280–283 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

Organising multi-dimensional biological image information: The BioImage Database

J. M. Carazo*, E. H. K. Stelzer[1], A. Engel[2], I. Fita[3], C. Henn[4], J. Machtynger[5], P. McNeil[6], D. M. Shotton[7], M. Chagoyen, P. A. de Alarcón, R. Fritsch[1], J. B. Heymann[2], S. Kalko, J. J. Pittet[4], P. Rodriguez-Tomé[6] and T. Boudier[7]

Centro Nacional de Biotecnología-CSIC, Campus Univ. Autónoma, E-28049 Madrid, Spain, [1]European Molecular Biology Laboratory (EMBL), Postfach 102209, D-69012 Heidelberg, Germany, [2]M. E. Müller Institute for Microscopic Structural Biology, Biozentrum der Universität Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland, [3]Centro de Investigación y Desarrollo-CSIC, Jordi Girona Salgado 18–26, E-08034 Barcelona, Spain, [4]Advanced Technology Center Silicon Graphics Inc. Chemin des Rochettes 2, CH-2016 Cortaillod, Switzerland, [5]Informix Software Ltd., 6 New Square, Bedfont Lakes, Feltham TW14 8HA, UK, [6]The European BioInformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [7]Department of Zoology, Univeristy of Oxford, South Parks Road, Oxford OX1 3PS, UK

Received September 8, 1998; Revised and Accepted October 21, 1998

ABSTRACT

Nowadays it is possible to unravel complex information at all levels of cellular organization by obtaining multi-dimensional image information. At the macromolecular level, three-dimensional (3D) electron microscopy, together with other techniques, is able to reach resolutions at the nanometer or subnanometer level. The information is delivered in the form of 3D volumes containing samples of a given function, for example, the electron density distribution within a given macromolecule. The same situation happens at the cellular level with the new forms of light microscopy, particularly confocal microscopy, all of which produce biological 3D volume information. Furthermore, it is possible to record sequences of images over time (videos), as well as sequences of volumes, bringing key information on the dynamics of living biological systems. It is in this context that work on BioImage started two years ago, and that its first version is now presented here. In essence, BioImage is a database specifically designed to contain multi-dimensional images, perform queries and interactively work with the resulting multi-dimensional information on the World Wide Web, as well as accomplish the required cross-database links. Two sister home pages of BioImage can be accessed at http://www.bioimage.org and http://www-embl.bioimage.org

INTRODUCTION

It is often said that an image is worth a thousand words. However, implicit in this statement is the capability to access the specific image portraying the situation we are interested in studying, as well as a precise understanding of the informational content of this image. The BioImage Database Project is designed to make this process of accessing, retrieval and understanding image information of multi-dimensional biological specimens an easy and user friendly task for the scientific community at large.

On technological grounds, nowadays a broad range of approaches, chiefly the diverse forms of microscope techniques, can produce either images, volumes, or time series of images and volumes that help unravel key biological processes ranging from macromolecular structure to subcellular and cellular structure and organisation. Collectively, we will refer to these different types of images, volumes and time series as ‘multi-dimensional images’.

In spite of its importance in biology, the situation has been that such multi-dimensional data were not organised in any database. In fact, there has been no way either to query specifically for multi-dimensional images or to automatically retrieve any such images. This clearly unsatisfactory situation led to the start of several exploratory projects from 1993 to 1996, which helped define the general need within the scientific community for a database for such multi-dimensional images. A proof-of-concept prototype was developed and was described by Marabini *et al*. (1). Subsequently, the design and implementation of the BioImage database of multi-dimensional microscopic images of biological specimens has been the work of a multi-national consortium of research laboratories together with technical industrial partners. Its first implementation was released in September 1998, in the 14th International Congress on Electron Microscopy, held in Cancun, Mexico. Further information about the BioImage database can be found in the two sister home pages of the project at www.bioimage.org and www-embl.bioimage.org. An open discussion forum for BioImage has just been set up

*To whom correspondence should be addressed. Tel: +34 91 585 45 43; Fax: +34 91 585 45 06; Email: carazo@cnb.uam.es

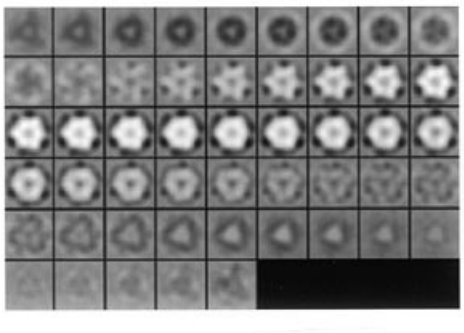

Figure 1. Example of a BioImage entry (only the multi-dimensional image—and not their accompanying metadata—is shown): Volume of the helicase DnaB of *E.coli* obtained by cryo-electron microscopy and image processing at 3.2 nm resolution. The volume is presented as a series of 2D sections parallel to the electron microscopy grid from ref. 4. (© Current Biology Limited, reproduced with permission).

through an Email distribution list at the address bioimage@listserv.cnb.uam.es.

INFORMATION CONTENT OF BIOIMAGE

Focusing on multi-dimensional digital images, we can define them as samples of a given magnitude over either a 2D, 3D or 4D matrix of elements generically refereed to as 'pixels'. The information contained in these images relies upon both the pixel values as well as the spatial and temporal relationships between them.

An example of the type of volume information we refer to is presented in Figure 1, where we reproduce 2D sections parallel to the specimen support grid corresponding to the 3D structure at pH 7.8 of the helicase DnaB from *Escherichia coli*, as obtained by cryo electron microscopy and 3D image processing techniques at 3.2 nm resolution (2). It is clear that, in spite of the fact that we know the precise amino acid sequence of the protein and its quaternary architecture as a homohexamer, we do not have sufficient resolution to interpret this volume in terms of its biochemical constituents and obtain, for instance, the C-α coordinates of the protein backbone. Therefore, the volume itself is the particular piece of experimental information we have to work with.

Having defined the informational core of BioImage as the multi-dimensional images it contains, it is also clear that a substantial amount of textual information must accompany these images in order for them to be understood within their correct biological and instrumentation context. Thus, each multi-dimensional image is accompanied by metadata specifying the biological organism being studied, together with experimental details about the sample preparation, the instrumentation setup and the image processing steps. A schema depicting in general terms the relation between the multi-dimensional images and their associated metadata is shown in Figure 2.

The image data stored in BioImage certainly have many interrelations with other pieces of biological information which are normally organised in other databases. Therefore, links with these other databases are also included. In general, bibliographic and taxonomic information is vital for all entries. In particular, links to PDB and Swiss-Prot (3,4) are essential for entries on the structure of macromolecules while, for example, links to tissue collections and whole organisms databases are key in many cell biology studies.

INTERNAL ORGANISATION OF THE DATABASE

The organisation of BioImage has been an international effort aimed at developing this 'missing' database of key biological information of a complex data type: multi-dimensional images. In this way the consortium developing BioImage represents a blend of scientific knowledge in different biological areas, plus technical expertise in the fields of information technology and of the visualisation and manipulation of multi-dimensional data.

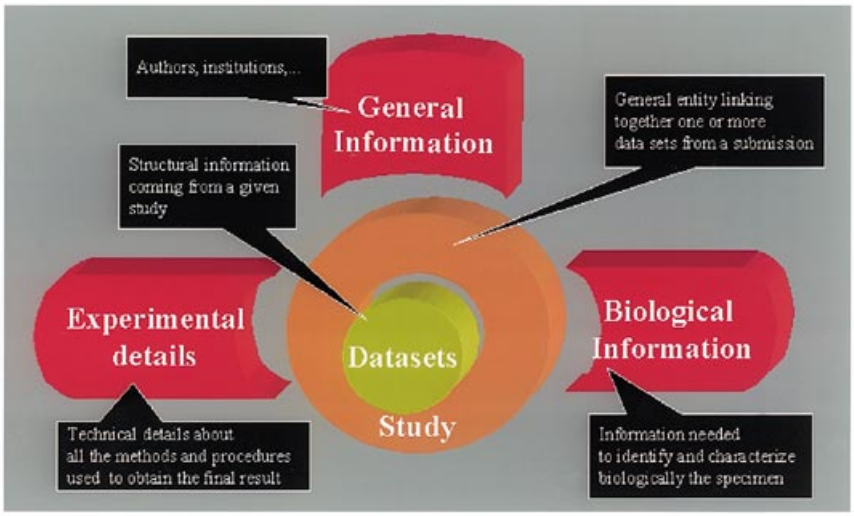

Figure 2. Schema of the information content of BioImage. The core of the database is formed by the multi-dimensional images organised in 'Studies'. Each study may have several datasets, each of them being a multi-dimensional image. Describing each dataset there is textual information (metadata) about the specimen being studied, as well as the experimental conditions in which the study was carried out.

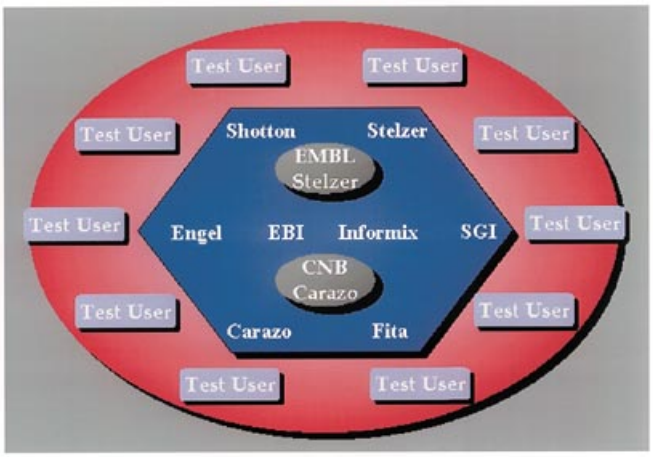

Figure 3. Schema of the internal organisation of BioImage. There are seven partners in the project, two of them providing the data servers. Around the partners there is a network of 'Test Users' (these Test Users are or have been the laboratories of Drs J. Bereiter-Hahn, A. Brisson, S. Fuller, H. Gross, R. Hegerl, M. J. Miles, P. Shaw, H. J. Tanke, J. M. Valpuesta, S. Vilaro and R. Wade).

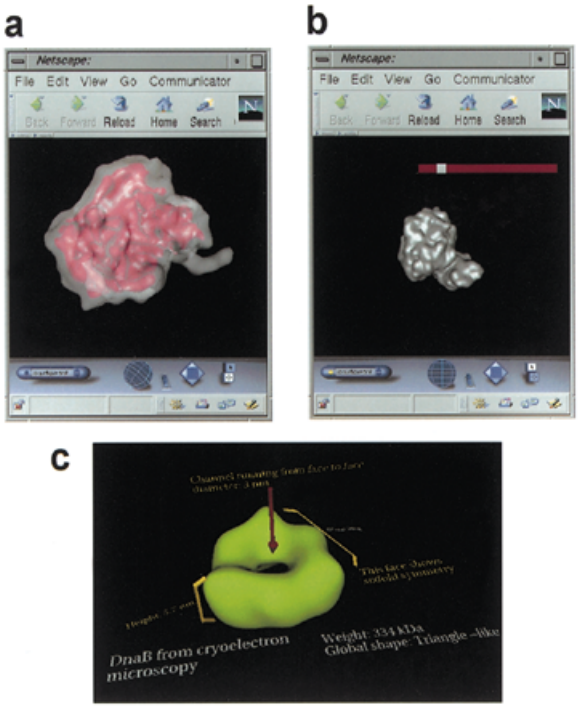

Figure 4. Data manipulation tools to interactively study multi-dimensional data over the Web. (**a**) Interface for volume visualisation. (**b**) Interface for automatic isosurface generation by thresholding. (**c**) Example of the volume annotation capabilities under development.

The image information in BioImage is quite homogeneous from the viewpoint of data types. However, it clearly addresses several distinct levels of biological organisation, ranging from structural biology of macromolecules to lower resolution 3D images and videos often found in cell biological studies. To permit access to all types of multi-dimensional image information in a homogeneous way, while paying attention to the specific needs of the different levels of cellular organisation under consideration, BioImage has been set up using two database servers specialising in two broad areas of biological interest: structural biology of macromolecules and cell biology. A diagram of the internal BioImage organisation is shown in Figure 3. The two database servers have the same data model, as well as common interfaces for query and visualisation. However, the actual multi-dimensional image data are kept only on the database server to which they have been submitted. Also, the submission interfaces and the curation processes of the two servers are more oriented towards a given community or other.

The data server at Madrid specializes in structural biology of macromolecules, while the one at the EMBL-Heidelberg addresses the cell biology field. The laboratories at Madrid, Barcelona and Basel provide scientific expertise in macromolecular structure, while those at the EMBL and Oxford are recognised in cell biology. The European BioInformatics Institute at Cambridge provides their applied expertise in running sequence and structure databases. Silicon Graphics is an industrial partner of the project, providing key technology for multi-dimensional image display and manipulation. The second industrial partner, Informix, provides the needed experience in information technology in the emerging area of large distributed databases of complex data types.

Another key feature that is proving most valuable in the development of BioImage is the organisation of a network of research laboratories around the project chosen in such a way to cover a wide spectrum of potential users of the database. They are referred to as 'Test Users'. Their input helps the BioImage developers to remain close to the needs of the scientific community for which this database is being developed, as well as to increase the visibility of the project in that moment in which we initiate an outreach program and involve the wider scientific community in the use of BioImage (this Test Users network is also shown in Figure 3).

ELECTRONIC ACCESS TO BIOIMAGE

From the very beginning, BioImage has been designed as an electronically accessible database, with all its interfaces developed for Web-based usage. This design starting point follows modern technological trends. However, due to the complex nature of the multi-dimensional images, this approach has demanded some special new technical developments that will be discussed in the next section.

During the development stage of the database, access will be granted upon request. Refer to the BioImage Web Sites for further details.

MULTI-DIMENSIONAL IMAGE VISUALIZATION AND MANIPULATION OVER THE INTERNET

The core of BioImage has been defined as multi-dimensional images. It then follows that the tasks of visualising and manipulating such complex information over the Web have to be addressed from the onset. To start with, it is a fact that there exists no common image format across the range of techniques embraced by BioImage. Therefore, one of the first tasks was to extend basic I/O system libraries in such a way that they would be able to handle the different formats in common usage in the

different research fields. This goal has been accomplished by extending the Image Format Library (IFL) of Silicon Graphics in such a way that new formats such as the so-called MRC, CCP4, Spider and Imagic formats are now recognised, in addition to the industry standards already recognised by the original IFL. This shared-object approach has the distinct advantage that all applications using IFL can now handle the different formats in a manner transparent to the application programmer, and that further new formats can be added to the basic IFL as needed.

The next task was then the development of specific Web browser plug-ins that allowed the manipulation of multi-dimensional images within the Web. Volume information that matches a given (textual) query can now be interactively studied within a VRML environment in such a way that volume rendering as well as isosurfaces derivation and surface rendering can be performed interactively. Tools are in development to allow the data submitter to 'annotate' the input data in such a way that specific biological information, such as point symmetries, specific contact surfaces and so on, can be incorporated at the time of data submission (Fig. 4).

Interesting volumes can then be retrieved from the servers and be used for the specific purpose of the user who has performed the query to BioImage. This software, developed within the BioImage consortium, may be accessed via the home page of the project. More complex queries, in which the actual structural content of the volumes can be addressed, in addition to their accompanying textual information, are starting to be implemented.

PRESENT SITUATION AND PLANS FOR DATA POPULATION

The first implementation of BioImage has just been released. The database is implemented using the Informix Universal Server running on Silicon Graphics servers at Madrid and Heildelberg. The query and visualisation interfaces have been developed using the Informix Web Datablade. To enhance the visualisation interface, the image handling tools developed by Silicon Graphics and described above have been incorporated. The Web-based submission interface has gone through several cycles of refinements with the help of the Test Users.

During the initial development process, the number of multi-dimensional images stored in the database has intentionally been kept small. The logical steps for data population have involved first the data supplied by the partner laboratories, then that of the Test Users (in order to test the complete implementation), and finally the opening of the database to a wider participation which we now invite.

BioImage is now interested in collecting high quality multi-dimensional microscopic images of biological specimens, and their associated metadata. Thus, all producers of these types of data are encouraged to contact either Madrid (for structural data on macromolecules acquired by electron microscopy or scanning probe microscopy) or Heidelberg (for multi-dimensional cell biological images, particularly those obtained by all kinds of light microscopy). At the same time, and in order to assure a fast population in a number of key fields, three areas of special attention have been defined in which active data collection schema are being followed. They are viral structures (coordinated by Stephen Fuller and with I. Fita as contact point: ifrcri@cid.csic.es), membrane proteins (coordinated by Andreas Engel and with B. Heyman as contact point: heymann@ubaclu.unibas.ch), and the cytoskeleton (coordinated by Ernst Stelzer and with E. Lindek as contact point: bioimage@embl-heidelberg.de). As a rule, multi-dimensional image data being submitted to BioImage should be supported by a peer-reviewed paper describing the work, which is then taken as the 'scientific validity check'. In future, we hope to move to a situation in which submission of such image data to BioImage will automatically accompany publication of scientific papers, as is presently the norm for protein and nucleic acid sequences and macromolecular coordinates to their respective databases.

The final goal of BioImage is to help to make available such multi-dimensional image information to the scientific community rapidly and in a user-friendly manner. However, it is acknowledged that the scientific community may develop specific policies about confidentiality of data, as it is the case for other types of information. Therefore BioImage incorporates the capability to set embargos upon the publication of specific submissions.

CONCLUSIONS

BioImage is already a reality. Multi-dimensional images can now be organised and accessible in this new database. So far, most of the effort of the BioImage consortium has gone into the design and implementation of the database, leading to the present prototype. The database is now open to the interested scientific community, and its general data population on a wider scale has already started. Information about BioImage can be found at the two sister home pages at http://www.bioimage.org (Madrid) and http://www-embl.bioimage.org (Heidelberg). A discussion forum has been established at bioimage@listserv.cnb.uam.es, in which all comments and opinions about BioImage are most welcomed.

ACKNOWLEDGEMENTS

The European Union through grant PL960472 has mainly financed work on this project. Special actions or extensions are being made possible in some partners laboratories by core national funding (J.M.C. acknowledges grants BI097-1485-CE and BIO98-0761). Initial support from the European Molecular Biology Network, as well as the many vivid discussions and proof-of-concept work in the field of macromolecular structure performed in collaboration with Dr Joachim Frank (Albany, US) is acknowledged by J.M.C. J.M.C. and J.F. would also thank the participants of the 1993, 1995 and 1997 Gordon Conferences in Three-dimensional Electron Microscopy for their support and ideas during all these years.

REFERENCES

1 Marabini,R., Vaquerizo,C., Fernandez,J.J., Carazo,J.M., Engel,A., Frank,J. (1996) *J. Struct. Biol.*, **116**, 161–167.
2 San Martin,C., Radermacher,M., Wolpensinger,B., Engel,A., Miles,C.S., Dixon,N.E., Carazo,J.M. (1998) *Structure*, **6**, 501–509.
3 Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
4 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.