron spectroscopy make it possible to quantify protein dynamics underlying functional flexibility in complex samples, including the situation *in vivo*.



The three-dimensional structure of a DNA translocating machine at 10 A resolution. Structure, *7*, 289-296. J.M. Valpuesta, J.J. Fernandez, J.M. Carazo and J.L. Carrascosa (1999).

Much effort is now going into the identification of the partners in protein-protein interactions on which the proper functioning of the cell depends, for example in the signal transduction processes that control the cell cycle, apoptosis or hormone stimulation. The identification of such partners is made by a

proteome approach or by the yeast two-hybrid screens. Other methodologies, such as fluorescence resonance energy transfer, can be envisaged to extend the analysis to the interaction *in vivo*. It is the quantitative data obtained in this way on which a proper understanding of, say, a signal transduction process, and its relationship to other signal transduction pathways, must ultimately rest.

Another important area of protein-protein interaction is that of motility, where great strides are being made in the molecular understanding of muscle contraction and cellular transport processes. The investigation of single molecules, made possible by the invention of techniques such as molecular tweezers in the study of actin and myosin, is a notable example. The increasing understanding of molecular motors and of multifunctional protein machines in general, is of value not just in its own right but in the prospect of generating molecular devices based on these newly discovered design principles.

Glycobiology

Many proteins are found as conjugates with bound carbohydrate. Such proteins are always extra-cellular and some of the most important are involved as hormones in signal transduction processes or as cell surface proteins in cell-cell interaction. Indeed, unless the proper carbohydrate moiety is attached, the proteins will not exert their proper biological function at all, since much of the essential specificity resides in the nature and fine structure of the carbohydrate component. The addition of the carbohydrate takes place at the post-translational stage and is brought about by the action of individual enzymes, each capable of catalysing one step of the overall biosynthesis and acting sequentially in different permutations and combinations on different target proteins.

The presence of carbohydrate in glycoproteins can be a major impediment to the determination of the structure of the protein, hindering or even preventing crystallisation and interfering with the NMR analysis if that is the chosen route. Moreover, the generation of glycoproteins poses major difficulties for the recombinant DNA approach to protein production: bacteria do not normally attach carbohydrates to proteins. Thus one is limited to generating the recombinant protein in an eukaryotic cell, and even then the correct carbohydrate component required for the ultimate biological function may not be added. Considerable progress is being made in the determination of the structures of the carbohydrate components of glycoproteins and in the analysis of the biosynthetic pathways involved. Clearly a major goal of bioinformatics, as with other post-translational modifications, must be the ability to predict the sites of carbohydrate addition and the nature and structure of the carbohydrate added. This will impact also on the proteomics field, where it is of course the glycoprotein, with carbohydrate attached, that exerts its biological activity *in vivo*.

Integrative biology

Detailed knowledge of the structure, dynamics and function of individual proteins is essential for a better understanding of the function of the living cell or the whole organism throughout its life cycle. In physiology, for example, modern biology now opens up totally new opportunities for linking genomic information and the regulation of gene expression with the role of proteins and their regulation in the context of complex cellular

processes such as metabolism, differentiation and cellular motility, to mention just a few examples. Gene regulation, feed-back regulation of metabolic processes, and signal transduction processes governing cellular differentiation all require detailed and precise knowledge of how proteins function and how these functions can be modified by the local physical environment. Knowledge of the structure and the function of proteins has to be supported by bioinformatics to help in understanding the complex multitude of molecular interactions that govern and execute the various functions of the living cell, whether the cell be engaged in motility as exemplified by the muscle, specialised to convert light into chemical energy as in green plant photosynthesis, or involved in host-pathogen interactions as for bacteria with other organisms such as man. In medicine and in various biotechnological applications, information on protein structure and function is opening up the diagnosis and treatment of various diseases such as Alzheimer's, spongiform encephalopathies, cystic fibrosis, etc., as well as the engineering of novel proteins for a wide range of purposes such as new biosensors and the synthesis of new antibiotics.

Revisit of the 1998 SR Review

The past three years

In 1997 the European synchrotron facilities were asked to report on the beam-lines and facilities available to biological users and the typical usage for biology at that time. The results were tabulated in the 1998 ESF review of the *Needs for European synchrotron and related beam-lines for biological and biomedical research*. The synchrotron sites have been asked briefly to update their current facilities and a summary is attached in Appendix 6. The detailed reports are not repeated here.

Two major changes have transformed the situation since 1998. First, the four QUADRIGA beam-lines at the ESRF have been fully commissioned and are now in routine use. This alone has had a massive effect on SR usage for protein crystallography in Europe. In addition, the two new crystallography wiggler lines are in use at SRS, together with fully commissioned lines at ELLETTRA, MAXLAB and SLS. New lines are under construction at ELETTRA and MAXLAB.

Second, efficient and reliable commercial CCD detectors have been installed on the ESRF beam-lines and at several other sites. This has enormously reduced the read-out time per image from the time-scale of minutes to a few seconds. As a consequence, in particular at the ESRF, the time required to collect a complete data set on a sample has fallen, in some cases, from several hours toless than half an hour.

The throughput of such third-generation beam-lines has shown exactly the type of increase foreshadowed by the 1998 ESF report. It is clearly possible to record 20 complete data sets per day, equivalent to 4000 per annum on a single beam-line. Even allowing for a failure rate of 50%, somewhat pessimistic with present pre-screening and cryogenic skills, this implies that each line should be capable of providing 2000 data sets or more annually.

The future

The pressure on beam-lines has thus been ameliorated in the short term, but it will rise again to unacceptable levels as post-genomic projects lead to increased numbers of crystalline samples. The needs for advances in technology and resources is apparent in several directions:

- Automated sample changers on the beam-lines. These are already under development at the ESRF as well as in the USA.
- Fast and intelligent software for image evaluation and assessment, providing rapid feedback to the user regarding the usefulness of the current crystal.
- Yet faster and more sensitive detectors. Pixel detectors are likely to be the next generation of X-ray detectors on the beam-lines and funding for their development must be ensured. This will be especially important for handling the weak and sharp data from microcrystals.
- Provision of more user-friendly and unified graphical user interfaces which control the whole experiment, including beam-line optics, sample alignment and detector, and provide a reliable record of the whole experiment in the image and subsequent data files.
- Fully automated crystallography suites for structure analysis and refinement, allowing the user to ascertain whether the

appropriate data have been recorded while the samples are still at the synchrotron site.

- Remote access synchrotron usage. This implies that most of the above-mentioned automation is already in place. Transport of vitrified crystals in Dewar vessels to a synchrotron facility, collection of data by remote user or by a scientist based at the site, will in all probability be essential to handling the numbers of samples anticipated in the future.
- The planned third-generation synchrotrons with a number of beam-lines dedicated to life sciences, such as DIAMOND, SOLEIL and LLS, are vital in this scenario. The new lines will allow European scientists to maintain their position in structural biology.

Continued coordination between the various facilities through, for example, the European Commission Round Table is essential. This helps in ensuring that beam-lines are developed with the appropriate, sometimes complementary, parameters such as wavelength range and tunability.

Multidisciplinary working group on Protein Structure and Function

- **Professor Gunnar Öquist (Chairman)**
  Umea University
  Deparment of Plant Physiology
  Sweden

- **Professor Francesc X. Aviles**
  Autonomous University of Barcelona
  Department of Biochemistry
  Spain

- **Professor Gunnar von Heijne**
  Stockholm University
  Stockholm Bioinformatics Centre
  Sweden

- **Dr. Diane McLaren** succeeded by **Dr. Mark Palmer**
  MRC Head Office
  Research Management Group
  United Kingdom

- **Professor Richard N. Perham**
  University of Cambridge
  Department of Biochemistry
  United Kingdom

- **Professor Reidun Sirevag**
  ESF/LESC member
  and University of Oslo
  Biological Institute
  Norway

- **Professor Joël Vandekerckhove**
  Ghent University
  Department of Medical Protein Research
  Belgium

- **Dr. Keith S. Wilson**
  University of York
  Protein Structure Group
  United Kingdom

- **Dr. Giuseppe Zaccai**
  Centre National de la Recherche Scientifique – Commissariat à l'Energie Atomique
  Institut de Biologie Structurale
  France
  and
  Institut Laue Langevin
  France

European Science Foundation:

- **Dr. Marianne Minkowski**
  Senior Scientific Secretary for Biomedical Sciences

Appendix 1

Group I Report

Protein Production in the Structure-Function and Structural Genomics Fields

20 March 2000, Frankfurt Airport Conference Center
(updated in March 2001)

***Participants*:**

Francesc X. Aviles (Univ. Autonoma Barcelona, ES) ***(Chairman)***
Arie Gerlof (EMBL, Heidelberg, DE)
Marianne Minkowski (ESF-Strasbourg, FR)
Luis Moroder (MPI-Biochemie, Münich, DE)
Luis Serrano (EMBL, Heidelberg, DE)
Andreas Plückthun (Univ. Zürich, CH)
Anthony Watts (Univ. Oxford, UK)

Summary

The thematic group was comprised of scientists who have been and are very actively involved in the systematic and preparative production of proteins, protein variants or labelled forms for structure-function (X-ray, NMR, other biophysical approaches, etc), folding or engineering studies, and for biotechnological or biomedical applications. It was unanimously agreed that the efficient production of such molecules, in a properly folded state, is an essential step in the above mentioned studies and applications (frequently the first step), and that not sufficient attention is paid to it at the levels of research, providing of facilities, training and finance. Although many biologically based or chemically based methods have been developed in the last two decades for the non-natural production of proteins, there is neither a general and precise single strategy to select such methods for any particular case, nor data banks or facilities from which the non-specialised researcher can devise a reliable planning. The high-throughput systems for the structural analysis of proteins planned for the present large-scale structural genomics projects will be seriously hampered without first solving the problem of producing well-folded proteins in a fast and efficient way. If the present trends are not modified, the massive structural information gathered from these projects might be biased towards proteins, which are easier to produce and analyse.

Several proposals and general comments arose from the meeting:

Practical courses

The present state-of-the-art techniques for protein production require a review and systematisation effort, and should be regularly disseminated through the organisation of practical courses and meetings at the European level.

Workshops/Symposia

The organisation of workshops or symposia dealing either with the whole field, as an overview, or with specialised subjects (that is protein expression and complementary molecular biology in *E. coli*, in baculovirus, in 'cell-free translation' systems, membrane protein production, etc.) would all be very valuable. The complementation of these subjects with protein engineering approaches (redesign of stability, specificity and biological action, chimeras, minimisation, etc.) is desirable, given that many engineered proteins are required nowadays for structure-function studies and biotechnology uses and specific proposals are given at the end of the report.

Facilities

The establishment of facilities/services specialised in such a subject, either in a few selected sites (that is, in certain leading research institutions), or spread in many research centres, would be a positive action. Ideally, such facilities should play a supporting role by storing and disseminating information and valuable material (such as building data banks, advising about adequate strategies, providing proper and efficient vectors, organising practical courses and meetings), rather than being directly involved in the production of particular proteins.

Research priorities

Such actions, together with the establishment of appropriate priority lines dealing with protein production and engineering in the research programs funded by the ESF, EC and at national level, should stimulate European investigators to focus their research in this field, and make it valuable by itself, not mainly dependent on other research fields.

Specific comments on new or particularly valuable approaches or acute needs were also collected, in order to propose their inclusion in practical courses, workshops and priority lines. Noteworthy among these were:

- the increasing use and potential of cell-free translation systems or membrane proteins produced *in situ*;
- the ligation-based approaches for the segmental specific labelling of proteins with given isotopes or non-proteinogenic probes;
- the need for efficient methods for the production of large multidomain proteins and large protein complexes.

Introduction

At the ESF Multidisciplinary Working Group meeting on the Protein Structure and Function field, held at the Heathrow Airport (London, UK) on 7 December 1999, the topic 'Protein production' was selected as the one out of five that require stimulation measures to potentiate the field. The thematic group that was subsequently set up to evaluate this topic met at the Frankfurt Airport on 20 March 2000.

All members of the group have been and are very actively involved in the systematic and preparative production of proteins for structure-function studies, at the both the fundamental and biotechnological/ biomedical levels, dealing with a wide array of molecules with different properties (for example., soluble proteins, single-domain/multidomain, disulfide-lacking and disulfide-rich glycoproteins, protein complexes, membrane proteins, etc.), which usually require different methodological approaches for their production. Also, they are greatly experienced in the production of protein engineered variants (mutant, labelled, cyclised, chimeric, minimised forms, etc.), which are of great use nowadays to purify and characterise proteins, facilitate their structural-functional analysis, and obtain molecules with different properties.

1. Subjects of concern for protein production in the European context

A general agreement was reached that an efficient production of proteins or protein variants, in a properly folded state, is required for the high-throughput approaches devised for the structural and functional studies to be carried out in the large scale structural genomics projects. This is the first step in such projects, which are usually quite demanding in terms of the quality and quantity of the required protein. It is evident that, unless general approaches for the non-natural production of proteins are developed, or the present ones are systematised better, a detailed structural information of an important percentage of the proteins that should be analysed in such projects -those which are difficult to produce- will probably never be obtained. In fact, such a constraint has biased some of the launched projects, in the sense that they are restricted to proteins which are easy to produce and crystallise, a restraint that will clearly limit the expected benefits and derived structural data-bases. This is therefore a clear bottleneck in the structural genomics projects world-wide.

It is worth remembering that in such projects (1,2,4) there is a trend towards 'robotisation' of the different steps usually followed in the functional characterisation and structure determination, such as bioinformatic analysis (that is with sequence and conformational predictions carried out over hundreds of proteins, in a quite automated way), differential expression mapping (at the nucleic acid and protein levels, with microchips, 2D-PAGE gels, MassSpec., etc.), detection of protein-protein interactions, protein crystallisation, X-ray and NMR data collection and interpretation, and comparative analysis of the derived structures and functionalities. However, the step with the poorest prospects for automation is the large-scale production of proteins or derived forms. Only the further development of the recombinant expression methodologies based in 'cell-free translation systems' seems to bring hopes for the solution of this problem.

It is frequently necessary to engineer a protein to make it suitable for structural characterisation. At present however, obtaining non-natural proteins either by biologically-based heterologous expression or by chemical synthesis is still considered a time-consuming and labour-demanding task subject to many uncertainties. It frequently requires the assay of many different variables (vectors, hosts, codon-usage, fusion to solubilisers and carriers, co-expression with folding helpers etc, in recombinant systems, and orthogonal strategies, blocking and de-blocking strategies etc, in chemical synthesis). This is particularly acute when the proteins are very large, not compact, structured in flexible domains, with different hydrophobic characters, and containing a significant number of disulfide bridges and/or post-translational chemical modifications, such as carbohydrates, characteristics found in many cases. One particularly well-documented and difficult problem is the production of membrane proteins, a task that fulfils several (sometimes all) of the above criteria. In fact, obtaining such proteins in a preparative, well-folded and homogeneous state is still one of the key reasons for the very limited number of full characterised 3D structures. Indications that at least 20% of the genes encode for membrane proteins in most cell types (3), and that these are very important for cell regulation and drug design, make solving such problems yet more urgent.

In spite of the aforementioned difficulties in obtaining many proteins, and the resulting detrimental effect in their structural analysis, there is a general lack of appreciation for such approaches in the scientific community. This is evidenced in the lower impact index attained by publications mainly dealing with protein production approaches, and in the lack of acceptance of the corresponding research papers - even when they describe new productive approaches - in high-impact journals. As striking is the lack of prominence of the researchers responsible for the production of proteins in the authorship of collaborative papers with colleagues specialised in X-ray, NMR or other structural approaches. In certain cases, the production of a given protein, protein variant or protein complex suitable for high resolution structural studies, can take years of dedicated work, yet researchers receive little recognition when compared with other efforts involved in the structure-function determination.

The lack of actual scientific appreciation and rewards for the tasks associated with the production of proteins is being exacerbated in Europe by the practical disappearance from the EC Framework programmes of priority lines and specific finance devoted to the development of new strategies and tools for the production of proteins. This is reflected in the priority lines of programmes for financing research at the national levels. Quite a number of academic groups working in such a field, or related ones, are surviving by being associated (as protein producers, in a secondary role) with projects related to the structural-functional analysis of given molecules or in association with the solution of a given biotechnological or biomedical problem, in the above mentioned devaluated manner. Fortunately some industrial research is still kept in such a direction in Europe, although this suffers from the problems of confidentiality, patenting and time-delays in relation with the dissemination of the information and improvements. In this respect, the USA-based industrial research is much more beneficial for this field because it involves relatively small biotech firms, which are interested in quickly commercialising kits for the easy production of proteins or protein variants. However, such offers do not solve the problem of large-scale production of difficult proteins.

2. Present large-scale projects, trends and experiences to be taken into account

The programmes directly or indirectly related to the support and financing of structural and functional genomics and proteomics large-scale projects, launched in different developed countries (USA, Japan, Germany, EC, etc.)(5-12), have clearly considered the need to stimulate the development of technologies for the production of proteins and protein variants (such as isotope-labelled forms), although with different emphasis. The issue has been partially circumvented in certain projects through the selection of protein targets belonging to organisms from which they can be isolated or produced easily. Why? Because they are simpler organisms or they contain proteins, which are more stable, easily foldable, more soluble and easily over-producible in the same organism (such as thermophilic bacteria), or they are usual hosts for recombinant protein production (such as yeasts). Most of these projects explicitly or implicitly consider that the analysis of proteins, which are difficult to produce or crystallise, such as membrane proteins, should be left for a second stage. This does not take into account the fact that they are an important fraction of the protein world and probably contain new and specific folds. Nevertheless, it is worth mentioning that in 1998 the NIH launched a specific research on the structural biology of membrane proteins (13-14), in which the development of improved methods for the over-expression of native and modified proteins was one of the goals.

One of the main structural genomics projects in Japan, based in the RIKEN Center (11), intending to produce an important part of the required proteins in a 'cell-free translation system' is more innovative in this context. Theoretically, at least, such a procedure should be less dependent on the particular type and origin of each protein to be expressed, and greatly facilitate the production of isotope-labelled material. Several reports indicate that it is actually feasible to produce simple wild type or isotope-labelled proteins by such an approach. However, by now it seems to be quite expensive, very difficult to extend to a high-throughput scale, and still not tuned for large multi-domain structures. In spite of this, it is a very promising methodology, which merits further efforts for its development (see section 6).

In Europe, several structural genomics projects (previously considered, or already launched or in a very advanced stage of preparation) have different approaches with relation to the material to be analysed, the way to produce it, and the priority of this research field. The Berlin consortium group (the Protein Structure Factory) is mainly working with proteins produced in thermophilic organisms or in *E. coli*-based systems. A general structural genomics project was considered in Switzerland: this decided on a more focussed approach to study particular structural families of direct biological importance because it would be a more economic approach, given the realistic available resources. Nevertheless, high-throughput protein production has been identified as a clear area where further research is required. In the UK, protein production is being considered as a supporting activity, and a facility site being sought. The First International Structural Genomics Meeting, held in Hinxton (UK) on April 2000 (8a), did not suggest a restriction in the types of protein targets selected, besides their having to be a representative set of macromolecular structures, including medically important human proteins and proteins from important pathogens. One of the goals emphasised at the meeting was the development of high-throughput methods for the production of target proteins suitable for structure determination, particularly for membrane proteins. Very similar conclusions were reached in subsequent meetings, such as the Structural Genomics Conference: *from Gene to Structure to Function*, held in

Cambridge (UK) in September 2000 (8b), and the International Conference on Structural Genomics, held in Yokohama (Japan) in November 2000 (8c).

In the structural genomics context, it is worth commenting on recent reports (15,16) relating to the preliminary results of a high-throughput project for heterologous expression and purification of 500 proteins from a thermophilic bacteria (*Methanobacterium thermoautotrophicum*) in *E. coli*: it was found that only about 15% of small proteins (less than 200 residues) and about 10% of large proteins were produced and purified in an adequate state for high-resolution structure analysis. This means that a great deal of effort is still required to improve and systematise such productive methods.

Two additional valuable conclusions were also derived from those reports (15,16), the fact that the percentage of produced soluble proteins was much larger in single domain than in multidomain proteins, and that the co-expression of protein partners in oligomeric complexes strongly increased the solubility of each protein. This emphasises the importance of parallel protein-protein interaction analysis, and the difficulties and hopes for the structural analysis of large proteins and protein complexes (see section 5).

3. Feasibility, advantages and advisable design of facilities and training sites for protein production

To what extend would the creation of central facilities for the production of proteins be successful? Although there are quite a number of research centres that have large-scale facilities for culturing cells in bioreactors, as a means to overproduce proteins, the number of facilities that are involved in the production of a wide array of proteins is much smaller. This is probably due to the fact that, as mentioned above, there is no general and precise strategy to select the approach to produce a protein by biological (that is recombinant heterologous expression) or chemical- based methods. A detailed knowledge of the properties of a given protein, and the assay of various productive systems (vectors, hosts, codon-usage, co-expressed helpers, etc.) is by now the most effective way to overproduce a protein, and this is feasible only with the direct involvement of an expert research group in such a problem. Some structural genomics projects use high-throughput expression methods, adequate for facilities but, as mentioned in the previous section, they have a limited success. The recombinant methodologies based on 'cell-free translation systems' seem to be the more promising for use in facilities, but would require further investigation. Only a few industrial firms have a long experience in the establishment of internal facilities for protein production. However, in spite of the rigid division of tasks in such organisations, they do not seem to be more successful than academic institutions for difficult proteins, as the problems are usually multidisciplinary and require an integrated approach leading from cloning to expression, purification and crystallisation. Furthermore, the industrial style of dropping a project when milestones have not been reached in time may not be suitable to work out intrinsically complex problems and challenges.

However, it seems feasible and highly beneficial to establish such protein production facilities in public research centres if their role is to provide advice to research groups about the best and most advanced methodologies in this field, and in storing and disseminating related information and valuable material (for example, building data banks and informative web sites, providing proper and

efficient vectors, strains, genetic constructs, organising courses and meetings), rather than being directly involved in the production of particular proteins. Another useful task could be the performance of comparative analysis between different methods and methodological variants to produce proteins, in collaboration with research groups, and publish and store them, making the best advice available. Also, computer-based approaches to find the most appropriate experimental strategy to produce a given protein (that is to detect those regions with low propensity to fold, based on sequence analysis, that should be eliminated to avoid expression and folding problems), should be encouraged. It would also be very useful to have advice on expression systems not subject to patent restrictions, an issue by itself of potential importance in large-scale production facilities.

Such facilities would probably be more easily hosted and established in large and leading research institutes, but they could also be established in small research centres. A major responsibility for such facilities (particularly those in the former centres) would be the regular organisation of practical courses on protein production, and to host researchers from other institutions and countries for short stays to train them in this field. These facilities, courses and visits, should be financed by EC, ESF and other transnational and national agencies. A facility of this kind and with similar aims has been recently created at the EMBL in Heidelberg (17).

All the participants in the thematic group meeting agreed that the efficient collection of the current and future information about the production of proteins by biological or chemical-based approaches is essential. These data must be stored in repository banks, and its dissemination to the scientific community, together with a parallel systematisation, would be essential tasks for solving the main problems in such approaches, and facilitate their high-throughput applications. The generation of a white or blue-book collating such information, and the links to expert groups in Europe, would also be a desirable initiative.

4. Organisation of general and specific meetings

An excellent way to facilitate the systematisation, update and dissemination of the knowledge about protein production would be through the organisation of meetings on this field, either in workshops with a specialised audience or in larger attendance symposia. These meetings would be positive at both the general level, by providing overviews of the field (particularly at the beginning and at the end of launched initiatives), and at the specialist level, providing information on specialised subjects such as protein expression in *E. coli* and related molecular biology, and similar ones on baculovirus or eukaryote expression systems, cell-free translation systems, membrane proteins, etc.

The need for such meetings is clearly exemplified by the most well known and still most frequently used expression host, *E. coli*. In spite of the abundance of literature about, many aspects of its molecular biology related to protein expression have never been fully understood and require regular updating and systematisation: replicons, promoters, mRNA half-lives, nucleases, proteases, chaperones, disulfide isomerases, and so on, as well as many aspects of fermentation and purification. *E. coli* should be re-visited! Experts in each bottleneck should be identified and gathered in meetings to remedy this deficiency. The stimulation and financing of all these meetings by appropriate European organisations would be very valuable. A definite proposal for instituting a series of meetings is given at the end of this report.

5. The delayed problem of the production and purification of large multidomain proteins, and large oligomeric complexes

Although most of the present high-throughput structural genomics projects concentrate their efforts on the production and analysis of small or medium size, relatively simple proteins, it is evident that many interesting larger proteins or protein complexes are omitted. This is mainly due to the difficulties of producing such proteins in well folded/soluble and active state, in sufficient quantities, rather than to the limitations of the present methods for analysing their structure. Although high-resolution NMR in solution is limited to medium-size proteins, solid-state NMR, X-ray crystallography and cryo-electron microscopy are able to tackle with very large structures.

Given that many interesting proteins, which are essential for the comprehension of the structure, function and dynamics of living organisms and for biotechnological applications, are either very large or involved in protein complexes, the systematic production of such complexes in a stable and functional state is a delayed goal that should be fulfilled as soon as possible. In fact, the NIGMS-USA recently launched a research programme (18,19) on very complex biological systems, that aims to deal with the above problem, among others, derived from the intrinsic complexity of living systems.

The use of transfection and over-expression methods that allow the production of large and difficult proteins in locations and organisms similar to the ones in which they are naturally found or even expressed in an amplified way, is a great advance in the proper production of such proteins. However, the subsequent purification and handling of such material is still difficult, as in the case of the production of membrane protein receptors *in situ*. In this context, the use of new approaches that allow the gentle isolation of labile complexes, such as those recently developed by EMBL researchers (20), based on tagged proteins and affinity columns, could be of great use in this field.

6. Comments on particularly promising approaches or acute problems

Comments and suggestions were made on several topics related to particularly difficult protein types and novel, potentially powerful, approaches, which would require significant efforts and stimulation measures in the field of protein production:

- disulfide-rich proteins;
- membrane proteins;
- large multidomain and oligomeric proteins;
- cell-free translation systems;
- ligation-based approaches for segmental labelling.

The disulfide-rich proteins, which frequently give rise to misfolded forms or inclusion bodies, particularly when over-expressed in *E. coli*, is one of the challenging cases but not the most severe. Such misfolding problems of difficult fitting in high-throughput productive strategies, can be partially alleviated by using different approaches: for example, special *E. coli* strains favouring disulfide formation, co-

expression with disulfide isomerases or other disulfide forming catalysts, the use of wall-less bacteria (21) as expression systems, heterologous expression in higher eukaryotes, *in vitro* refolding, and so on.

In the case of membrane proteins, particularly those with large membrane-embedded parts or multidomain structures of different hydrophobicity, the problem is much more acute, and their production usually requires large efforts not easily automated. In several cases where membrane proteins led to high-resolution structural data, they were obtained from natural (non-recombinant) systems, solubilised (and even crystallised) in the presence of non-chaotropic detergents, lipids, lipidic 2D-arrays (for cryo-EM) or lipidic cubic phases (for X-ray crystallography). The very recent case of the high-resolution 3D structure derived for Ca-ATPase from sarcoplasmic reticulum is a good example of such methods. Unfortunately, these methods have not proved to be universal, by now.

In recent years a few membrane proteins have been produced by recombinant transfection and over-expression approaches *in situ*, in the membranes of cell lines similar to the species in which they are genetically encoded, or by amplified expression in the same ones. In such environment, membrane proteins can be produced in a well-folded and active state (22), although rarely at the amplified levels, which are required for structural studies. Besides, unless they are structurally and functionally studied in such a location when produced in sufficient quantities (that is, by EM or solid state-NMR), they usually suffer significant conformational changes and aggregation when extracted and transferred to simpler environments (that is, artificial membranes, liposomes, etc.) or solution media for high-order crystallisation trials or high-resolution structural analysis. Frequently these proteins are found to be heterogeneous, as a result of post-translational modifications, such as glycosylations for eukaryotic proteins. Such heterogeneity is one of the main problems for structural studies. The avoidance of such modifications, when not essential for the folding and activity, might improve the potential of this approach.

One of the few successful high-level amplified expression systems for large integral membrane proteins has been designed for sugar transporters (by Henderson, Poolman and colleagues)(23). Here, an ideal 20-50% amount of total membrane protein is expressed, his-tagged and can be purified in large (over 10 mg) amounts for biophysical studies. Since this protein is expressed in *E. coli*, it can be isotopically labelled (24) either with non-perturbing isotopes for NMR and/or neutron spectroscopy works, or presumably with non-natural labels (for example Se-Met) for crystallographic studies. Since *E. coli* is metabolically lazy, either uniform or specific labelling of all residues of one or two type(s) can be achieved. The success of this expression is almost certainly due in no small part to the non-pathological nature of the protein and the resemblance of proteins in nature to the bacterium, but also relies on well-chosen promoters. What is now required is that similar expression strategies be adapted for other proteins of wider diversity and sources, and for expression systems such *E. coli*, in view of their versatility and adaptability.

An important case for such approaches is the vital 7TMD-GPCRs (seven transmembrane domain G-protein coupled receptors), which constitute the major family of mammalian receptors as judged from genome analyses. They have not yet been expressed in large (over 10 mg) amounts, although some research groups have successful expression in *E. coli* at low levels. This family of proteins is a major focus for protein production and structure-function analysis both in the academy and pharmaceutical industry fields, given their tremendous biological and biomedical implications.

In the longer term, newer methodologies may arise both for expression and isolation and purification of membrane proteins. The needs are however different according to the analytical methods to be used, some requiring highly purified membrane proteins, such as crystallography, whilst other are successful with impure or mixed protein environments, such as solid-state NMR and activity measurements.

In a conclusion, much more research should be devoted to the topic of membrane protein production, particularly when it is taken into account that probably at least 20% of genes encode for such proteins (3), and that new folds could arise from them.

In recent years, much hope has arisen from the use of cell-free translation systems for the universal recombinant production of proteins, independently of the origin, codon usage and sequence of the nucleic acid taken as a template, and without the addition of post-translational modifications. Several reports indicate that it is feasible now to produce small wild type or isotope-labelled proteins by such an approach (25). However the caveat is that it is very difficult to extend this approach to a high throughput scale, because of the enormous amounts of extract needed. The main problem is that all state-of-the-art translation systems, including those based on flow reactors, use extracts which are about 10% as concentrated as the *E. coli* cytoplasm in terms of ribosomes and all other factors. As a consequence, to obtain the same productivity as that of *E. coli* cytoplasm, 10 times more cells are needed to produce the extract containing the ribosomes than if the expression was directly carried out *in vivo*. Therefore, while it is conceivable to produce proteins by *in vitro* translation for structural studies, to do so for a large number of proteins in structural genomics would require an incredible effort in production of ribosomes and co-factors, with very large costs. In spite of such difficulties, the RIKEN structural genomics project (11) uses this technology extensively, and they claim that it has been automated to produce in parallel hundreds of different proteins. Nevertheless, in this project it has also been shown that the small single-domain proteins are much more easily produced than the large, multidomain proteins. Significant research and improvements are therefore still required to make this approach of general utility, as well as to decrease its present high financial costs.

The labelling of proteins with isotopes or probes is one of the requirements for certain types of high-resolution structural analysis of proteins (either in solution, in crystals, or *in vivo* systems), by methods such as NMR, EPR or X-ray (nD-NMR, MAD, etc.). Such labelling or substitution is usually performed at sites where certain residues are located (that is, Met, Cys, Trp, etc.), frequently giving rise to a dispersion of the label or probe throughout the protein, unless a single residue type is present in it. For years, it has been feasible to concentrate the label in a given region of a protein, if it is synthesised by chemical methods, but this approach is restricted to small proteins, in the practice. This approach has been recently extended to large proteins. A potentiation of semi-synthetic methods may derive from "intein-based" ligation approaches (26) that allow the precise ligation of different regions or domains of a protein that were previously obtained by chemical or recombinant DNA methods. This should allow us to specifically visualise regions or domains of larger proteins by the above mentioned analytical methodologies, providing a promising strategy for studying large protein structures in solution, when associated in oligomers, or in large macromolecular environments (such as artificial or reconstituted membranes).

The production and structural analysis of large multidomain or oligomeric proteins (specially mentioned in section 5) would greatly benefit from the development of several of the approaches previously

mentioned in the present section. However, even if these approaches improve significantly, it is unlikely that the production and structural-functional analysis of these proteins can be easily accommodated to the high-throughput strategies of the structural genomics projects. In fact they may not be essential targets for such projects, because the main protein folds probably will already be identified in single-domain structures, or in oligodomain structures dissected by excision, production and structural analysis of the individual domains in isolation, although solubility problems might arise for this approach (as mentioned in section 2). Nevertheless, the great importance that large multidomain and oligomeric proteins have in the understanding of the living macrostructures and processes, fully justifies the efforts that could be made in their production and analysis.

The various issues treated in this section, among others, would be important topics for the repository data banks, practical courses, workshops and symposia recommended in sections 3 and 4. They also should conform to the EC and European national agencies' priorities included in future funding of research programmes to properly stimulate the protein production, structure-function studies and structural genomics fields.

Acknowledgements

The contributions from the different members of the ESF Study Group have been essential for the productive development of the meeting and the preparation of this report. The suggestions, comments and information provided by Prof. T. L. Blundell and by Drs. K. Mizuguchi, M. Hyvonen and K. Phillips (Dept. Biochemistry, Univ. Cambridge, UK), have also been very useful and are most acknowledged.

Proposal of an initial series of meetings within the protein production and related fields

Time schedule: two for the first year (1 and 2), and two for the second year (3 and 4)

Format: Symposia or workshops, according to the financing possibilities and supporting institutions.

1. Protein Production and Engineering: An Overview
2. Expression in *E. coli*: where do we stand?
3. Protein Expression in Eukaryote Systems
4. Membrane Protein Expression for Structural and Functional studies.

Bibliographic and electronic-based references

1-4 – Overviews and basic references in the scientific literature :

- Service R.F. "Structural genomics offers high speed look at proteins". (2000), Science, *287*, 1954-1956
- Blundell T.L. & Miziguchi K. "Structural genomics: an Overview". (2000), Progress Biophys. & Mol.Biol, *73*, 289-295
- Adams M.D. "The genome sequence of *Drosophila melanogaster*". (2000), Science, *287*, 2185-2195.
- Supplement of Nature Structural Biology, on "Structural Genomics, vol 7, Nov. 2000

5-12 – WEB addresses with information about the present structural genomics projects (USA, Europe, Japan):

http://grants.nih.gov/grants/guide/pa-files/pa-99-117.html
http://www.nigms.nih.gov/funding/psi.html
http://www.nigms.nih.gov/news/announcements/psi_international.html
http://www.nigms.nih.gov/news/meetings/hinxton.html
http://www.cordis.lu/fp5/home.html
http://www.esf.org
http://www.gsc.riken.go.jp/
http://www.sta.go.jp/life/e-life.html

13-14 – WEB addresses with information about the NIH-USA initiative on structural biology of membrane proteins:

http://grants.nih.gov/grants/guide/pa-files/PA-99-004.html
http://www.nigms.nih.gov/news/announcements/structbio.html

15-16 – References about the limitations of the high-throughput expression approaches used in structural genomics:

● Christendat D., Yee A., Dharamsi A., Kluger Y., Gerstein M., Arrowsmith C.H. & Edwards A.M. "Structural proteomics: prospects for high-throughput sample preparation". (2000), Progress Biophys. & Mol. Biol., *73*, 339-345 (Analysis of a research project of expressing 500 proteins from a thermophilic bacteria).

● Edwards A.M., Arrosmith c.H., Christendat D., Dharamsi A., Friesen J.D. Greenblatt J.F. & Vedadi M. "Protein production: feeding the crystallographers and NMR spectroscopists". (2000) Nature Struct. Biol., *7*, 970-972.

17 – Information about facilities for protein production set up or planned in Europe:

The EMBL-Heildeberg facility (Drs. Arie Gerlof & Luis Serrano, personal communication).

18-19 – WEB addresses with information about the NIH-USA initiative on the study of complex biological problems:

http://www.nigms.nih.gov/news/reports/complexbio.html
http://www.nigms.nih.gov/news/announcements/complexity.html

20-26 - References about particularly promising approaches or acute problems in the field of protein production:

● Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Séraphin, B. "A generic protein purification method for protein complex characterisation and proteome exploration". (1999), Nature Biotech., *10*, 1030-1032 (Purification of protein complexes).

● Gumpert J. & Hoischen C. "Use of cell wall-less bacteria (L-forms) for efficient expression and secretion of heterologous gene products". (1998), Curr. Opin. Biotechnol., *9*, 506-509 (high expression of proteins, particularly disulfide-rich).

● Biochemical Society Transactions on the Colloquium on "Expression and Purification of Membrane Proteins" (669th BS Meeting, Univ. Keele), (1999), *27*, 883-962 (15 papers).

● Watts A. "NMR of drugs and ligands bound to membrane receptors", (1999), Current Opin. Biotech., *10*, 48-53.

● Ward A., Sanderson N.M., O'Reilly J., Rutherford N.G., Poolman B. & Henderson P.J.F. "The amplified expression, identification, purification, assay and properties of histidine-tagged bacterial membrane transport proteins", in 'Membrane Transport - a Practical Approach', (2000), Chapter 6, pp. 141-166. Blackwell's Press, Oxford, UK.

● Kigawa T., Yabuki T., Yoshida Y., Tsutsui M., Ito Y., Shibata T. & Yokoyama S. "Cell free production and stable-isotope labelling of milligram quantities of proteins". (1999) FEBS Lett. 442, 15-19.(protein production based on cell-free translation systems).

● Otomo T, Ito N, Kyogoku Y. & Yamazaki T. "NMR observation of selected segments in a larger protein: central-segment isotope labeling through intein-mediated ligation". (1999), Biochemistry, *38*, 16040-16044 (Protein labelling by the intein-ligation method).

Appendix 2

Group II Report

Biophysical Chemistry Methods in a European Context

***Participants*:**

Richard N. Perham (University of Cambridge, UK) ***(Chairman)***
Alan Cooper (University of Glasgow, UK)
Stephen E. Harding (University of Nottingham, UK)
Leonard C. Packman (University of Cambridge, UK)
Giuseppe Zaccai (CNRS–CEA, IBS & Institut Laue Langevin, FR)

Executive summary

Methods in biophysical chemistry have paved the way for the success of structural biology in recent decades. To the conventional methods based on fundamental discoveries from the nineteenth and early twentieth centuries, have now been added completely new approaches such as Atomic Force Microscopy or Surface Plasmon Resonance. However, probably because of spectacular progress in high-resolution X-ray crystallography and NMR, the promotion of the other methods as well as the support for major associated instrumentation projects has been neglected, especially in Europe. The challenge of the post-genomic era lies not only in solving the high-resolution structures of as many gene products as possible, but also in addressing questions to bridge the gaps between molecular and cell biology and between thermodynamics and structure. In fact, the aim of understanding molecular function and signalling within the living cell is more likely to be attractive and stimulating to bright young scientists. Questions such as - How/where are proteins and other macromolecules located in the cell? How do their structures depend on their environment and how do they change during function? What are their interactions with functional partners? What is the role of macromolecular dynamics in these processes? - can only be addressed by methods in biophysical chemistry. For example, there is increasing evidence that metabolic pathways involve large multi-enzyme transient complexes, the 'metabolons'; such complexes were named 'quinary' structures, as an extension of quaternary structures that correspond to more stable multicomponent entities. 'Solving' quinary structures will be difficult, because of their weak binding energies, and will require the application at a very high level of biophysical chemistry methodologies combined with clever biochemistry. This report summarises the state of affairs for the field in Europe and suggests areas in need of strong development, if maximum benefits are to be reaped from post-genomic biology.

1. Methods in biophysical chemistry

The methods of biophysical chemistry probe macromolecular interactions, dynamics and low resolution structures in various environments. They fall into the general areas of calorimetry, hydrodynamics (analytical ultracentrifugation, dynamic light scattering), microscopy (electron microscopy, confocal laser microscopy), radiation scattering (small angle X-ray and neutron scattering), spectroscopy (fluorescence, infra-red, circular dichroism, etc.), near field microscopy (atomic force microscopy, tunneling electron microscopy), mass spectrometry, surface plasmon resonance, and so on. In the context of the high-throughput high-resolution structure-solving approaches that will be implemented in X-ray crystallography, methods in biophysical chemistry have important roles to play at three levels:

- In the pre-crystallisation stage - to characterise target molecules;
- In the post-crystallisation stage, when the high resolution structure is solved - to relate it to function and to characterise dynamics and interactions;
- In cases where there is no hope of crystallisation (or of a NMR structure), as for quinary structures, for example – to characterise low-resolution structure, dynamics and interactions.

2. Inventory and needs

We focus on methods requiring sophisticated instrumentation and/or a high level of expertise. These are not distributed evenly across the European continent. In each case, an important question to be resolved is whether to concentrate instrumentation and expertise in central facilities or distribute them more widely. Note that methods associated with synchrotron radiation or neutron beams are naturally associated with large installations, which often spawn a concentration of other facilities around themselves. Examples of this are the development of a spectrum of biophysical methods such as NMR, mass spectroscopy, electron microscopy, ultracentrifugation, and so on, around the ILL-ESRF-EMBL site in Grenoble, or the DESY site in Hamburg.

Analytical ultracentrifugation

Present: Analytical ultracentrifugation (AUC) can be used to probe the oligomeric state and overall conformation of biological macromolecules in solution (ranging from very dilute - <0.1 mg/ml up to the limits of solubility and or gel formation) and their interactions with ligands (small molecules or other macromolecules). A new generation of instrumentation, matched by some innovative developments in on-line data capture and software, make it a very versatile tool for the protein scientist. The drive in instrumentation has come from the USA, whereas European scientists (particularly groups in Spain, Germany and the UK) have been prominent in advances in software (such as the sedimentation velocity analysis programme LAMM and the sedimentation equilibrium programme MSTAR) and methodologies for molecular weight measurement and solution conformation determination.

Future: The great potential of this techniques for structure analysis is in its use in combination with high-resolution X-ray crystallography and NMR, as well as other low resolution solution techniques such as fluorescence depolarisation and X-ray, neutron and light scattering. Advantage can be taken of its ability to study proteins in dilute solution and the availability of clearly laid out protocols (and associated software such as HYDRO, SOLPRO and ELLIPS) for triaxial structure (regularly shaped molecules) and surface shell-bead models (complicated structures like antibodies) largely established by laboratories in Spain and the UK.

The problem of thermodynamic non-ideality currently restricts the application of the ultracentrifuge towards the analysis of interacting systems (stoichiometry and strength). Groups in the FRG, Spain and the UK have already made significant advances (including the use of crystallographic 'Protein Data Bank' (PDB) data to predict co-exclusion effects with the COVOL software, and also molecular crowding treatments of concentrated solutions) and it is expected that extensive advantage will be taken of this.

AUC is a highly specialised technique and it would be prudent to follow the example of the USA in establishing clearly identified centres. The existing one in the UK (the NCMH at Nottingham) effectively operates as a North European Centre. Similar centres could be established in Spain (jointly between Murcia and Madrid) to serve Southern Europe and in Germany to serve central and Eastern Europe. In France, there are single instrument AUC centres in Grenoble associated with the Structural Biology campus (IBS, ILL, ESRF, EMBL), on the Gif-sur-Yvette CNRS campus near Paris, in Toulouse and in Montpellier. As with microcalorimetry (discussed below), it will be important to provide training for new/younger researchers, to facilitate the spread of expertise from such centres into the wider community.

X-ray and neutron scattering

Small angle X-ray and neutron scattering provide low-resolution structural information on macromolecules, their complexes and interactions in solution. These studies used to be stand-alone. New approaches, however, combine small angle scattering with high-resolution data from crystallography or NMR to examine conformational changes in solution under different conditions or to place components inside large complexes or molecular machines. These are based on powerful simulation computer programs developed in Europe.

The theoretical basis and appropriate instrumentation for small angle scattering was developed in Europe, mainly in France, Austria and Germany. X-ray apparatus was available in a number of laboratories and the field saw a period of strong interest up to the late 1970s. With the advent of synchrotron radiation sources, which were initially exploited for fibre and muscle diffraction, most laboratories abandoned their equipment in favour of user-oriented beam-lines at the large-scale facilities. This allowed the provision of state-of-the-art equipment for the scientific community as a whole, but limited expertise to a few locations. Furthermore, it has been shown in a number of instances that the high brilliance achieved on SR beams allows to kinetic experiments on functioning systems. There are centres at the Hamburg synchrotron (EMBL outstation), at the LURE synchrotron in Orsay, at the Daresbury synchrotron, at ELETTRA, Trieste and at the ESRF, Grenoble.

Neutron experiments can only be performed at large scale facilities. For small angle scattering, neutrons are complementary to X-rays, providing for contrast variation methods based on deuterium labelling that allow different components in a complex solution to be distinguished. Because of a number of reasons (such as more favourable contrast, the possibility of using longer wavelengths with negligible absorption, the absence of radiation damage), it is possible to compensate to a large extent for the low neutron beam flux compared with X-rays and to perform experiments under very favourable conditions. The small angle neutron scattering facilities with the highest incident flux in the world are at the ILL reactor in Grenoble. There are small angle scattering cameras suitable for biology also at the LLB reactor in Saclay, the HMI reactor in Berlin, the Risø reactor and the ISIS spallation source in Didcot.

Neutron scattering with energy resolution (spectroscopy) is uniquely suited to the experimental study of thermal dynamics and motions in macromolecules that have been shown to be essential for biological function. Again the premier facilities for such studies are in Europe, at ILL, and there is suitable instrumentation also at LLB, at HMI and at ISIS. A new neutron beam reactor is being built in Garching, near Munich with more instrumentation both for small angle scattering and spectroscopy. It is hoped that the European Spallation Source (ESS), representing the next generation neutron source, will soon proceed beyond the initial planning phase.

It may appear that the current situation is adequate for X-ray and neutron scattering needs in biology, but this is misleading. As testified by the regular international conferences held in these fields, X-ray and neutron small angle scattering and neutron spectroscopy are very useful in many areas of physics and chemistry (especially for the study of synthetic polymers, colloids and materials in general). Biological applications account for only a small part of the experimental time at central facilities.

A proactive approach to X-ray and neutron scattering in biology in Europe should be undertaken at different levels:

- Training: providing for up-to-date text-books in biophysics accessible to students in biology; integrating the university teaching of biophysics and biochemistry; organising high level training schools such as the yearly HERCULES course in Grenoble.
- Encouraging central facilities to adapt to the special needs of biological experiments: for example by reserving instruments to biology, developing in-house biology groups and setting up closer collaboration with outside users (providing fellowships to spend extended periods at the facility to best adapt sample preparation to measuring time). It should be debated whether it would be worthwhile to set up close to high-throughput small angle scattering analysis at a large facility for the more routine characterisations. This would require a radical change in current practice. The organisation of such analytical facilities is well understood and practised in industry, where they run with engineers and highly trained technicians, not with research scientists.
- Promoting a European effort to coordinate expertise in deuterium labelling in order to make it more widely available. Major neutron scattering studies of large complex structures will require advanced labelling techniques that will be specific to each system. Furthermore, deuterium labelling is also useful in other methods of biophysical chemistry, such as mass spectrometry and NMR.
- On the instrumental side, special efforts are required to develop suitable detectors for neutron spallation source experiments (proposals are currently being invited for this in the USA), but also to maintain and develop the special detectors required for small angle X-ray scattering. At present many of these detectors are made at the EMBL, but it is unclear whether or not this organisation will continue development and support in this area. In general, development of new experimental techniques requires a parallel effort in instrumentation. Few academic institutions in Europe have engineering facilities suited to this, whereas at the large facilities, such capabilities are usually fully occupied by maintenance and machine development. If this situation is not corrected to some extent it will certainly hamper the European development of new technology in the coming years.

Microcalorimetry for protein structure-function studies in solution

Introduction and background

Biological microcalorimetry has a long history, with original experiments dating back to A.V. Hill in the 1930s, but until recently has been a rather specialised and technically difficult technique to implement. However, during the 1980s, developments in recombinant DNA techniques, which gave convenient access to larger quantities of pure protein and, more importantly, improvements in instrument sensitivity and baseline stability led to much wider use. Excellent instruments and associated software are now commercially available and are found in many laboratories. Calorimetry is a very low-resolution technique for structural purposes, yielding information generally only on stoichiometry and domain/sub-unit structures for example, but has enormous analytical potential for functional studies.

Two basic types of calorimetry are currently in use: (1) DSC (differential scanning calorimetry) for the study of thermal transitions in solution, and (2) ITC (isothermal titration calorimetry) for measuring interactions in solution brought about by mixing or dilution at fixed temperature.

Calorimetry is, in principle, non-destructive and non-invasive, relying on the heat energy changes that naturally accompany most biomolecular transitions or interactions. Consequently it has widespread application as a straightforward analytical technique, without recourse to extrinsic probes or modifications. It has the further advantage, of course, that the experimental signal relates to absolute thermodynamic quantities (ΔH or ΔCp), and so may lead to information on the fundamental thermodynamic forces underlying the process.

Protein folding, protein-ligand interactions

DSC is typically used to follow thermal unfolding/refolding processes of proteins in solution, with sample requirements of 0.1 mg or less per experiment (concentrations down to 0.1 mg/ml in a DSC sample volume of 0.5 ml). Information is returned on T_m, co-operativity of folding, subunits and intermediates, as well as ΔH and ΔCp values. Interactions with other molecules (large or small) can be detected by effects on T_m, but this is usually better studied by ITC (see below). One drawback to DSC experiments - especially with larger proteins near neutral pH - is the distortion of thermograms by aggregation of unfolded protein and other irreversible effects at higher temperatures. This reflects the inherent stickiness of unfolded or misfolded protein - also a significant problem *in vivo*. Some groups are beginning to use DSC to study the energetics and mechanism of aggregation of misfolded proteins. Kinetic effects can sometimes be detected by scan-rate variation.

ITC is rather more versatile for studying molecular association interactions in solution, and the data are usually easier to interpret. Typical experiments require a sample volume of 1-2 ml (ca. 10 µM concentration for proteins) into which is injected, sequentially, small aliquots (2-10 µl) of the second component (ligand), which usually must be 15-20 times more concentrated than the protein to allow for dilution. Analysis of the subsequent heat pulse train gives, under optimal experimental conditions, the stoichiometry of the reaction (n), the enthalpy (ΔH) and the equilibrium constant or affinity (K). Typical systems studied include: protein-ligand, enzyme-inhibitor, enzyme-substrate, protein-DNA/RNA, protein-protein, antibody-antigen, membrane receptor interactions. Because of the concentrations used (which may match those found within living cells), the upper limit (tight binding) for K is around 10^8 M^{-1} (K_d = 10 nM), though this can be extended to tighter binding in some cases by competitive inhibition techniques. The lower (weak binding) limit is around K_d = 0.1 M, or weaker using inhibition methods. Dimer or oligomer dissociation may be measured in the ITC by dilution experiments. The kinetics of slow processes (from minutes to hours) can sometimes be followed.

Where we stand?

European expertise in the applications of biological microcalorimetry is equal to or better than in the US or elsewhere, with strong groups in the UK, Germany and Spain. Instrument development, however, is mainly restricted to just one (Microcal Inc., Northampton, MA - http://www.microcalorimetry.com/), possibly two (Calorimetry Sciences Corporation, Utah - http://www.calscorp.com/) companies in the US. Both these organisations have benefited from migration to the USA of former Soviet Union personnel from the Privalov group in Poustchino. One EU company (Setaram - http://www.brookhaven.co.uk/thanda.html) has potential for calorimetry instrument development, but currently lags behind in applications related to proteins in solution. Thermometric AB (Järfälla, Sweden - http://www.thermometric.com/) produce sensitive calorimeters best suited for

longer time-scale studies. Ultra-micro, nanotechnology 'calorimeter-on-a-chip' devices are under development in the UK and USA, but are unlikely to be of significant wide-scale use in protein solution studies until problems associated with sample delivery and mixing have been addressed.

Protein identification

Calorimetry, especially in combination with other techniques, may allow the attribution of function to newly identified proteins. One recent example (from the Glasgow group, UK) involved the use of DSC, ITC and fluorescence methods to identify a hitherto unknown function of a protein allergen from a parasitic nematode. Subsequently a protein with close sequence homology was identified in the *C. elegans* genome, thereby allowing a possible function for the *C. elegans* protein to be identified.

Membrane proteins

Detergent-solubilised proteins can be studied by both DSC and ITC, and numerous protein-receptor interactions have now been examined in this way. DSC can be used to monitor thermal transitions in intact membranes, and sometimes the unfolding *in situ* of intrinsic membrane proteins can be followed. Study of intact membranes by ITC requires some care to avoid artefacts due to the high concentrations of membrane material required to give sufficient receptor concentrations. Reconstituted lipid vesicles containing the required species are useful here.

Protein dynamics

Thermodynamic fluctuations in mesoscopic systems such as individual protein molecules are fundamentally linked to the heat capacity (Cp) of the system. Consequently, calorimetric methods supply the basic parameters within which molecular dynamic processes must operate.

Supramolecular complexes

The assembly/disassembly of protein complexes can be followed by ITC titration or dilution methods, to give stoichiometries and thermodynamic parameters for the complex.

Integrative biology

Whole organisms can be studied by calorimetric techniques, though experiments are usually confined to the monitoring of metabolic processes. The methods are rarely, if ever, of sufficient resolution to identify individual molecular processes, though the non-invasive nature of the technique has advantages in some instances. Nanoscale calorimeters on a chip have been built to follow heats of metabolic processes in single cells, but the results so far have not been very encouraging and much more development is needed here before the technique can gain general acceptance. One potential advantage of calorimetry, yet to be properly exploited, is the ability to study complex mixtures at concentrations relevant to those possibly found within living cells, without the need for additional spectroscopic probes nor any requirement for optically clear samples.

Mass spectrometry

No technology has exceeded, and a few other technologies have equalled, mass spectrometry in its rapid growth in sophistication and widening applicability to protein science. Its importance in genomics and proteomics, from protein identification to sequence analysis to post transcriptional modification, is

stressed elsewhere in this report. But it should be noted that mass spectrometry is now beginning to be applied also to the study of proteins in solution. This has been made possible by exciting developments in both hardware and software. The introduction of new instruments, such as the Q-TOF and FTICR spectrometers, has raised the masses of molecular ions that can be manipulated and the accuracy with which they can be measured, to undreamed-of heights.

In combination with hydrogen-deuterium exchange, mass spectrometry can be used not only to detect conformational changes in proteins, but also to pinpoint the regions undergoing such changes. Protein-protein interaction can be studied by the use of chemical cross-linking agents, most elegantly by means of reversible cross-linkers, and the sites of cross-linking can readily be identified. In favourable instances, proteins that form complexes are found to remain complexed in the mass spectrum and this too can lead to the identification of interacting partners. All these are important aspects of the study of protein structure and function in solution; the capacity of mass spectrometry to yield valuable information, not least where other techniques are difficult or indeed impossible to apply, has unexpectedly moved it to the front of the stage in this area too.

Much mass spectrometry is relatively simple to undertake, given in particular the increasing simplification of the instruments by the manufacturers. The price of 'routine' instrumentation places it within the reach of most well-equipped laboratories. The corollary of this, however, is that it is becoming such a routine technology that few laboratories can afford to be without it and this will place its own burdens on the national funding agencies. When it comes to the most sophisticated instrumentation, the cost together with the level of skill required to operate it and to interpret the results, inevitably means that there will have to be fewer machines and that thought should be given to funding these properly and to ensuring reasonable access to other scientists with interesting problems to solve using this important methodology.

Other technologies

In this report on biophysical chemistry, we have limited ourselves to technologies that derive from the application of relatively sophisticated instruments, which in certain instances, most obviously synchrotrons, may best be sited in central locations and used on a shared basis. Our list was not intended to be exhaustive and we have deliberately left out well-known ancillary techniques such as spectrophotometry and fluorescence. This should not be taken to mean that we regard these as of lesser importance; on the contrary, proper protein science in solution cannot be conducted without laboratories well equipped for such purposes. Imagine trying to study an enzyme seriously without access, if needed, to rapid-reaction techniques.

Another methodology now beginning to make an increasing impact is the use of surface plasmon resonance detection to monitor the interaction of one protein (immobilised) with another or with a ligand, the major advantage being that the experiment can be conducted in real time to give kinetic data for both the on and off reactions. The inferred equilibrium data can profitably be compared with those measured by other means, such as microcalorimetry (see above). Atomic force microscopy also deserves a mention since it too can provide data that complement studies of proteins in solution by other means, for example a comparison of unfolding a single protein molecule with the bulk unfolding in solution studied by, say, NMR spectroscopy. Techniques such as these can confidently be expected

to grow in importance and applicability in the foreseeable future and to require adequate funding and support.

Manpower and training

As with all instrument-based techniques, and probably even more so as instruments become more automated and apparently easier to use, it is important to maintain a background of expertise and experience, together with appropriate training facilities, in order to avoid pitfalls in measurement and interpretation. It would be unsafe to rely solely on instrument manufacturers for this, because of potential conflict of interest. It is also unrealistic to assume that every laboratory or group should have this expertise in all areas - even though they may have access to the instruments. National, regional, or European facilities specialising in key technologies will act as a reservoir of talent and provide training for new/younger researchers, as well as providing instrument access to outside users where appropriate. It is important, however, that any such facilities be truly open.

Appendix 3

Group III Report

Macromolecular 3D Structure Analysis in Europe

20 March 2000, Schipol Airport Conference Center, Netherlands

***Participants*:**

Keith Wilson (University of York, UK)
(Chairman)
Christian Cambillau (CNRS – Université de Marseille, FR)
Jose L. Carrascosa (Universidad Autonoma Madrid, ES)
Maria Armenia Carrondo (Instituto Tecnologia Quimica e Biologica, PT)
Rob Kaptein (Utrecht University, NL)
Marianne Minkowski (ESF- Strasbourg, FR)
(Guiseppe Zaccai, CNRS–CEA, IBS, FR, *Apologies for absence*)

Introduction

The remit of the thematic group III was to assess the needs for support for 3D analysis of macromolecular structures on a European as opposed to a national level. This was particularly directed to the needs for improvement in methodology including both hardware and software, and the needs for central facilities and access to those facilities on a European scale. The discussions were limited to the three major areas currently in use, X-ray crystallography, Nuclear Magnetic Resonance (NMR) and Electron Microscopy (EM), plus a more limited area of application, diffraction with neutrons. Bioinformatics was covered in a separate thematic group. The group recognised that major limitation for all three techniques lies in the overexpression and purification of proteins in large amounts for structural research, and whilst this was being considered by another thematic group, these problems will also be emphasised here.

The group aimed to identify bottlenecks in the current analysis of structures by the three techniques and to suggest how these might be alleviated or overcome by European initiatives.

1. Structural studies in the post-genomic era

Structural studies in the post genomics age can be subdivided in two main streams that share the same techniques, but with different aims:

- ***Structural genomics.*** We took this to be the systematic study of potentially all the proteins produced by a genome with the aim of identifying new folds and new functionalities. This approach is currently applicable only to those proteins that are easy to over-express and purify in large amounts, and is not likely to be successful for complex systems such as multi-domain flexible structures, glycosylated eukaryotic proteins or membrane proteins.
- ***Target oriented projects.*** This is a continuation of the work which most academic groups have been concentrating on to date, that is the identification of interesting biological targets for study, followed by a detailed structure-function analysis of the system. This whole area is also transformed by the availability of complete genomes, as this provides a wealth of related target structures in any particular functional area.

In the meeting at Erice (Sicily) in May-June 2000 there was a general discussion on the impact of Structural Genomics in Macromolecular crystallography (MX). Several speakers summarised some key differences between true structural genomics projects and the present target-oriented approaches as shown in the following Table:

Some key differences between true structural genomics and present target-oriented approaches

	Structural genomics	**Biologically driven**
First aim	Observation of nature	Understanding of nature
Questions addressed	Easy ones at first	Difficult, involves function
Academic interest	Medium	High
Structure and function	Separated	Combined
Education	Efficient but lower teaching value	Educational
Organisation needed	As big science	Many independent labs, institutes or consortia

Experimental methods of structure analysis remain central to molecular biology: the current *ab initio* prediction methods for macromolecular structures are simply inadequate for the level of accuracy required for such activities as drug design.

For a true structural genomics project, high-throughput is essential, but it is clear that the target-oriented projects will also be greatly aided by high-throughput techniques.

For all three techniques (MX, NMR, EM) it is likely that an increasing number of biologists will wish to use the techniques as non-experts. This requires that the infrastructure in terms of software and hard-ware be made as simple as possible to use, with user-friendly interfaces and guidance for the inexperienced. However, it is vital that a core of researchers be supported to develop and maintain the methods at the heart of all this techniques, which continue to pose significant intellectual challenges. **Training and recruitment of young researchers to these fields is essential if Europe is to retain a powerful position.**

2. The major techniques

We now briefly discuss each of the three major techniques (MX, NMR, EM) in turn, with a short discussion on neutron scattering and the free electron laser.

X-Ray crystallography

Overview

Of the three techniques, X-Ray crystallography is the one which is most easy to gear up to high-throughput using automated crystallisation methods, increasingly automated data collection in-house and at synchrotron radiation sites and better algorithms for structure solution, refinement and deposition. X-Ray crystallography continues to play the major role in the determination of 3D structures of macromolecules and is expected to do so for the foreseeable future.

Concerning technology, overexpression of purified and stable protein remains a bottleneck and, as discussed by another thematic group, must be an area of central importance. We discussed in particular, the possibility of reinforcing efforts about membrane protein crystallisation. Membrane studies are an area where Europe has a strong tradition, and since roughly 30% of proteins appear to be membrane-associated, this should be targeted in the future.

In addition, we could mention detectors, since while MAR in Hamburg is a leading player, we face stiff competition from the USA. Europe has very good facilities for MX at its various synchrotron sites, especially the ESRF, but we note that the number of beam-lines for MX, current and planned, remains well below that in the USA and Japan. Automation of these lines in terms of, for example, sample changing, is also vital. Europe has a strong tradition in crystallographic methodology and software and this must be supported in the future.

Over-expression

Macromolecular crystallography (MX) generally requires large quantities of soluble, homogeneous material. Furthermore, the most recent techniques require the introduction of labelled methionines.

Therefore, high-throughput protein structure determination required for structural and functional genomics, is feasible using only recombinant proteins in *E. coli*. Techniques for eukaryotic proteins using baculovirus or related expression systems are being developed apace, but the *E. coli* systems can be expected to dominate in the short term. Scaling-up involves automation of sub-cloning, solubility checking and small-scale expression. The automation of each of these processes requires independent studies for the different types of protein (cytoplasmic, glycosylated, membrane etc.) and genome. While they will form part of any structural genomics project, they will be of great value to all structure function research.

These steps are common to MX, NMR and EM.

- ***Crystallisation:*** Crystallisation of *soluble* proteins is largely based on standardised procedures available in molecular biology kits. Robots are available but are still expensive and require a good deal of protein. For a structural genomics project, micro-techniques should be developed and used (based on handling 0.05-0.1 micro-litre drops) as well as automated techniques for checking the evolution of the crystallisation drops. However, membrane, glycosylated or insoluble proteins will be difficult and require further study.
- ***Data collection:*** This will be largely based on the use of third generation synchrotrons. Automated procedures should be developed to install crystal-handling carousels and to provide remote control of sample exposure and data collection control.
- ***Phasing:*** to date, phasing is mainly based on Multiple Wavelength Anomalous Diffraction (MAD) techniques. Single Wavelength Anomalous Diffraction (SAD) should be developed in order to reach high-throughput phasing. *Ab initio* methods should be further developed and implemented, especially with the increasing availability of atomic resolution data from third generation synchrotrons. Computer tools are available which allow automatic phasing and phase improvement, but these need to be streamlined to facilitate communication between the different packages.
- ***Model building and refinement:*** Computer tools exist which have made it possible to build automatically a large part of a well-phased protein model (70-90% of the model) for high-resolution structures. Refinement techniques have improved greatly. However further development is needed to increase the level of automation and to handle low-resolution structures better.

The steps of X-ray crystallography, common to both structural genomics and target-oriented projects are listed in the summary below (* indicate the level of difficulties, which may be anticipated in the scale-up necessary for structural genomics):

Summary

Task	Automation	Difficulty
Over Expression	Automation of subcloning	**
	Automation of small scale screening of expression and solubility	***
	Introduction of seleno-methionines	*
Crystallisation	Automation: robots available	*
	Systems not presently amenable for structural genomics:	
	• Big ensembles	***
	• Floppy domains	***
	• Membrane proteins	***
	• Glycosylated proteins	***
	• Insoluble proteins	***
Data Collection	Automation of crystal handling	**
	Multiple (or Single) Anomalous Scattering (MAD or SAD)	*
	Remote access to synchrotron data collection	**
Phasing and Phase Improvement	Automation with M/SAD techniques and *ab initio* methods	**
Model Building and Refinement	Automated or semi-automated	***
Model Checking	Automated	*

Actions needed

a) Structural genomics

Apart from a few comparable efforts to date in Europe such as the Protein Structure Factory in Berlin, the Swedish programme, the French Genopoles (Lyon, Marseille, Paris-Evry and Strasbourg), the Wellcome Trust and the BBSRC initiatives, we are already far behind the USA.

Europe urgently needs to set up infrastructures for structural genomics. This requires central facilities for data collection (synchrotrons), bioinformatics and possibly for protein expression. Further discussion is needed about the best distribution of the latter into smaller centres or individual laboratories as opposed to centralisation. Examples of genomes to study might include: hyperthermophiles, pathogenic bacteria, yeast, individual genes of higher eukaryotes (*C. elegans*, a plant, human, etc.). This range of targets includes basic science, and the pharmaceutical and agricultural industries. The size of such projects may well imply that a number of well-funded consortia (approximately ten laboratories in each) be funded. The management of such consortia will need considerable planning.

b) Key biological problems

Many key biological problems are not currently addressable by structural genomics: membrane proteins, large assemblies of proteins, proteins-DNA/RNA complexes. Here again the limiting steps are over-expression and crystallisation. They will benefit greatly from expertise gained in the structural genomics applications.

c) Technology to be strengthened

Technology advances and improved infrastructures are needed in structural biology (MX, NMR and EM) as there are several bottlenecks. In bioinformatics, the European Bioinformatics Institute (EBI) should be supported as a central facility fundamental for structural biology in Europe. For high-throughput approaches, protein expression needs to be automated, as does crystallisation in MX. In MX, beam lines and software for structure analysis need to be enhanced. In NMR, improved labelling techniques and cryogenic probes need to be implemented, as well as more high field machines.

The technologies need advances in two senses. An example is in high-throughput screening crystallisation. This is firstly a problem of technology and engineering: techniques for rapid screening for large numbers of established conditions require use of robotics and image recognition techniques. Secondly however, there are scientific problems to be solved: the subsets of proteins currently resistant to crystallisation such as membrane proteins, need truly scientific research programmes. The technology developments need highly skilled scientists to be recruited to and retained in the field, not only those carrying out applications of currently topical biological programmes.

d) Education

It must be recognised that the development and use of automated methods and equipment is detrimental to training in the fundamentals of this field. It is important that funding is available for people and especially students in the area of structural biology. It is vital that Europe retains its competitiveness: most of the techniques of structural biology originated in Europe. We have to retain a good European-wide education and advanced training system in this field, and actions should be taken to retain skilled people in Europe and prevent a drain to the USA.

Nuclear Magnetic Resonance (NMR) spectroscopy

High-resolution Nuclear Magnetic Resonance (NMR) spectroscopy is able to provide 3D structures of biological macromolecules in solution at atomic detail. In this aspect the technique is therefore complementary to X-ray crystallography. Currently, about 20 to 25 % of the new structural folds of proteins are determined by NMR. These are structures of small (< 40 kDa) often non-crystallisable proteins.

The present size limit of approximately 40 kDa is due to the fact that both line-width and complexity increase with molecular weight. However, recently developed methods such as TROSY and the use of residual dipolar couplings observed for partially aligned biomolecules are likely to push this limit, possibly up to 100 kDa. Also, NMR provides information on the dynamic behaviour of biomolecules and their complexes. Via the measurement of ^{15}N relaxation parameters such as T1, T2 and H-^{15}N NOE, the dynamics of a complete protein backbone can be probed in the picosecond to nanosecond time range. Furthermore, conformational exchange processes in the milliseconde to microsecond range can be determined from their contributions to T2 or T1r information on protein dynamics is often crucial in understanding the catalytic activity of enzymes and the ability of proteins to bind to other proteins or nucleic acids.

Advances have also been made in developing novel NMR technologies for drug discovery. With new automatic flow NMR techniques or automatic sample change techniques the analysis of chemical libraries for suitable ligands to proteins has been considerably facilitated. The so-called 'structure-activity relationship' (SAR) and other selective detection procedures (transfer NOE) are applied in many laboratories of the pharmaceutical industry. In connection with the determination of the target protein structure it will be possible to explore the binding sites of the ligands and to improve the chemical structure of the ligand for a more efficient binding, using computer-modelling procedures. In this context NMR may be used as a high throughput for drug targeting and drug design.

Another area where NMR is making important contributions is that of protein folding and misfolding. This is a central issue in the heterologous expression of proteins for pharmaceutical and medical purposes. And it is increasingly recognised that folding is coupled intimately with localisation and regulation of biological activity. Misfolding events therefore lead to malfunctioning of living systems and an increasing range of diseases from cystic fibrosis to Creuzfeld-Jacob disease, and Alzheimer's is now associated with such problems. NMR has the unique ability to characterise the structure and dynamics of non-native states of proteins and thereby to study folding intermediates.

Structure determination by NMR: limiting factors

The determination of protein structures by NMR involves the following steps.

a) Protein expression

A limiting factor (common to X-ray crystallography) is the over-expression of soluble proteins. High-yield expression is particularly important in view of the need of isotope labelling (^{15}N and ^{13}C). Good solubility at 0.1-1 mM concentration is essential since aggregation is detrimental to the spectral resolution. Improvements can be envisaged in the greater use of automation in screening for expression conditions.

b) Data collection

NMR data collection for resonance assignment and structure determination currently takes several weeks of instrument time. There are several ways to reduce this time.

Firstly, higher magnetic field spectrometers provide both increased sensitivity and resolution. This year a new generation of NMR spectrometers will become available operating at 900 MHz. Another boost in sensitivity comes from the so-called cryo-probes, now available at 500 and 600 MHz. By operating at a low temperature for the receiver coil (~20° K) a major increase in sensitivity of a factor 3-4 is obtained.

Secondly, the development of novel protocols that are less dependent on exhaustive NOE analyses and involve more rapidly obtainable parameters such as chemical shifts and residual dipolar couplings provides yet another way to shorten the structure determination process.

c) Structure calculation

Several good computer programs exist for the generation of 3D structures based on NMR data. Newly developed programs will be intimately linked to the new protocols mentioned above.

d) Analysis and validation

For the quality assessment of NMR, derived protein structures programs are available such as PROCHECK and WHATCHECK. These programmes address the issue of co-ordinate validation. However, validation with respect to how well structures fit to the experimentally derived NMR parameters is less well developed. Some attempts have been made (for example, the AQUA program), but more systematic validation suites are badly needed, also in view of the 'new' structural parameters mentioned above. In particular, the problem of data uniformity is acute, since thus far the consistency in nomenclature of NMR data and coordinates in the Protein Data Bank (PDB) has been poor.

Solid state NMR

Recently solid state NMR has started to impact on structural biology. The nature of the information is twofold. On the one hand precise but limited structural information is obtained, for instance, on the conformation of retinal in bacterio-rhodopsin or on cofactors or inhibitors complexed to proteins. On the other hand progress has been made towards full structure determination of membrane associated proteins albeit thus far limited to small single or double helical units. Solid state NMR holds particular promise for the study of membrane proteins. This is especially important since these proteins are difficult to study by X-ray or electron diffraction methods. In Europe a consortium of research groups has obtained EU funding for a wide-bore 750 MHz NMR spectrometer, which is currently the top for solid state NMR. The instrument is located at Leiden University in the Netherlands.

European NMR facilities

The new generation NMR instruments necessary for biomolecular NMR, 900 MHz for solution-state and 750 MHz wide-bore for solid-state NMR, are extremely costly (about 5 million euros). In addition, the large stray magnetic field and shear size of the magnets usually require separate building facilities. This makes it necessary to concentrate high-field NMR instrumentation in a few centres in Europe. These centres would typically have a range of NMR spectrometers available plus the infrastructure for computing and protein expression. High-level expertise in biomolecular NMR should also

be maintained. The European Union has played an important role in establishing this type of centres through the Large-Scale Facilities programme. Thus, a cluster of three NMR Large Scale Facilities exist in Frankfurt (Germany), Florence (Italy) and Utrecht (The Netherlands). Related facilities are located in Wageningen (The Netherlands), focused on plant biochemistry, and in Jena (Germany), where a combination of NMR, X-ray and EM is supported. The EU programme provides access to the facilities for research groups from all EU member states, while investments in instrumentation and building costs are born by the local governments and funding agencies. The NMR Large-Scale Facilities typically serve 40 to 50 groups in Europe and are the focal point for NMR expertise and technical development. The three facilities will all have the most modern equipment available (900 MHz in 2001). This model of a cluster of de-localised but inter-linked facilities has worked very well for Europe. EU support has provided access to expensive equipment for groups that otherwise would not have had this possibility. On the other hand the EU funding also worked as a leverage to secure the large investments from local governments. However, there is some concern whether this model will work in the future considering the staggering rise in costs that can be expected for the next generation equipment. Financial support not only to provide access but also for infrastructure at the NMR centres by the EU or other supranational European bodies should be seriously considered.

Electron Microscopy (EM)

Electron Microscopy (EM) has a wide range of applications due to its unique capability to provide structural information from the atomic level (electron crystallography) up to the cellular level (electron tomography). These possibilities have induced the USA and Japan to heavily invest in infrastructures related to new electron microscopes: NIH, Harvard and several Japanese initiatives are investing hundreds of million of dollars in imaging facilities based on new Field Emission Guns at intermediate (200-300 KV) and high voltages (1MV) equipped with cryogenic-facilities. The use of cryogenic-EM in structural biology can be considered at three levels: electron crystallography, single particle analysis and electron tomography.

Electron tomography

Use of the EM to collect data of thick sections of cells or tissues (or even thin cells) and to obtain a 3D reconstruction of the supramolecular organisation of cells. It demands a high degree of automation and data analysis packages. Its potential can be considered at the cell biology level, due to the inherent low resolution (5 nm at best). Nevertheless, it has unique possibilities to study the overall morphology *in situ,* and bridges the medium resolution with optical microscopy.

Electron crystallography

It is based in the analysis of (thin) 2D crystals of proteins. These crystals can be either natural or artificially induced (2D crystallisation techniques are available, and are the subject of intensive development). It is presently the best-suited technique to solve structures of membrane proteins crystallised in their natural environment (that is, the lipid bilayer). It has also been applied to soluble proteins.

- *Limitations*: around 4Å resolution.
- *Requirements*: Demands over-expression systems at a moderate level and crystallisation of the complexes under study. The possibility to use either detergents or lipid layers in the crystallisation process allows the examination of the membrane proteins in a near-physiological environment.

Single particle techniques

These are, in principle, universally applicable to objects in a range of 100 kDa to several millions kDa. It offers the widest range of application of 3D cryogenic-EM, and it is the only EM-analysis system that is susceptible of medium-throughput scaling.

- *Limitations*: Presently the routine resolution limit is around 15Å, best cases around 9Å. Possibility of optimisation up to 5Å.
- *Requirements*: Demands moderate over-expression and purification of the objects to study, but it does not require crystallisation.

Requirements for developments and exploitation of the possibilities of 3D EM techniques

A European working group nucleated in a Network during the EU Fourth Framework Programme has been discussing the possibilities for the development and wide implementation of the possibilities of modern EM. The main ideas are the following:

1) There is a need of automating the data acquisition to allow extensive analysis of samples. This will require development of new methods and expert systems for specimen selection, area analysis and data selection and acquisition.

2) Development of new programmes for single particle reconstruction at high resolution. Treatment of physical properties of the optic system (CTF), and problems derived from the massive image input. There is a need for new algorithms for tomography (single particle and whole-cell) and a merging of the different packages already available for electron crystallography.

The best way to attain these objectives would be the generation of an European Group of EM centres, that should include three or four of the leading laboratories already equipped (that is, W. Baumeister, R. Henderson, W. Kühlbrandt and A. Brisson), plus five others requiring investment in top-of-the-line microscopes (200-300 FEG, cryo-EM, energy filter, computer control: 1-2 million euros per group), and selected because of their experience, particular expertise in complementing activities and a regional distribution covering most areas in Europe. This network would co-ordinate the different developments, boosting the interchange of methods, programmes and personnel. Furthermore, these centres would act at a regional level by activating the use of these techniques, as well as offering the instrumentation and methods developed along these lines. In a way, it would be a 'distributed' Large-Scale Facility that would be based on state-of-the-art international facilities in European EM.

The main Areas of Application:

- Membrane proteins.
- Macromolecular aggregates with large dimensions and (or) complex morphological features that prevent their crystallisation.
- Structural transitions induced by ligands.
- Complementarity with X-ray diffraction (and NMR): use of atomic data to locate components and their configuration in medium resolution EM structures.
- Use of EM data to phase X-ray diffraction data sets.

Limitations:

- Poor resolution.
- Limited output per group: five structures per year.

Neutron scattering

Due to the expense of neutron sources, their use is likely to remain small in comparison to PX, EM and NMR. Nevertheless, neutron scattering (NS) has applications in structural biology that range from atomic resolution crystallography to scattering studies in solution to study large macromolecular complexes. NS is complementary to MC and NMR at high resolution and EM at low resolution, providing information that cannot be obtained as readily. Its main disadvantage is the scarcity of appropriate sources and instrumentation. Europe, in this context, has the premier neutron scattering centre in the world, around the high flux reactor of the Institut Laue Langevin in Grenoble. The best neutron spallation source centre is ISIS at the Rutherford Laboratory, near Oxford. A new neutron beam reactor is being built in Munich. A decision concerning the ESS (European Spallation Source) project is expected within a year. The ESS is planned with two target stations to provide new generation pulsed and steady state beams, respectively. Outside Europe, a new spallation source is currently being built at Oak Ridge, in the USA, and there is a spallation source project in Japan. Biological applications are essentially neutron flux limited and will greatly profit from these developments.

Neutron crystallography

The definition of the ligand-macromolecule interface, which requires the positioning of functionally important H atoms in the macromolecule and bound water, represents an important challenge in structural biology. Neutron crystallography provides information on these H atoms in crystal structures. Such information can be obtained from X-ray crystallography only if the H atoms are ordered to atomic resolution (better than 1Å), whereas, by using H-^{2}H exchange, information on hydrogens not ordered to very high resolution can be obtained from neutron crystallography. A disadvantage of the technique is that relatively large crystals are required, but new multi-wavelength methods are being developed to increase data collection efficiency. The new spallation sources are expected to provide very important gains in this field.

Intermediate resolution studies

They provide information on partially ordered systems such as membranes and fibres and on the organisation of disordered complex components within crystal structures. H-^{2}H labelling and contrast variation allow to interpret the low-resolution data and to obtain a highly reliable structural description. Examples of applications are the positioning of different lipid types in membranes, of hydration in membranes and DNA fibres, of the detergent distribution in membrane protein crystals and of the organisation of nucleic acid inside viruses.

Large complexes at very low resolution

Progress in macromolecular crystallography has stimulated an interest in the next level of organisation: structures in their cellular context. Often this involves large complexes or macromolecular machines. NS, again because of the power of H-^{2}H labelling and contrast variation, is a proven

method to obtain reliable low resolution information on the internal structure of such complexes, e.g. on the relative location of various protein or nucleic acid components.

Free Electron Lasers (FELs)

FELs were not extensively reviewed as none exist at present that operate in the relevant wavelength range. These devices are likely to develop as a direct result of the next generation of linear particle colliders, the limit of circular systems having been reached for practical purposes. FELs are planned at several sites world wide, the most relevant site being DESY in Hamburg. They are proposed to produce radiation in the wavelength range of interest to structural biology, in an extremely small focal spot with intensities around 10^{20} photons per mm^2 per second, but in very short pulses of length of the order of femtoseconds. In addition, the incident beam will be coherent, giving the possibility of holographic experiments with no phase problem.

Many potential difficulties need to be addressed if FELs are to be applied to problems in structural biology. These include sample preparation (most samples are likely to be single particles or small arrays of particles avoiding the need for crystallisation), detectors (the present 2D detectors are clearly insufficient for pulses of this intensity and duration) and sample stability. The latter is likely to pose the most challenging problem: the question is whether the diffraction data can be successfully emitted (and recorded) before the absorbed energy causes complete destruction of the sample.

This really is a field for the future in which biologists have no experience. However we have as a community always made rapid and effective use of new radiation sources of ever increasing intensity, and it is vital that the potential of FELs be seriously investigated during the next years, so we are best placed to do so yet again. A FEL in the relevant wavelength range is planned in Hamburg, and structural biology groups should contribute to the scientific activities at this site.

3. Summary

It is clear that many of the actions defined in the draft are truly transnational projects in essence. These should be the real targets for a European financing programme, better that replicating national objectives. At the least, truly structural/functional genomics projects and maintaining the basic infrastructure for structural research to assure European competitiveness could be good examples to justify the need of a change in the European Commission financing policy. It is particularly urgent to allow the acquisition of equipment with EC funds to ensure an even geographical distribution of resources in Europe. Another alternative could be a combined programme to mobilise national resources based on transnational projects that would have been selected and (partially) financed by the EU (for example distributed centres for NMR and EM). This is another key role that European agencies should play in the structural biology/genomics field.

The concept of infrastructures has evolved significantly in recent years and now concerns facilities at an international level and centres of excellence at a regional level. These are evaluated by their inter- and transdisciplinary scientific activities, their technological competence, their relationships with companies and also their capability for training young researchers at basic/applied and science/industry interfaces. The main requirement of any such graduate training programme is to produce

trainees that have an excellent basic core competence in one of the key scientific technological areas of relevance while, at the same time, being able to reach out across sciences and technologies which interface with their own, as such 'lateral' thinking is the key to innovation. The competencies of today can then become the new technologies of tomorrow.

Centres of excellence with these characteristics play a critical role in the transfer of technologies from academia to industry (and vice versa) and thus networks of regional centres of excellence and training programmes with special emphasis for training at interfaces should be promoted and encouraged at a European level.

4. Actions needed

Some developments can and should continue to be carried out at a national level. Others are sufficiently challenging that they clearly require co-operation at the European level if we are to make the best use of resources and compete with the USA and Japan.

Central facilities and other experimental centres

MX will remain the prime technique for fast throughput 3D structure determination and synchrotron beam lines will continue to play an essential role. The accessibility of these to scientists from all over Europe is a priority. This requires:

- a sufficient number of beam lines most with facilities for MAD data collection;
- the further development of high speed detectors, such as pixel based devices;
- automated sample handling of the crystals on the beam line, that is, robotics for crystal changing;
- automated and user friendly software at synchrotron sites, ideally making it transparent to the use which beam line they are using.

Due to the recent increase in the number of MX beam lines at ESRF in particular, and the installation of commercial CCD detectors on most beam-lines we are for the moment in a relatively good position with regard to over-subscription of beam-lines. This is likely to take a downturn over the next few years as structural or functional genomics projects vastly increase the number of structures to be determined.

NMR will require more high-field instruments. The present small number of European centres will need to be increased.

More **EM** centres will need to be established, spread more widely over Europe. Although the throughput of EM is likely to remain low in terms of numbers of structures, its ability to look at large complexes and complement the other techniques makes this essential.

The development of the FEL must be closely followed, its potential for structural biology thoroughly investigated, and if necessary biological groups established at the appropriate centres.

Methodology development

All the structural techniques rely heavily on general techniques for high-throughput of large amounts of protein: this is addressed by the other thematic groups. They all in addition rely on computer software.

Needs here include:

- integration of software across Europe to create efficient and user-friendly packages for academic and commercial scientists;
- development of new algorithms adapted to high-throughput;
- integration of the results of NMR, EM and MX into a high-quality 3D database. This needs collaboration with the EBI and the PDB;
- for MX, collaboration on the problems of crystallogenesis, including the creation of a database of the approaches and their success;
- Macromolecular 3D Structure Analysis in Europefor NMR, improved labelling techniques.

Exchanges and training

- European wide networks and discussion groups especially on the methods involved in structural research, not just on the applications;
- training workshops on the complex methods;
- exchange fellowships, ranging from a few weeks to pre-doctoral and post-doctoral fellowships to allow young scientists to gain hands-on experience in state-of-the-art centres.

All these areas are complex and challenging. Their use should be made as straightforward as possible, but a core set of scientists is needed who really understand the basics of the techniques and can develop them in the future. It is essential that the area is made sufficiently attractive to encourage young scientists to enter it and see a long-term future in it.

Coordination

In all the techniques, a European level of co-ordination would be beneficial.

- In MX, this would ensure there is a complementary set of beam-lines with the appropriate properties such as wavelength, intensity and detector.
- In NMR and EM, it would provide centres with sufficient resources to solve a range of problems, or perhaps concentrate on one sub-set.
- The complementarity between our battery of approaches must be emphasised. We must fight together for the importance of structural biology, and not become divided by technique.

Appendix 4

Group IV Report

Proteomics

19 April 2000, Sofitel, Brussels Airport, Belgium

***Participants*:**

Joël Vandekerckhove (Ghent University, BE)
(Chairman)
Francesc X. Aviles (Univ. Autonoma Barcelona, ES)
Denis F. Hochstrasser (Hôpitaux Universitaires, Genève, CH)
Marianne Minkowski (ESF-Strasbourg, FR)
Peter Roepstorff (Odense University, DK)
Matthias Wilm (EMBL, Heidelberg, DE)

Introduction

The group met to discuss the relation of proteome research in the context of protein structure-function research, to define their viewpoints and to answer questions raised by the multidisciplinary working group on Protein Structure and Function. The latter met on December 7, 1999 in London.

1. Definitions

In order to direct the discussions the panel first formulated a definition of 'proteomics'.

Proteomics is a quantitative and qualitative analysis of the global protein pattern of a tissue, a cell, a cell-organelle or a supra-molecular protein organisation present or synthesised at a given moment.

***Global*:** methods used, should measure as many proteins as possible.

***Quantitative*:** should reflect the ratio of proteins present in the original sample (cell, organelle, etc.)

***Qualitative*:** methods used, should be able to pick up protein modifications, single amino-acid exchanges, etc.

***Present or synthesised*:** the protein pattern of a given sample at a given moment may be different from the pattern being synthesised (expressed) at that moment. The former reflects the history of an expression pattern, the latter provides a snapshot of gene activity. The actual protein composition is generally visualised by protein staining while the protein synthesis is measured by ^{35}S-incorporation.

It is realised that none of the technologies that are available today can fulfil all these requirements (non-ideal proteomics). This strengthens the need to search for novel methodologies, which will approach the definition of proteomics in a better way.

There are sub-definitions in proteomics:

***Expression-based proteomics*:** measuring protein patterns in a quantitative and qualitative manner.

***Functional proteomics*:**

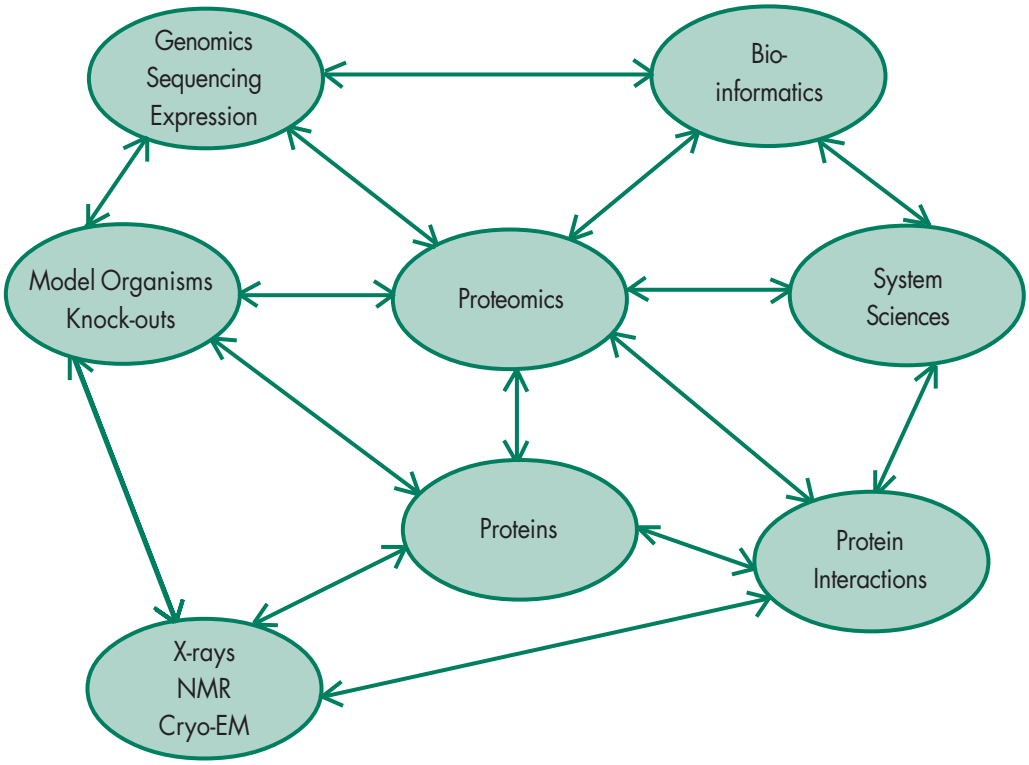
- identifying components of interconnecting pathways;
- identifying partners in protein organisations;
- Post-translational modification (PTM) linked protein-interactions.

***Structural proteomics*:** Assigning three-dimensional structures to given protein sequences; either by computational structure prediction or by Nuclear Magnetic Resonance (NMR), X-ray crystallography or cryo-Electron Microscopy (cryo-EM). This is considered by the group as another level, integrative with structural genomics and is therefore not taken up in the definition.

2. Proteomics in protein science

Based on the interrelation scheme provided by the multidisciplinary working group on structure-function, the proteomics thematic group agreed on a scheme shown below:

Interrelation scheme showing proteomics connected with different protein sciences



Proteomics plays a vital and unreplaceable role in genomics. For instance, in verifying predicted proteins from DNA sequences, in predicting the splicing boundaries, and in identifying tissue-specific alternative splicing. In addition it is the only avenue by which to identify post-translational modifications.

A global analysis of expressed proteins provides clues to the dynamics and activities of the genes in different tissues, under different stimuli.

The massive amount of data generated in the course of proteome studies, creates the need for efficient exploration of these data.

While bioinformatics has mainly been employed to create and to search sequence databases (the fast A package), the thematic group put forwards a novel important approach called 'system sciences'.

'System sciences' is a science dealing with a multitude of information, analysing this information in real time (pattern recognition) or with time (able to stimulate, reconstruct and remodel complex biological

systems). 'System sciences' is dealing with highly complex biological systems by 'intelligent' analysis, deducing rules from this and predicting the behaviour of these complex systems.

'System sciences' is a highly multidisciplinary science at the intersection of mathematics, engineering and biology. It is one of the most important sciences at the edge of proteomics and probably also of functional genomics.

While the other connections between proteomics and other related sciences are self-evident, the relation with structural sciences (X-ray, NMR and cryo-EM) is considered only of a secondary nature.

In conclusion, proteomics is a pivotal entity in the global context of protein sciences.

Proteomics is not only important in fundamental research, where it primarily contributes in understanding the enormous complexity of gene activity in different cells, but it has also acquired a crucial position in applied sciences.

Several illustrations are cited here:

- in Medicine to identify disease-related protein-patterns, used as diagnostic tool;
- in Pharmacology to analyse effects of drugs on cells in a global manner;
- in Microbiology to study pathogen-host interactions;
- to study potential alterations in gene expression in GMO's;
- as a quality control.

3. Signals to be transmitted to the 'decision makers' in society

a) The actual money-stream from political instances in Europe has been limited so far. However the thematic group welcomes recent initiatives supporting proteome research in several European countries.

Concerning private initiatives, most of the financial support does not seem to be anchored to European capital.

In contrast to the fact that European scientists were very often leaders of, or in at the development of, proteome-related technologies (for example, MALDI, nanospray, fragmentation nomenclature, protein databases, electroblotting, peptide mass fingerprinting, sequence tags, etc.), commercialisation of their products was rarely undertaken in a European context.

A typical illustration of the situation was summarised in the following statement: "*Europeans develop the machines and technologies, US venture capitalist commercialise and take the patents; Europeans buy back their own technology in a user-friendly tool box.*"

b) Access to a facility where protein identification is done in a fast and routine manner, appears very important in several areas of biological sciences.

For instance: to verify the nature of the expressed product, to check for post-translational modifications, as guidance during protein-purification etc. The thematic group therefore strongly supports the concept of having automated (MALDI) machines at the disposal of nearly every laboratory.

c) Private companies or organisations should provide more and easier access to their databases at the profit of non-private research. This could be encouraged by financial support from the European Union.

4. Recommendations to the scientific community

a) The thematic group members feel there is general lack of acquaintance with the possibilities of proteome research, although they agree that the term proteomics is generally spread.
They recommend the organisation of courses and better publicising of courses, the organisation of visits, etc.

b) Proteome research will eventually generate data and results that have to be integrated with data from other areas. This horizontal linkage (also called interfacing) can be done only in a unified framework with other areas. Therefore, the thematic group is aware of the urgent need to design a unified framework to manage the data flow.

While the group considers this as an important task, it advises the creation of a "proteomic Esperanto", containing a minimal set of data, fully defining a protein or a gene product, for example:

- chromosome/gene linkage;
- predicted pI, mass;
- calculated pI, mass;
- measured pI, mass.

In order to make the information quickly available, the thematic group recommends that submission to databases should be in a standardised format concomitant with publication.

c) Finally, the group stresses the need for the development of novel technologies, realising that current techniques show severe limitations. While discussing different emerging techniques, the group members summarise the future objectives as follows:

- Quantification is of high importance. They realise that the 2D-gel approach may miss proteins and that mass spectrometry is far from being quantitative. They welcome strategies using differential isotope labelling.
- They stress the need for highly automated high-throughput approaches.
- They point out the need for alternative protein or peptide purification methods.
- They realise that understanding post-translational modifications will form an important item, which can best be approached by mass spectrometry.

Appendix 5

Group V Report

Bioinformatics in a European Context

***Participants*:**

Gunnar von Heijne (Stockholm University, SE)
(Chairman)
Søren Brunak (Technical University of Denmark, DK)
Graham Cameron (EBI – EMBL, UK)
Anna Tramontano (IRBM "P. Angeletti", IT)
Gert Vriend (CMBI, NL)

Summary

There is a pressing shortage of trained bioinformaticists in Europe. This shortage is felt in industry and academia alike and is crippling for a successful dissemination of the results from genome projects. Unless Europe is able to very quickly vitalise its bioinformatics community, all fruits of large genome sequencing and functional genomics/proteomics efforts will go to the USA.

This report summarises the national and international bioinformatics efforts in Europe and makes suggestions for the granting of existing bioinformatics groups and for future procedures that should lead to a rapid increase in trained bioinformaticists in Europe.

The field of bioinformatics is undergoing a very rapid development at present, and it should be noted that this Annex reflects the situation as of the early fall 2000.

1. What is bioinformatics?

Bioinformatics is often called *in silico* biology - its main defining characteristic is the use of computational techniques to handle, analyse, and add value to the flood of data coming out of modern genomics and proteomics. Theoretical analysis of macromolecular sequences and structures, analysis and comparison of genome data, modelling of protein structures, so-called 'rational drug design', prediction of mutations that will give macromolecules new characteristics, protein design, the search for regulatory or other elements in DNA, data-mining in biological databases, and handling of macromolecular sequence and structure data and databases are the classical core of bioinformatics activities. In addition, molecular dynamics, metabolic pathway analysis, modelling cellular functions, computational aspects of DNA-array related work, etc., are now also part of bioinformatics. However, bioinformatics is not equivalent to computational biology and areas such as small-molecule databases, quantum chemistry, neuroinformatics, biological image analysis, theoretical epidemiology, biostatistics, and mathematical modelling in population biology and ecology are not generally considered parts of bioinformatics. Given the mounting interest in bioinformatics and the temptation to 'join the bandwagon' to cash in on the money that is becoming available, there is a danger that the term itself will be eroded into meaninglessness. It is thus important to be very clear about what is and what is not bioinformatics.

2. Inventory

Most European countries now have programs in bioinformatics. In the following, we provide short descriptions of the main initiatives known to us to date.

International centres

EBI. The European Bioinformatics Institute (EBI) near Cambridge is an outstation of the central EMBL laboratory in Heidelberg. It is the *de facto* core of bioinformatics in Europe. EBI is the central node of EMBnet, and as such has good means to distribute its data and expertise. Unfortunately, the EC has decided that international organisations such as EBI cannot be funded under the Fifth Framework Programme (FP5). This rather abrupt decision is likely to destroy any hope of using existing organisations for a structured vitalisation of bioinformatics in Europe.

EMBL. Five bioinformatics groups currently work at EMBL in Heidelberg, and these groups are actively collaborating with about 100 other bioinformaticists that work in Heidelberg (including the EMB, a venture capital bioinformatics/linguistics institute, the German cancer institute DKFZ, the University of Heidelberg, the venture capital company Lion Ag, and a BASF research institute).

EMBnet. EMBnet was founded in the late 1980s by bioinformaticists at EMBL. In those days the Internet did not exist as a medium for rapid transport of massive amounts of data, but several countries (that is Netherlands, Germany, Scandinavia) already had good national networks. To distribute the data from the EMBL databanks (now maintained by EBI) to scientists in the field, a layered distribution structure was built. National nodes would receive the data from EMBL and would then be responsible for the distribution to the scientists who needed the data. The rapid internationalisation and speed improvements of the Internet have done away with some of the need for decentralised data storage, but the EMBnet still has a role as a formal point of contact between bioinformatics activities in the different countries. At present, there are national EMBnet nodes in Argentina, Australia, Austria, Belgium, Canada, China, Cuba, Denmark, Finland, France, Germany, Greece, Hungary, India, Ireland, Israel, Italy, Netherlands, Norway, Poland, Portugal, Russia, Slovakia, South Africa, Spain, Sweden, Switzerland, Turkey and the United Kingdom.

National initiatives

Austria. There is no dedicated bioinformatics programme in Austria. Groups in Salzburg and Vienna working on structure prediction and general sequence analysis problems are funded through regular research grants.

Belgium. There are no dedicated bioinformatics programmes in Belgium, but the subject is to a limited extent covered by groups in molecular modelling and genomics (for example, groups at ULB in Brussels, Gembloux, Antwerpen, Ghent, and Namur). ULB is also hosting the Belgian EMBnet node.

France. There are two important recent initiatives in France:

- The Genopoles (www.infobiogen.fr/agora/menrt/ao-genopoles.html) are programmes co-ordinating several activities in genomics. The programmes support activities that involve bioinformatics in the different Genopoles, such as Lille, Lyon-Grenoble, Marseille, Montpellier, Paris and its surroundings, Toulouse, and Strasbourg.
- A large genomic centre at Evry (see www.aietek.com/business/genopole and http://www.genopole.com/whatis.htm). The Evry project involves the Ministry of Research, the three main research organisations in biology (CNRS, INRA, INSERM) as well as in informatics (INRIA), plus several private companies. It is also in Evry that the new INFOBIOGEN national biocomputing facility for service, data base maintenance and management as well as support for software developments is located; it is also the French EMBnet node.

In Marseille, the 'Information Générique et Structurale' Unit (IGS) (CNRS-Aventis) is officially the only CNRS unit entirely dedicated to bioinformatics (structural biology, genome analysis, experimental validation, etc.).

The development of bioinformatics is supported through new programmes set up by research organisms (CNRS, INSERM, INRA, INRIA and CEA). As an example, the 'Action bio-informatique pour la periode 2000 - 2003' conducted by the CNRS and the Ministry of Research supports both academic and joint industry/academic projects. The Ministry of Industry has also launched a programme of three to four years' duration to support collaborative projects between academic and small/medium size enterprises with a budget of 3 million euros per year.

New teaching units have been developed like the IUP ('Institut Universitaire Professionalisé') and the DESS ('Diplôme d'Etude Supérieure Spécialisé') in the field of bioinformatics including genomics.

A network for research and technology innovation was set up to promote innovation and transfer of technologies (R2IF) including bioinformatics between the Genoplante and the Genome programmes.

High-performance computing will also be developed through a national participation in the European project of the high-throughput network communication grid.

Globally, the overall financial effort in the field of bioinformatics consists of an additional support of about 10 million euros every year for the next four years. In addition to this financial supports are associated about 15 to 20 pre-doctorate research allocations (for all research organisations) as well as 20 post-doctorate research and university positions per year for four years.

Germany. In Germany the BMFT started about eight years ago an initiative aiming at the growth of bioinformatics. This initiative of about 25 million euros involved stimulation of collaborations between informatics groups, (chemical) industry and the few already existing bioinformatics institutes. The DFG has just solicited grant applications to establish centres in bioinformatics research and education. About 25 million euros have been attributed to this over a period of 5 years. Another German project is the bio-regio initiative. A bio-regio project is by definition big and geographically focused. The bio-regio project around Heidelberg has led to the creation (by ex-EMBL scientists) of Lion Ag, which is now probably the largest bioinformatics company in Europe.

The major bioinformatics groups are:

- Munich. MIPS: specialising in protein sequences, databases, genome analysis;
- Berlin. The Max-Planck-Institute for Molecular Genetics and the Resource Centre for the German Human Genome Project: specialising in support for genomics (clone administration, sequencing, mapping) and also RNA expression studies and proteomics.
- Heidelberg. EMBL, DKFZ: specialising in sequence analysis, functional genomics, genome analysis.

Greece. The major bioinformatics centre in Greece is the Institute of Molecular Biology and Biotechnology (IMBB) at Iraklio. The IMBB is host to the Greek EMBnet node and hosts and curates the *Anopheles* database. A recently established bioinformatics group focuses on a wide range of software development efforts for advanced annotation capabilities, automated literature search and metabolomics. A second, less developed, bioinformatics focal point in Greece, exists in Athens (Fleming Institute, Athens University) with main interests in protein-ligand interactions, macromolecular crystallography and the development of structure prediction and sequence comparison algorithms.

Ireland. There is no dedicated national programme in bioinformatics in Ireland, although Irish research groups are eligible to apply to Wellcome Trust (UK) initiatives including Mathematical Biology. Several university research groups working in the area of molecular evolutionary sequence analysis have been relatively successful. The well-known multiple sequence alignment program CLUSTAL was originally developed at the Irish EMBnet node, as was the database alerting service www.PubCrawler.ie.

Italy. The Italian EMBnet node is located in Bari and offers a number of services, mainly in the area of sequence analysis. It organises courses both in Bari and elsewhere.

Italy lacks a proper bioinformatics programme although there are a number of active groups in several universities, national research council institutes, and private research centres that interact frequently and organise schools and workshops. Recently, members of the Italian Society of Biophysics and Molecular Biology (SIBBM) have met to create to a 'Gruppo di cooperazione in Bioinformatica' that is actively discussing a number of aspects of bioinformatics in Italy.

The major problem in the country is the lack of a formal educational programme in bioinformatics. Only a few universities offer formal courses on the subject.

The financial support for bioinformatics activities in Italy is low. A rough estimate would be less than 250 thousand euros a year in total.

Netherlands. The Netherlands Government realised about a year ago that the country was dramatically behind in the field of bioinformatics. The simple announcement by the granting agencies that special grants for bioinformatics will become available in 2000/2001 has spurred initiatives at about five universities. It seems likely that money will be available for at least five bioinformatics groups (the one at the Netherlands EMBnet node, called CMBI, will soon consist of four professors with staff), but hiring the people is the main problem. The Netherlands organisation that distributes research funds has announced that they will make a one-off exception and will use research money for bioinformatics teaching activities.

The Nordic countries. Excluding the national EMBnet nodes, *Denmark* was first among the Nordic countries to fund a dedicated bioinformatics programme. The programme has been built around the Centre for Biological Sequence Analysis (CBS) at the Technical University of Denmark, Copenhagen (www.cbs.dtu.dk). CBS was started in 1993 by the National Danish Research Foundation, a new funding agency working in parallel with the conventional research council system. The first five-year grant awarded in 1993 amounted to around 4 million euros and the grant was subsequently renewed for another five-year period covering period 1998-2003. The CBS funding in this period (including contributions from other funding agencies) amounts to around 7 million euros. The main focus at CBS is basic research in bioinformatics, but the centre also offers training in the form of courses at the PhD and master levels.

In *Sweden*, two major bioinformatics grants have been awarded during 1999, both running for five years. One is for the Linnaeus Center for Bioinformatics (LCB) at Uppsala University (www.linnaeus.bmc.uu.se). The second is for Stockholm Bioinformatics Center (SBC), run jointly by Stockholm University, the Royal Institute of Technology, and Karolinska Institutet in Stockholm (www.biokemi.su.se/sbc). The total grants are around 8 million euros for LCB and 6 million euros for

SBC. Beyond these major programmes, dedicated funding exists for a postdoctoral programme in bioinformatics.

In *Finland* in 1996, an EMBO evaluation of biotechnology pinpointed the lack of bioinformatics skills and training. Subsequently two graduate schools have been formed in Turku and Helsinki and some courses have begun at the Åbo Akademi University, but the overall situation has not improved. In Turku, a detailed plan is now being formulated aimed at raising 32 million euros for key computational technologies, of which about 4 million euros would be dedicated to bioinformatics. The major emphasis will be in areas complementary to the developing biotechnology and pharmaceutical industry.

So far, *Norway* does not have a dedicated bioinformatics programme, although such an initiative is being considered by the Norwegian Research Council.

In *Iceland*, there is very little academic bioinformatics, though the 'DeCode Genetics' company has a rather large bioinformatics group of 15 to 20 people. There is a great need for more trained bioinformaticists, both in academia and industry.

Most Scandinavian universities offer some kind of basic courses in bioinformatics.

Portugal. There is no dedicated programme.

Spain. Spain has an EMBnet node and a national network on bioinformatics that organises several activities (annual meeting, workshops and courses). The network has also actively promoted a 'white paper' document about the importance of bioinformatics in the future of genomics and proteomics, and is preparing a proposal for a pre-doctorate program on bioinformatics that could be common to several universities, to alleviate the shortage of expertise in the country. There are two other networks also partially related to bioinformatics, one in protein structure and one in molecular modelling (proteins and drugs).

At a more general level, the new national programme for science (similar to the EC framework programmes) just starting will fund a horizontal activity in genomics and proteomics with a special emphasis on bioinformatics. Bioinformatics (sequence analysis and protein modelling) has had high priority in the national programme on biotechnology for years and all the main bioinformatics groups have national grants.

Switzerland. There is a strong bioinformatics programme at the Swiss Institute for Bioinformatics (SIB) (www.isb-sib.ch) in Geneva, centred around SWISS-PROT and the ExPASy facility. SIB is a non-profit academic institution, which receives funding from the Federal government, industry and academia.

The missions of the SIB include the development of software and databases in the field of bioinformatics, and to reinforce and increase the leading role of the databases (SWISS-PROT, SWISS-2DPAGE, PROSITE, EPD, etc.) developed by the SIB groups as well as the www ExPASy server. Since its creation in August 1993, ExPASy has been accessed more than 80 million times by more than 1 million computers from 120 different countries.

The SIB organises pre- and postgraduate courses in bioinformatics. In 1999, together with the Universities of Geneva and Lausanne, the SIB set up a Master's degree in bioinformatics.

Research activities of the SIB are carried out in collaboration with other groups in Switzerland, in France, and elsewhere in Europe (mainly with groups belonging to EBI and EMBL), in the USA (for example, the National Center for Biotechnology Information), in Australia, and in Japan. Today, the SIB counts more than 80 collaborators world-wide.

United Kingdom. Bioinformatics research is mainly funded through the Biotechnology and Biological Sciences Research Council (BBSRC), the Engineering and Physical Sciences Research Council (EPSRC), and the Medical Research Council (MRC). The Natural Environment Research Council (NERC) is increasingly interested in bioinformatics and the Particle Physics and Astronomy Research Council (PPARC) has had some cooperative involvement in bioinformatics funding.

The BBSRC has a joint bioinformatics funding initiative with the EPSRC currently calling for proposals. The BBSRC can consider bioinformatics proposals within its life sciences programmes. From time to time, the BBSRC issues calls for proposals for specific initiatives, some of which may include bioinformatics within their remits. Past examples include 'Technologies for Functional Genomics' and 'Investigating Gene Function'. There is a current call for applications in theoretical biology, which covers the mathematical content of bioinformatics. The bioinformatics initiative also awards studentships, and there has been a recent call for studentship applications (for take-up in October 2001).

There was a joint initiative on behalf of BBSRC, EPSRC, MRC, NERC and PPARC to fund new bioinformatics research groups in July 1999.

The MRC funds bioinformatics both through direct support of MRC establishments, as well as by awarding grants. It is hard to get a complete overview of bioinformatics spending; however, using approximate figures from different years, we can estimate the following annual breakdown:

- MRC Establishments (direct support) — approx. 3 million euros
- Grants (indirect support) — approx. 1.5 million euros
- Studentships — approx. 100 thousand euros

In addition there is a new bioinformatics programme at the MRC Functional Genetics Unit in Oxford with an annual spent projection of 230 thousand euros.

The Wellcome Trust funds a range of biological and medical research, which often has a bioinformatics component. It specifically funds four-to five-year fellowships in mathematical biology, which can include bioinformatics. It has also highlighted bioinformatics in its provision of four-year PhD fellowships, but with disappointing uptake. Its most recent initiative is in functional genomics, with a total budget of 160 million euros. This initiative will aim to foster a cross-disciplinary, co-ordinated approach to the exploitation of sequence data for medical benefit, and will include database development and management and other bioinformatics-related issues.

The Sanger Centre, on the genome campus at Hinxton, which is the world's largest public-domain genome sequencing centre, also has a huge bioinformatics component (some 40 staff). On the same campus, the MRC's Genome Project Resource Centre provides bioinformatics services to UK researchers, and hosts the UK EMBnet node, which is jointly funded by the MRC and the BBSRC. Finally, the European Bioinformatics Institute (EBI) is located on the same campus, making it a very significant centre of bioinformatics expertise.

As far as we have been able to ascertain, there are no dedicated bioinformatics programmes in the former eastern bloc countries, although there are a number of good groups working in the field.

Funding on a European level – EC programmes

Over the years, very little money (relative to other scientific disciplines) has been reserved for bioinformatics. In the Fifth Framework Programme (FP5) there is no longer a special heading for bioinformatics.

Bioinformatics in FP5. In FP5, bioinformatics must compete with many other topics. The infrastructure funds which in previous frameworks were mainly spent on bioinformatics have been merged with funds for medicinal test, pathology, teaching, etc. Apparently, there is a possibility to fund some bioinformatics-related projects under the Information Technology programme.

Ongoing EC-funded projects in bioinformatics. There are very few ongoing European bioinformatics projects because we are in-between frameworks. To our knowledge, only one bioinformatics project was awarded in the early rounds of FP5. A major problem is that the EC changed the rules of what is fundable in a way that was misunderstood by most bioinformaticists. In FP4, priority was placed on data collection and data-mining software. In FP5 data collection is explicitly not fundable. These rapid 180 degrees turns of policy make it hard to plan the direction of scientific work that relies on EC money.

3. Scientific strengths and weaknesses in European bioinformatics

Generally speaking, Europe is reasonably competitive in most of the key themes in bioinformatics, especially in protein structure and function prediction, protein interaction analysis and database annotations.

European academic research is less competitive in fields such as protein-ligand interaction and drug design, which is left almost completely to industry. New tools and databases in these areas end up being proprietary most of the time and the acquired knowledge is not spread among the scientific community.

Compared to the US and Japan, Europe has not been able to efficiently initiate 'big' genomics projects where bioinformatics is an integrated part.

A major weakness, especially in southern Europe, lies in the educational aspects. University courses seem to be centred around 'classical' topics, and bioinformatics has not yet been accepted by the academic world as a subject on its own. The negative effects of this on the field are twofold. First, it is at present extremely difficult to find skilled manpower, both for academia and industry. Second, many scientists enter the field without having a sufficient background in either informatics or modern biology. This leads, respectively, to the implementation of tools where already well-established 'informatics wheels' are reinvented, or to the dispersion of efforts towards not particularly relevant or novel biological problems.

4. Needs

Bottlenecks in skilled manpower, infrastructure and technologies

European bioinformatics is desperately down on skilled people. One reason is that industry is even more desperate than academia and hires people even when they are not as yet very skilled. The computing and network infrastructure is just about strong enough. To increase it would help, although this is not vital.

European collaboration

European groups make every individual effort to collaborate. Many publications in the field arise from collaborations between groups in different countries, and have produced excellent and competitive results. However, as long as this is left to individual efforts and relies on individual budgets, it will necessarily be scattered and discontinuous. This also leads to the paradox that collaboration among groups with sufficient resources of their own are easier than between groups in countries less evolved in the field. The scientists are willing, but the granting is weak.

Better integration between industry and academia?

This is more a national problem. For example, in the Netherlands the few academic bioinformaticists available collaborate intensively with industry. In many other countries this is not the case. Since the EC projects require collaboration with industry, this may be a problem in some areas of Europe, but not in others. There is a need for a European forum of discussion since the two worlds rarely meet on an equal basis. Academic science is mostly organised on the national level whereas the pharmaceutical and biotech industry is multinational.

5. Actions

Promoting exchanges and collaborations

- Create networks with the resources for financing exchanges between laboratories.
- Organise workshops on specific topics where industrial and academic scientists are equally represented.
- Organise a network of professionally maintained websites to address the issue of academic-industrial interaction.
- Strengthen national centres, taking advantage of the infrastructure of the EMBnet nodes.

Promoting training

- Make a large number of PhD studentships and post-doctoral positions available. Both the quality of the student and the quality of the institute where the PhD work will be done should be judged. It is important that the next generation of bioinformaticists in Europe get their training in the very best environments.
- Create a professional website with lectures, tutorials and practicals. Many universities in Europe have sites that partially meet this need, and it is important to collate the material on these sites under a common format.

Political actions

- The EC is advised to create a dedicated heading for bioinformatics at a level of 100 million euros for five years. The funding should be for training PhD students and connecting the existing bioinformatics centres to other groups by making funds available for training periods at those few good bioinformatics institutes. This must be implemented now, under FP5, if Europe is not to lose the fruits of the genome projects to others.
- Bioinformatics must be prioritised in FP6, both as regards collaborative project grants and in the Marie Curie fellowships programme.
- The future of the database activities at EBI must be secured.
- Contacts between academia and industry must be promoted on a European level.

Appendix 6

1. Synchrotron Radiation Sources in Europe

2. Structural Biology Beam-Lines of European Facilities

1. Synchrotron Radiation Sources in Europe

Machine and Location	Energy GeV	Storage ring circumference m	Injected current mA	Lifetime hours	Hard X-ray range	Emittance (*) nm-rad
LURE DCI – Orsay (FR)	1.85	–	320	200	Yes	1600.0
ESRF – Grenoble (European)	6.0	844.0	200	65	Yes	4.0
BESSY II – Berlin (DE) (**)	1.7	240.0	220	10	Yes	5.5
DORIS III – Hamburg (DE)	4.45	289.2	120	12	Yes	404.0
SRS – Daresbury (UK)	2.0	96.0	250	>20	Yes	130.0
ELETTRA – Trieste (IT)	2.0	259.2	300	20	Yes	7.0
MAX II – Lund (SE)	1.5	90.0	200	>10	Yes	8.8
DELTA – Dortmund (DE)	1.5	115.0	200	–	Yes	5.2
ASTRID – Aarhus (DK)	0.58	–	200	10	No	140.0
Under Construction						
SLS – Villingen (CH)	2.4	288.0	400	–	Yes	4.1
ANKA – Karlsruhe (DE)	2.5	110.0	400	>17	Yes	40-80
Planned						
DIAMOND – (UK)	3.0	500.0	300	>20	Yes	2.5
SOLEIL – (FR)	2.5	336.0	500	30	Yes	3.0
LLS – Barcelona (ES)	2.5	251.8	250	–	Yes	8.0

(*) The lower the emittance, the smaller the electron beam and therefore photon beam giving rise to a higher brilliance.
(**)A new protein cristallography facility is expected to be operational at BESSY II in the spring 2002

2. Structural Biology Beam-Lines of European Facilities

- SRS – Daresbury
- ESRF – Grenoble
- ELETTRA – Trieste
- ELSA – Bonn
- EMBL-HH (Outstation at DESY) – Hamburg
- LURE/DCI – Orsay
- LURE/ACO – Orsay[#]
- MPI-GBF – Hamburg
- MAX-II – Lund
- SLS – Villingen

[#] All facilities claim to have adequate biological support laboratories

Laboratory	Station	% Biology	Brilliance (x10^{12}) [***]	λmin(Å)	λmax	λfix	Δλ/λ (x 10^{-4})	Typical beam size as sample HxV (mm^2)	Detector	Comments[+]
1. Macromolecular Crystallography – Essentially Fixed Wavelength										
SRS	7.2	100	NA	1.5	–	1.488	3.0	0.2 x 0.2	IP (MAR)	Development station
	9.6	100	NA	0.87	–	0.87	3.0	0.2 x 0.2	CCD (ADSC)	–
	14.1	100	NA	1.2	1.5	Choice	3.0	0.2 x 0.2	CCD (ADSC)	Small wavelength changes
	14.2	100	NA	0.98	1.2	Choice	3.0	0.2 x 0.2	CCD (ADSC)	Small wavelength changes
ESRF	ID02	<5	10.0	0.7	1.6	0.97	3.0	0.05 x 0.05	IP	Now dedicated to SAS[ε]
	ID13	33	1 x 10^{-4}	0.78	2.0	Fixed	1.0	0.001 x 0.030	IP/CCD	Microfocus with special goniometre
	ID14-1	100	10.0	–	–	0.93	2.0	0.05 x 0.05	Mar CCD	
	ID14-2	100	10.0	–	–	0.93	2.0	0.05 x 0.05	CCD (ADSCR4)	
	ID14-3	100	10.0	0.89	1.44	0.93	2.0	0.05 x 0.05	Mar CCD /LARGE IP	IP 80 x 80 cm robot controlled
EMBL-HH	X11	100	0.2	–	–	0.91	5.0	0.2 x 0.2	CCD	
	BW7B	100	1.0	–	–	0.83	5.0	0.2 x 0.2	IP	
	X13	100	0.2	–	–	0.90	5.0	0.2 x 0.2	CCD	Available April 2001
LURE DCI	W32	100	0.3	0.9	1.6	0.90	10.0	0.7 x 0.2	IP Mar345	Becomes CCD in 2001
	D41a	50	0.016	1.2	1.8	1.38	10.0	0.6 x 0.6	IP	Closed down end 2000
2. Macromolecular Crystallography – Tunable Wavelength										
SRS	9.5	100	NA	0.45	2.60	–	2.0	0.3 x 0.3	CCD (MAR)	
ESRF	ID14-4	100	50	0.35	1.70	–	2.0	0.05 x 0.05	CCD (ADSC)	
	ID29	100	10	0.70	2.17		0.7	0.05 x 0.05	CCD (ADSC)	Available 2001
	BM14	100	1.0	0.60	1.80	–	2.0	0.15 x 0.15	Mar CCD/IP	Becomes CRG in 2001[χ]
EMBL-HH	X31	100	0.02	0.7	1.80	–	0.25	0.35 x 0.35	IP	
	BW7a	100	0.5	0.50	1.80	–	0.25	0.20 x 0.20	CCD	
MPI/GBF	BW6	100	0.5	0.6	2.0		3.0	2.0 x 0.40	IP	
LURE/DCI	W21b	30	0.3	0.62	3.0	–	0.1	0.7 x 0.2	IP Mar300	Becomes IP Mar345 in 2001
MAX-II [α]	1711	75	1	0.8	1.55	–	5	0.6 x 0.4	IP/CCD	CCD is Quantum 210
ELETTRA [β]	5.2R	90	10	0.5	3.1	–	2.0	1.2 x 1.2	IP/CCD	Mar165CCD from October 2000

Laboratory	Station	% Biology	Brilliance (x10^{12}) (***)	λmin(Å)	λmax	λfix	Δλ/λ (x 10^{-4})	Typical beam size as sample HxV (mm^2)	Detector	Comments(+)
SLS	PX	100	1 x 10^5	0.7	2.5	–	2.0	0.015 x 0.025	CCD/Pixel	*In-vacuum* ondulator Beam-line available in 2001, pixel detector available in 2002
ANKA	PX	90	1	0.6	3.1	–	3.5	0.3 x 0.2	IP/MarCCD	
3. Laue, and Time-resolved Measurements										
ESRF	ID09	50	2.0 for Laue 0.8 for fixed λ	0.3	2.0	Tunable	2.0	0.2 x 0.19	CCD/IP	CCD – Image Intensifier IP 30 x 40 cm off-line MD
4. SAXS – Small-angle Scattering										
SRS	16.1	27	2.0	–	–	1.41	40	5.0 x 1.0	RAPID/IP	Water bath/LINKAM
	2.1	30	0.5	–	–	1.54	40	3.0 x 0.8	MWPC/IP	Water bath/LINKAM
	8.2	7	0.125	–	–	1.54	40	3.0 x 1.0	MWPC/IP	Water bath/LINKAM
ESRF	ID2	20	10.0	–	–	0.97	3.0	0.05 x 0.05	MWPC/ CCD/IP	a) Continually variable detector-sample distance b) Off-line IP
EMBL-HH	X33	100	1.0	–	–	1.5	50	–	MWPC	SAXS/WAXS
LURE DCI	D24	55	0.032	1.2	1.9	Fixed	10	3 x 0.5	MWPC/IP	
	D43	20	0.016	0.7	1.8	Fixed	100	0.07 x 0.07	CCD/IP	Diffuse Scattering Disordered Systems IP is off- line MD
ELETTRA	5.2L	50	5.0	0.77	2.3	–	25	5.4 x 1.8	ID & 2D MWPC	SAXS/WAXS; microfocus 3λ = 0.7, 1.54, 2.30 ≈
5. X-Ray Absorption Spectroscopy (only lines with significant biological usage are listed)										
SRS	8.1	11	0.3	1.1	3.5	–	1.4	2.0 x 2.0	SS-13 element	
	9.2	19	0.02	0.25	2.0	–	0.5	Unfocused	SS-13 element	
	16.5	27	0.45	0.30	1.4	–	0.5	2.5 x 2.5	SS–30 element	High-count rate electronics
	7.1	11	0.36	1.1	3.0	–	1.4	2.0 x 2.0	SS-9 element	Electronics

Laboratory	Station	% Biology	Brilliance (x10^{12}) (***)	λmin(Å)	λmax	λfix	Δλ/λ (x 10^{-4})	Typical beam size as sample HxV (mm^2)	Detector	Comments(+)
ESRF	ID26	New BL	5	0.5	2.5	–	1 – 10	0.2 x 0.02	Multi-channel Si drift	Ultra-dilute samples and QEXAFS
LURE DCI	D21	30	0.016	0.4	6.2	–	1	3to15 x 0.5to1	SS – 7 element	The beam is unfocused
ELSA	–	10	0.01	6.2	12.4	–	–	2 x 10	Ionisation chambers	
	–	10	0.01	6.2	12.4	–	–	2 X 10	Semi-conductors	
EMBL-HH	EXAFS	100	0.01	0.4	2.0	–	0.14	1 x 5	SS –13 element	Absolute energy calibrator
6. VUV and IR										
SRS	3.1	40	0.005	350	5000	–	5.0	6.0 x 1.0	Photo-multiplier	Circular Dichroism
	13.1a	80	–	2000	7000	–	200	250x250 nm	Photo-multiplier CCD array spectograph	Confocal Microscopy
	13.1b	50	–	1900	10000	–	10	1.0 x 1.0	Photo-multiplier	TR fluorescence & CD
	13.3	40	–	10^4	10^6	–	White beam	0.01x0.01 (microscope)	MCT (microscope) Bolometre (RAIRS)	Operate as MID-IR microscope or FAR-IR RAIRS
LURE/ ACO	SA1	50	–	2000	7000	–	10	0.1 x 0.1	μ-channel plates	
	SB1	50	–	2000	10000	–	10	0.1 x 0.1	μ-channel plates	TR fluorescence used in 2-bunch mode only
	SA4	50	–	2100	12000	–	10	10.0 x 10.0	Photo-multipliers	
	SA5	70	10^5	5 x 10^5	10000		100	0.008 x 0.012	Photo-conductor	Micro-spectroscopy & chemical imaging
MAX-I	–	20	–	2000	24000	–	–	2.0 x 2.0	μ-channel plates	TR fluorescence in UV and visible

(***) This column should be read with caution – It is not clear that all facilities have used the same definition of brilliance
(+) All PX stations appear to have cryo-cooling stations
(c) Retains the ability to undertake PX studies and can be used for such under unusual conditions or an emergency
(x) Several CRG beam lines are available on which ESRF has 1/3 total availability
(α) Other stations 100% dedicated to biology are planned at MAX-lab: a MAD beam line (for 2002) and up to 4 independent side-stations
(β) A second PX beam line is under construction (for 2002)

Further beam lines for structural biology applications will become available in the next 5 years at ANKA (Karlsruhe) and the Swiss National Light Source. At the SRS, Daresbury, one additional VUV SR circular dichroism beam line is in the construction phase to replace old beam line 12.1 (another new beam line for NCD on a multipole wiggler is also under construction but mostly for physical sciences). Replacement synchrotrons are agreed to be built in France (SOLEIL) and the UK (DIAMOND) and a new source is being planned for Spain (LLS at Bellatera). No information was made available from BESSY (Berlin).

IREG Strasbourg - Dépôt légal: juin 2001 - N° d'ordre: 011903