Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Inter-union bioinformatics group report

Herman J. C. Berendsen
Biophysical Chemistry, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands

1. Introduction

The IUBG was established on initiative of the IUPAB in 1998 as an Inter-Union activity addressing issues concerning the availability, maintenance, and free access of biological and biophysical scientific data. The availability of sequence data on the human genome and on gene sequences of other organisms, and of structural data on proteins and other bio-macromolecules, is growing rapidly. Combined with the growing impact of these data on applied sciences and medicine with their concomitant commercial interests, concerns have been raised about the proper archiving of such data, its quality control, the guaranteed unrestricted access, and the safeguarding of archival data for future generations. Such primary scientific data as sequences and structures must be considered as cultural assets that should be made accessible and kept available for future generations as part of the human heritage. In a way they are comparable to basic scientific data such as the physical properties of the elements, or to taxonomic data of biological organisms, which are well-documented and readily available to the worldwide community of pure and applied scientists.

A difference with the 'classical' data is that the latter are archived and available in the open literature, which is kept accessible by scientific libraries of universities and institutes. National libraries have the specialized task, recognized and supported by the governments of several countries, to safeguard the access of relevant documents for future generations. On the contrary, bio-data are presently not published in detail in scientific journals, but rather deposited in databases. These databases are maintained by institutions that do not have the support status of national libraries. It is not yet generally recognized at the government level that the archiving of such data needs protection similar to the archiving of literature, and that governments and supra-national bodies bear a responsibility to maintain the collections and safeguard their integrity and access into the far future.

As the quantity of primary information grows at an exponential rate, its proper archiving and maintenance become more and more a matter of concern. Organisations that maintain archival databases are confronted with an ever-growing work load. Without adequate public funding they may be forced to make selections based on non-scientific considerations, relax the strict requirements of quality control, or impose restrictions to access that are detrimental to public availability in a worldwide community. There is also a need for international cooperation of the institutions concerned, and for the definition of standards for data treatment and data exchange.

Another matter of concern is the national and international legislation on intellectual property rights, which is at present actively discussed. While there is no doubt that the author of a creative process possesses intellectual property rights on the product of his creative process, and that such rights need protection by law, there is also no doubt that primary biological data *per se* can never become the subject of intellectual property rights, and their 'fair use' for scientific and educational purposes must not be impeded by regulations. Careful distinction must be made between the data themselves and the added value of the form in which they are presented, or the processes to which they are subjected.

The rate of growth of the biological primary data and of the various techniques that are being developed to make use of these data, present challenges to scientists as well as to current university curricula. Not only are the changes fast and the applications novel, but the combination of skills required for managing these data is far from traditional. One must combine aspects of biological or medical sciences with mathematics and informatics in an unprecedented way. Apart from a selected group of specialists, the majority of scientists will be left behind if educational curricula do not quickly adjust to these changes. The IUBG feels that much can be gained by international cooperation in aspects of education, and addresses this topic as well in this report.

2. The IUBG and its mission

The IUBG is a Joint Initiative of the International Union for Pure and Applied Biophysics (IUPAB), the International Union of Biochemistry and Molecular Biology (IUBMB), the International Union of Crystallography (IUCr), the International Union of Pure and Applied Chemistry (IUPAC) and the Committee on Data for Science and Technology (CODATA). The IUBG seeks overlap with the nomenclature committees of IUCr, IUBMB, IUPAC, and CODATA. The IUBG has received support from the ICSU (International Council for Science) Grants Programme and UNESCO (United Nations Educational, Scientific and Cultural Organisation). [See the online Appendix for a list of acronyms]. Its mission statement is:

To monitor worldwide developments in Bioinformatics

To take measures as required to ensure and facilitate inter-process communication, such as standardization of data formats

To act when the continuity or reliability of key informatics providers is endangered

To act when the free access to data in the public domain is endangered

To catalyse actions by the appropriate authorities in areas of the world where Internet access to servers and data providers is technically inadequate

To organize relevant educational activities.

The steering committee consists of

Herman J. C. Berendsen (former president IUPAB), Secretary

Helen M. Berman (IUCr)

Richard Cammack (IUBMB)

Charles Cantor (former IUPAB Council)

Jean Garnier (President IUPAB), Chairman

Arthur Lesk (representing CODATA)

Alan McNaught (IUPAC)

Richard Roberts (ICSU)

M. Vijayan (IUPAB, IUCr)

In addition to the members of the steering committee, as mentioned above, the following advisors were present on invitation at one or more of the IUBG meetings, and have contributed to this report:

Rolf Apweiler (Swissprot/EBI)

Michael Ashburner (EBI)

Philip Bourne (PDB)

Nobuhiro Go (Kyoto)

Marianne Minkowski (ESF: European Science Foundation)

Carlos Martinez-Riera (EC: European Commission)

Peter Murray-Rust (Cambridge University)

Barry Robson (IBM)

Patricia Rodriguez-Tome (EBI/EMBL)

John Rumble (President CODATA)

Jay Russell Snoddy (ORNL)

Hideaki Sugawara (DDBJ).

The following meetings were held:

28-29 June, 2000: Whitehead Institute, Cambridge, Mass, USA

17 October, 2000: Baveno, Italy

22 October 2001: EBI, Hinxton, Cambridgeshire, UK

14 May 2002: ICSU, Paris, France

3. Primary scientific data: care and responsibility

3.1. Classification of data

There are two basically different *types* of databases: archival databases and derived databases. The former carry the primary public data; the latter use or recombine these data to derive new properties and relations. One can also distinguish databases by their *content:* there are structural databases and protein and nucleic acid sequence databases. There are databases on taxonomy, on metabolic pathways, on phenotypes, on biodiversity, and many more, as well as organism-specific databases. Finally, one may distinguish databases by their *ownership:* public and private.

The IUBG concerns itself primarily with the archival databases, in the fields of biophysics/biochemistry/molecular biology. The archival databases are those to which the researchers submit their data (or which collect data actively from primary sources); the proper archiving of primary biological structure and sequence data for the future is put into their hands by the scientific community. This imposes a great responsibility on the database managing authorities and their sponsors. Not only must they provide lasting storage; they must - as much as possible - provide for validation and annotation of the data, and cooperate to avoid duplication and ensure completeness. These responsibilities should be made explicit.

3.2. Obligations for journals and funding agencies

Since journal publications do not include all of the primary data described in research reports, it must become mandatory to deposit data in a public archival database, and journal editors must require such deposition before accepting a manuscript for publication. In the case of structural data of biological macromolecules, IUCr plays a vigorous role in providing guidelines for deposition and release of data. Its publications carry much weight and have resulted in the practice that such structural data are indeed deposited into the Protein Data Bank (PDB). Unfortunately such a requirement has not become common practice for structural data of small molecules. Neither does a requirement exist for public deposition of thermodynamic, kinetic, and spectroscopic data. For nucleic acid sequence data the situation has become unclear and the scientific journals do not at present follow a common policy. Although it would follow a long tradition that scientific publications are only acceptable if they include the data on which the publication is based, there are recent examples that publications have been accepted when the relevant data were only promised to be made available by the author or his (commercial) sponsor on a private site; this limits free access and lacks the guarantee of a permanent archive of the data. One restriction is that public archival databases are not allowed to copy such data into their database. These examples include the publications in Science about the human genome [Science **291** (2001) 1304-1351] and the rice genome [Science **296** (2002) 79; 92], in which cases the data are available from commercial databases. The acceptability of such publications is at least questionable, and the International Scientific Unions are urged to consider the principles and set out guidelines for data deposition related to scientific publications. There should be a balance between the commercial interest of the data producer and the requirement of accessibility of data which accompany scientific publications. The commercial interest could be protected by a well-defined limited period of denial of public access, and the requirement of public access could be met by the deposition of the data in a public archival database under a suitable embargo. There should be a wide acceptance of requirements of this type, and articles from authors who do not follow the agreed requirements should simply be rejected by scientific journal editors.

Adherence to the principles outlined above should be demanded by the *research supporting agencies* which are publicly funded, as an imposed condition on the recipient of the funding. To be precise, the sponsor should require that data generated during the course of a funded project, be deposited in public archival data bases, and allow submission of publications only to journals that adhere to the principle of public deposition. The funding agencies are – as many do in practice – urged to follow the recommendations of international bodies representing the scientific community, such as the International Unions. Such a policy of the major funding agencies will be an effective tool to enforce implementation of these principles in the scientific and publishing world. It will counteract any tendencies that may otherwise arise to exploit data generated by public funding to the benefit of local rather than public interests.

The general opinion is that archival databases should be public. Public means unrestricted access in the sense that *no moral judgments may be made who should get the data,* but 'public' is not the same as 'free'. The costs of maintaining the archival databases must be recovered somehow. The ideal is that data are both public and free, but this requires a worldwide commitment of national and international funding agencies.

3.3. International obligations

The maintenance of archival databases is in essence a supranational activity. The obligation to deposit data must be followed worldwide. There must be *a single archive* even if distributed over more than one site. Data must remain uniform. Therefore, ideally an international agreement should be reached for funding international archival databases. At present, there is an asymmetry in funding: there are different funding models for the sequence databases and the structural databases and there are different funding models in Europe, the USA, and Japan. The funding models change with time. The PDB is funded under a 'memorandum of understanding' by a consortium of US government agencies (NSF, DOE, NIH); GenBank is funded by a special allocation by the US Government. The deposition sites in Japan, the DDBJ for sequence data and the Institute of Protein Research at Osaka University for structural data, are financed by MEXT, the Japanese Ministry of Education, Culture, Sports, Science and Technology. The European deposition site at the European Bioinformatics Institute (EBI) functions as an outstation of the European Molecular Biology Laboratory (EMBL) with limited basic funding, extended with funding by the EU and the Wellcome Trust. Other national institutions do not have a recognized role in the international safeguarding of archival data. The ideal would be a wide international agreement at the

government level that can stabilize the situation and guarantee cooperation, consistency, and funding.

While several instances exist of international organizations with funding at the government level, directly or jointly through a United Nations organization, the establishment of such an organization is not at all trivial and is considered premature for the present purpose. The subject is likely to be considered too specialized for agreements at the government level, and referred to existing international bodies as EMBL and EMBO and to the national and supranational *research funding agencies*. It is in fact in the direct interest of fundamental and applied research worldwide, and thus of the research funding agencies, that scientific data are properly archived, validated and made accessible. Therefore the research funding agencies, both public and private, and both national and supranational, are in fact responsible for safeguarding scientific data. Although this applies to such organisations worldwide, the primary responsibility lies with the major funding agencies in the USA, Europe and Japan, who have the means to fulfil this task. It is important that they will also show their resolve to fulfil this task, which requires long-term commitments to support the archival databases.

Support by the funding agencies involves more than the support of short-term projects related to the archival data bases. It is mandatory that the funding agencies commit themselves to *long-term support* of the archival databases. This may involve long-term support of the data-archiving activities, but in the interest of flexibility and proper quality control the support should be made dependent on a periodic review of the performance. The exact terms are of course a matter of the funding agencies themselves; the essential ingredient is that the major funding agencies consider data archiving activities as their permanent responsibility. The responsibility extends – under strict quality control – beyond existing and recognized archival data bases to new archiving activities if the need arises.

4. The public databases

In this section we give a survey of the archival data bases as they presently function, followed by requirements related to the validation and quality of the data.

4.1. Major archival databases for biological molecules:

1. Nucleic acid sequences: *DDBJ/EMBL/GenBank International Nucleotide Sequence Database (INSD)*: NCBI (Natl Center for Biotechnology Information at NLM/NIH; 'GenBank Database')/EBI (European Bioinformatics Institute, Outstation of EMBL; 'EMBL Nucleotide Database')/DDBJ (DNA Data Bank of Japan, Natl Institute of Genetics; 'DDBJ Database'). Data can be submitted to any of these and their contents are daily synchronized. EBI and DDBJ each receive 10 to 15% of the entries, and GenBank receives 70 to 80%, but the three databases provide the same entries. Aim of the INSD is to record every publicly known nucleic acid sequence. Data are freely available.

2. Protein sequences (including post-translational modifications): *PIR-International* (Protein Information Resource) is a collaboration between NBRF (Natl Biomedical Research Foundation, Georgetown)/MIPS (Munich Information Center for Protein Sequences at the Max Planck Inst. for Protein Research, Munich)/JIPID (Japan International Protein Information Database). Data are freely available. *SWISS-PROT* is an annotated protein sequence database maintained collaboratively by the Swiss Institute of Bioinformatics (Geneva) and the EBI. Data from SWISS-PROT are freely available for non-profit institutions only. Commercial users pay a yearly license fee. SWISS-PROT is not strictly an archival database, but it is unique in its annotation, accuracy and links to other databases. TrEMBL is a supplement to SWISS-PROT providing computer-annotated entries derived from the translation of all coding sequences in the EMBL Nucleotide Sequence database. SPTR (SWISS-PROT/SP-TrEMBL) is a weekly updated combination of both databases and checks on additional data in PIR and PDB.

Biomolecular structure: The *Protein Data Bank (PDB)* of protein structures based on X-ray or NMR data is operated by the RCSB (Research Collaboratory for Structural Bioinformatics), funded jointly by the US Natl Science Foundation, Department of Energy, and two units of the Natl Institutes of Health. Partners are SDSC (San Diego Supercomputer Center at La Jolla, CA), Rutgers University (Piscataway, NJ), and NIST (Natl Inst. of Standards and Technology, Rockville, MD). Data can be submitted to the US site at Rutgers, the Institute of Protein Research at Osaka University in Osaka, Japan, or the MSD site at the EBI. The *BMRB* (*BioMagResBank* at the University of Wisconsin, Madison) collects primary NMR data. PDB and BMRB data are freely available. The *NDB* (*Nucleic Acid Database*) distributes data about nucleic acid structures and is managed by Rutgers University. It is funded by the National Science Foundation and the Department of Energy. The *CSD* (Cambridge Structure Database) is managed by the *CCDC* (Cambridge Crystallographic Data Centre), contains small molecules, and is available for a fee.

3. Other databases: there are many databases of biological interest that are not mentioned here (bibliographic data, taxonomic data, data on specific organisms, data on lipids, carbohydrates, protein ligands, vitamins, *etc.*, metabolic data, *etc.*). If they are archival, they serve in principle the same public interest as the major archival databases, and such databases may encounter similar problems related to free access and funding as those discussed by the IUBG.

4.2. Requirements for validation and quality of public databases. Problems of redundancy.

There are two aspects: *quality control* and *annotation*. Quality control is an essential process before data can be entered into an archival database. It involves error and redundancy checking, correspondence with authors, etc. Corrections are difficult and require skill and experience. A good partnership is needed between the curator and the depositor, and also between the curator and the programmer. Although error-free data can never be guaranteed, the usefulness of a database depends on its accuracy and reliability. Annotation, such as the addition of literature references, the comparison of gene sequences with other organisms, the relation of nucleotide sequences to proteins, the comparison with other sequences or proteins, and the reference to possible protein function, involves cross-referencing with literature and other data and, again, requires skill and experience. Archival databases need validation and annotation to be useful; however, the possibilities and requirements are subject to change and the standards to be applied to validation and annotation should be set by periodic peer review.

With the increasing rate at which data are deposited, curation and annotation will become a quantitative and financial problem. Some two hundred new genes (or rather genetic loci) come in each day: how can one keep pace with annotation? It is required to hire, train and retain skilled curators and annotators. There are good opportunities to involve countries around the world in distributed annotation activities and in database-related work; countries like

India have a large reservoir of well-trained people. Much of the work can be done in a distributed fashion while maintaining the integrity of the annotation standards and methods. A program of visits by experts to these countries and visits of scientists from these countries to the data centres could enhance the cooperation with and involvement of countries outside US, EU, and Japan.

4.3. Threats to continuity and availability of publicly funded biological databases

As mentioned above, there are several different funding mechanisms for the archival databases. They all operate on a short-term basis, and lack the long-term commitment needed for the guaranteed and continued functionality and maintenance. What is particularly lacking is an international agreement between sponsoring organisations and a commitment of such organisations in other countries than the USA, the EU and Japan. The international cooperation is now based on the goodwill of the people involved and not on agreed international cooperation, backed up by the commitment of sponsoring organisations. In such a *chaotic* situation, the integrity of the archival databases is threatened.

The IUBG recognizes the need to convince government bodies and publicly funded research organisations of their responsibilities to maintain infrastructure for archival databases in order to guarantee continuous archiving and international access to the archived data. The archived data are our heritage!

5. Intellectual Property Rights (IPR)

The concern of the IUBG over free access to primary data bears on the Intellectual Property Rights (IPR) issue that is widely debated at present. In 1996 the UN organization WIPO (World Intellectual Property Organization) published a draft 'Treaty on Intellectual Property in Respect to Non Original Databases', extending existing copyright agreements to databases and supporting an overly protectionist property rights regime. This draft was not adopted, however. Its far-stretching consequences were immediately attacked by the US National Committee for CODATA in a 1997 report 'Bits of Power. Issues in Global Access to Scientific Data'. The European Union has adopted a Directive on the Legal Protection of Databases in 1996 (supposed to be implemented in national laws by Jan 1, 1998) which extends copyrights to material contained in databases without making appropriate explicit exceptions for fair use for personal, scientific, and other non-commercial purposes. This Directive has raised much concern, because - if such exceptions are not specified - the resulting laws could impair the free scientific development by making original data not freely available. The matter has been taken up by ICSU/CODATA (see http://www.codata.org/codata/data_access/summary.html) and a meeting on the subject was held in Baveno, Italy, on Oct. 14, 2000. It is also a subject of debate in connection with the proposed Global Biodiversity Information Facility (GBIF, still in planning stage).

The IUBG is much concerned with over-protective measures that would limit the worldwide free sharing of primary biodata. The primary biological data, including all sequence and structure data derived from natural organisms, should be freely accessible and their fair use for scientific purposes should not be limited. The IUBG supports the activities undertaken by ICSU/CODATA.

A related topic is the *patentability* of nucleotide sequences. Measured sequences are facts and not inventions, and should not be possible objects for patents or copyrights.

6. Worldwide access to public data and inadequacies of the Internet

Bioinformatics and biotechnology services take the shape of 'interactive only' sessions in the vast majority of sites involved in the service provision of genomic and gene-related data. The network infrastructure that makes this possible is improving all the time and expands beyond European and North American borders. However, the pace at which this growth happens is directly related to the robustness in the relationships between network partners in various continental regions, and to the flow of funds required to establish, maintain and upgrade the networks. As a result, there are varying service levels for the biotechnology user community in respect to these resources over the Internet: some enjoy good connectivity and can submit jobs to interactive services, while others are limited to accessing these resources through email (this includes many parts of South and Central America, the Caribbean, Africa, Indo-Asian subcontinent, many countries in the Pacific rim, and sometimes even in the USA and Europe).

In the present situation the data-providing centers must make sure that the tools for information retrieval and analysis are available to the biotech community regardless of how well connected one is to the Internet. These services must have both interactive and email based interfaces, while availability on CD-ROM provides an additional service. While the interactive forms are self-explanatory due to their visual nature, the email interface is not, and the degree of experience that is required to become aquainted with all options is very high. Therefore users with insufficient Internet functionality do not in fact obtain a quality of service equal to that of the more fortunate. It is in the interest of the high-level participation in international science that internet facilities grow to high quality on a worldwide basis.

The general internet situation in third world countries is a subject that goes far beyond the possible interventions of IUBG. The problems are common to most scientific fields, although in the Bioinformatics field the insufficiencies are most apparent. The IUBG will support any action of ICSU bodies (like CODATA) to address this issue. Also in Europe, as in other non-US countries, academic networks should improve. Dedicated research networks are not a solution. The global message is: *good access is required.*

Access in the present context could mean not only access to information in the databases, but also *access to the process of creating and maintaining databases*. At present, the number of people involved in the process are few and originate from a few countries. Highly competent manpower exists in some reasonably developed, but comparatively poor countries in Asia and Latin America. Involvement of well-trained scientists and technicians from such countries in the development and maintenance of major databases could be mutually beneficial. On the one hand, these scientists and technicians would get an opportunity to participate in bioinformatics operations at the international level. On the other hand, the organizers of the databases would have access to trained manpower. In fact, it is possible that much of the work could be carried out by the scientists and technicians in their home laboratories in the course of cooperative projects, with occasional visits to the main sites of the databases.

7. Standardization issues

7.1. Standardization of data definitions and nomenclature.

Standardization of data definitions and nomenclature is important to prevent

confusion and to facilitate data exchange. There are two reasons to enforce a controlled vocabulary: *uniqueness* and *equivalency* on various levels. Requirements for classifications are: they should be unique, concise, informative, and easy for searching. At present naming and classification is missing for large classes of biomolecules. A systematic definition of the *ontology*, describing the hierarchy of terms in which the knowledge in a field can be expressed, is needed for the fields of genetics, molecular biology and bioinformatics. Compatibility in data definitions is just one step ensuring that the data are expressed in machine-readable form and that data can be unequivocally interpreted.

Standardization of nomenclature is not a matter for a committee like the IUBG. Experts are needed to establish standardization and classification. The Unions should take responsibility on behalf of the scientific community and establish working groups for the creation of a *controlled vocabulary* for archival bioinformatics resources. They should also help secure funds to finance such activities.

An example of existing standards is the EC number classification for enzymes. Some enzymes have more than 50 names, but a functional classification, based on reaction catalysed, had been found to be the most useful and robust system. It is restricted to enzymes that catalyse a distinct and well-defined reaction, and not every property of the enzyme is considered. The classification is given by a *number* like EC 1.1.99.238, plus a *common name*, plus a *systematic name*, plus a *comment section* with cross references to sequence, structure, cofactors, substrates, metabolism. The number is composed of

EC: defines the database
1 (first number): defines the enzyme type
1 (second number), and 99 (third number): define the reaction
238 (fourth number): is a catalogue number.

The databases should use standard nomenclatures for their contents if these nomenclatures have been defined.

7.2. Standardization of data formats and data exchange

Considering data formats, the most important requirement is that *the format should be specified completely and in detail*. For each type of data there must be an agreed format such that the data are machine-readable and their meaning is unequivocal.

8. Education

There is a demand for education in bioinformatics for students of life sciences and chemistry, at levels from undergraduate to specialist postgraduate, and for conversion courses for teachers. This can be through degree programmes or summer schools. Bioinformatics should be integrated into the life sciences curricula.

Training in diverse disciplines is required for those constructing databases. Biologists, computer scientists and software engineers will be involved in teams for creation of databases. Biologists need training in programming and scripting languages, while computer scientists need appropriate training in biology and chemistry.

9. Conclusion

Following the analysis presented above, the IUBG has formulated a series of *Statements and Recommendations* as an independent document (see below). These summarize the report and provide the necessary guidelines for further actions.

10. Statements and Recommendations from the Inter-Union Bioinformatics Group (IUBG)

10.1. Statements

Statement 1 on *the safeguarding of biological data.*

It is the obligation of the scientists and legislators of all nations to archive and support primary (i.e., fundamental experimental) scientific data, including, but not exclusive to, nucleotide sequences of biological organisms, amino-acid sequences of proteins, three-dimensional structures of biological molecules, as well as other primary data produced by genomics and proteomics studies. These data must be validated, stored, made publicly accessible, and safeguarded for future availability and access. Access must be public and unrestricted and no organization should have a monopoly on these data. These primary scientific data are crucial for the development of science and its applications.

Statement 2 on *the obligations of data generators*

It has always been the practice that those who claim scientific advances by publication of their work should support their claim by making openly available the objective data on which their claim is based. Thus it is the obligation of scientists who generate primary biological data in the course of publicly funded research to preserve these data for present and future reference and unrestricted access. Regardless of whether publication in journals is appropriate, such data must be deposited into the archival databases to guarantee their present and future availability. Primary data producers in the private sector are also urged to conserve and eventually deposit their primary data.

Statement 3 on *right to fair use of data*

Scientific advances rely on full and open access to data. Primary data that are accessible through the archival databases should not be subjected to any restrictions that would limit fair use of those data. Fair use includes the use for teaching and research purposes.

Statement 4 on *standardization issues*

There are four different aspects associated with primary data for which standardization should be considered: *content, nomenclature, data format,* and *data exchange protocol.* Standardization is an ongoing activity requiring high-level agreement among scientists of various fields in order to ensure understanding and knowledge exchange across borders of scientific disciplines.

Statement 5 on *education*

Considering the skills required for archiving, validation and dissemination of data, educational institutions should recognize the need for specific education in (bio)molecular informatics

10.2. Recommendations

Recommendation 1 *to International Unions and scientific societies*

(a) It is recommended that each Union, on a regular and ongoing basis, identifies and publicizes a list of key archival data bases.

(b) It is recommended that International Unions and other scientific societies actively encourage their membership to deposit primary data in recognized data repositories which provide unrestricted access to these data.

(c) It is recommended that journals of these Unions and societies ensure that these requirements are met before accepting publications in their journals.

(d) It is recommended that International Unions, specifically IUPAB, address the general issue of education in the field of (bio)molecular informatics.

Recommendation 2 *to funding agencies*

(a) It is recommended that funding agencies insist that all primary data produced by grants that they fund be deposited in recognized data repositories which provide unrestricted access to these

data. International Unions and scientific societies should work with funding agencies to create guidelines for this purpose.

(b) It is recommended that funding agencies actively encourage and adequately support existing and newly funded primary data repositories, including their updating and annotation, to provide the mechanism to preserve in perpetuity the data deposited therein and to preserve it in a form which is fully recoverable by future generations of researchers.

Recommendation 3 ***to for-profit organizations***

It is recommended that for-profit organizations deposit their data as early as possible in public archival databases.

Recommendation 4 ***to publishers and authors***

It is recommended that journal publishers make the primary data on which a publication is based available under the same conditions as they make the printed article available, and – if applicable – require that such data are deposited in a recognized key archival database. Authors are encouraged not to publish in journals that do not conform to these rules.

Recommendation 5 ***to legislators***

Following the recommendations of the ICSU/CODATA Ad Hoc Group on Data and Information, it is recommended that legislators take into account the impact of intellectual property laws on research and education, in order to allow fair use for scientific and educational purposes.

Recommendation 6 ***to scientific committees for nomenclature and standardization***

It is recommended that International Unions and scientific societies play an active role in the definition of standards in the fields they represent. This should be done through nomenclature and data standardization committees. They should be conversant with both the content and the technologies needed for a full definition of the field, in order to ensure the exchange of data without loss of information.

Recommendation 7 to ***educational institutions***

It is recommended that bioinformatics curricula should include specific education in the creation and curation of databases, as well as in their use. Life sciences curricula should include courses and training in bioinformatics.

The work of the IUBG was supported jointly by the ICSU Grants Programme and UNESCO. The report was adopted and its recommendations were accepted by the 27th General Assembly of ICSU in 2002. The report is available online at http://md.chem.rug.nl/~berends/IUBG-FinalReport.html, and *via* http://www.iupab.org. See also *Nature* (2002) **419**, 777.