

---

Universidad Autónoma de Madrid  
Escuela Politécnica Superior  
Departamento de Ingeniería Informática

Integration of biological data: systems,  
infrastructures and programmable tools

por

Mónica Chagoyen Quiles

. Tesis propuesta para el doctorado en

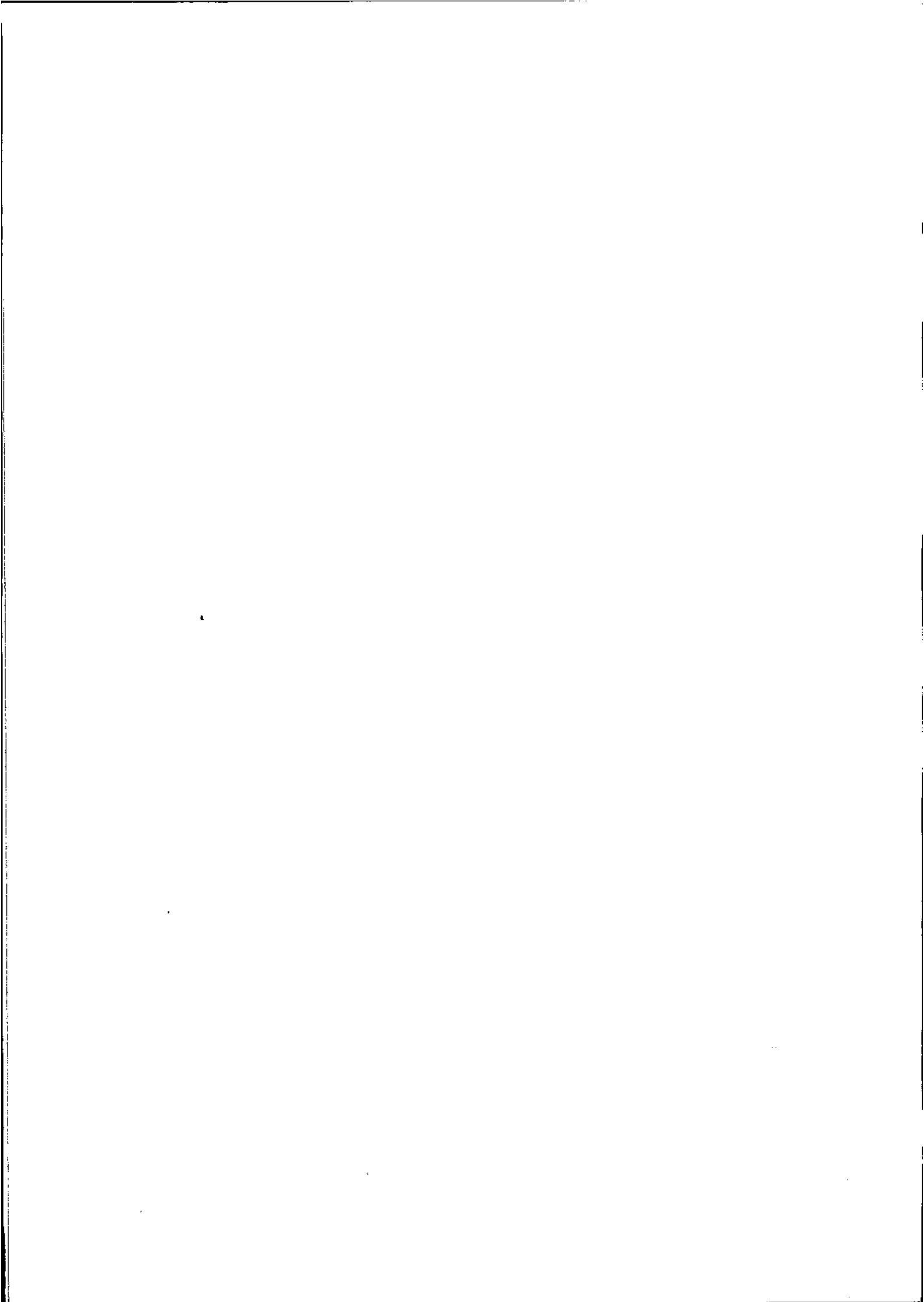
Ingeniería Informática y de Telecomunicación

Madrid, 2005

Director: José M. Carazo García



UNIVERSIDAD AUTÓNOMA MADRID REGISTRO GENERAL
Entrada 01 Nº. 200500004460 11/04/05 15:35:16



R.10817  
①

Tesis

J-34

As an agent of change, printing altered methods of data collection, storage and retrieval systems and communications networks used by learned communities throughout Europe.

Elizabeth L. Eisenstein. *"The printing press as an agent of change"*

REF-DOU-178



## Agradecimientos

En primer lugar mi reconocimiento y gratitud a José María Carazo, impulsor de mi trabajo durante estos años y quién me convenció finalmente a embarcarme en el doctorado. A Roberto Marabini, que me ha acompañado en la escuela.

Gracias también a todas aquellas personas con las que he trabajado a lo largo de los distintos proyectos, y con las que he tenido la oportunidad de aprender y compartir muchas horas de trabajo: a Roberto, Carmen, Montse, José y Luis Gerardo por enseñarme el día a día de la microscopía electrónica; a Pedro A., Susana y José Luis, compañeros de los tiempos de BioImage; a Phil, Kim, Richard, Peter, John y Mo del EBI, que me han guiado por el mundo de las bases de datos biológicas; a Amarnath, Erdem y Pedro A. (¡otra vez!), por perseverar en la difícil tarea de la integración; a Bernard y David, por su loable esfuerzo de estandarización.

A mis actuales compañeros de fatigas, Alberto y Pedro. ¡Espero que nos quede mucho por recorrer juntos!

Gracias a mis compañeros del CNB, que han hecho y siguen haciendo el trabajo mucho más agradable: Patrick, Sonia, Carlos Óscar, Yolanda, Diego, Ernesto, C-Manzana, María G., Luis Enrique, Rafa, Yacob, Mikel, Eva, Blanca, Julia, Esther, Carlos y los "niños" del museo: Jesús, Javi, Ángel, Ignacio, Román y especialmente Sjors que ha revisado pacientemente la memoria.

A Natalia y María, que traspasaron la barrera de buenas compañeras. Gracias por hacerlo:

Gracias también a los compañeros que hicieron mis estancias en el extranjero mucho más llevaderas: Jorge, Brian, Eduard y Simone.

A los "partners" y "test users" de los proyectos, a la gente de la Unidad de Microscopía del CNB que me descubrió todo un "nuevo mundo" en sus seminarios.

Finalmente, gracias a los que habéis aportado indirectamente a este trabajo: los que me acompañáis y enseñáis lo más importante.

*This work has been partially supported by: European Union (project grants BIO4-CT96-0472, QLRT-CT-2000-31237 and QLRI-CT-2001-00015) and CICYT (project grants BIO95-0768, BIO 98-0761, BIO 2001-1237).*



---

## **Integración de datos biológicos: sistemas, infraestructuras y herramientas programáticas**

### **Presentación**

El desarrollo de técnicas experimentales de alto rendimiento (como la secuenciación y los microchips de ADN), junto con el consiguiente desarrollo de la bioinformática y la biología computacional y la acumulación de gran cantidad de datos e información han convertido a la Biología Molecular en una ciencia dependiente en gran medida de las Tecnologías de la Información. Este torrente de información hace necesario, casi inevitable, automatizar el análisis integrado de los nuevos datos experimentales disponibles.

La integración de información biológica tiene diversas caras y, por tanto, los diferentes enfoques y soluciones existentes se revisan como una introducción a la materia. La segunda parte de la tesis presenta los pasos dados para solventar la carencia de infraestructuras para la gestión y almacenamiento de datos estructurales obtenidos mediante microscopía electrónica tridimensional, articulados alrededor de dos proyectos científicos de ámbito internacional: una primera conceptualización en la base de datos BioImage (integrando información de distintas técnicas microscópicas) y la creación de EMD (Electron Microscopy Database) en el European Bioinformatics Institute (integrando información de estructuras macromoleculares).

Finalmente se presenta el trabajo realizado (en colaboración con el San Diego Supercomputer Center) para proporcionar herramientas que respondan a las necesidades de análisis de un laboratorio experimental y/o computacional. Los procesos de análisis se modelan mediante flujos de trabajo (*workflows*), que intercalan accesos a fuentes de información (bases de datos estructuradas, ficheros locales, sitios Web) y ejecución de algoritmos y/o aplicaciones.

---

**Abstract**

The development of high-throughput experimental techniques (such as DNA sequencing technologies and DNA microarrays), together with the subsequent advance of bioinformatics and computational biology disciplines and the accumulation of a great amount of data and information, have made Molecular Biology a science heavily dependent on Information Technologies. This flood of information makes it necessary, almost unavoidable, to automate the integrative analysis of new experimental data.

Integration of biological information has several approaches and already existing solutions which are reviewed as an introduction to the subject. The second part of this thesis presents the works done in order to provide infrastructures for the management and archiving of structural data obtained by three-dimensional electron microscopy. These works are part of two international scientific research projects: a first conceptualization in the BioImage database (integrating information obtained by various microscopy techniques) and the creation of the EMD (Electron Microscopy Database) at the European Bioinformatics Institute (integrating information on macromolecular structures).

Finally, the development, together with the San Diego Supercomputer Center, of novel tools that respond to the analysis performed in an experimental and/or computational laboratory is presented. The analytical processes are modeled by computational workflows, which intersperse data source access (databases, data files, Web sites) and algorithm and/or application execution.



## Table of contents

---

Table of contents.....	v
Figures and tables .....	ix
List of figures.....	ix
List of tables.....	x
Glossary of acronyms .....	xi
<i>Chapter 1</i> Introduction .....	1
1.1. Scientific data management .....	1
1.2. OAIS Reference Model.....	4
1.3. Biological data.....	5
1.4. Molecular biology databases .....	6
1.4.1. Service organizations .....	6
1.4.2. Data quality .....	6



1.4.3.	Data types .....	6
1.4.4.	Data operations.....	6
1.5.	Biological data integration .....	6
1.6.	Contents.....	6
<i>Chapter 2</i>	<i>The path towards integration.....</i>	<i>6</i>
2.1.	The basics: establish means of data citation.....	6
2.2.	Integration approaches .....	6
2.2.1.	Use of software standards .....	6
2.2.2.	Developing standards .....	6
2.2.3.	Bridging the semantic gap.....	6
2.2.4.	Making use of the data .....	6
2.3.	Integrated systems.....	6
2.4.	Discussion .....	6
2.4.1.	Choosing the right approach .....	6
<i>Chapter 3</i>	<i>Organising biological multidimensional image data.....</i>	<i>6</i>
3.1.	Introduction .....	6
3.2.	Objective .....	6
3.3.	Methods.....	6
3.3.1.	Architecture.....	6
3.3.2.	Phases and methodology .....	6
3.4.	Results .....	6
3.4.1.	Information model.....	6
3.4.2.	Database model .....	6
3.4.3.	Functional model.....	6
3.4.4.	Combined studies .....	6

3.5. Discussion .....	6
<i>Chapter 4</i> New infrastructures for structural data.....	6
4.1. Introduction .....	6
4.1.1. 3D-EM data .....	6
4.1.2. Organising the information .....	6
4.2. Objectives.....	6
4.3. Methods.....	6
4.3.1. Establishing relationships with atomic models .....	6
4.3.2. Architecture.....	6
4.3.3. Phases and methodology .....	6
4.4. Results .....	6
4.4.1. Electron Microscopy Database (EMD) .....	6
4.4.2. Additional results .....	6
4.5. Discussion .....	6
<i>Chapter 5</i> Developing tools for biological data integration.....	6
5.1. Introduction .....	6
5.1.1. Computational biology workflows.....	6
5.1.2. Scientific workflow framework .....	6
5.2. Objectives.....	6
5.3. Methods.....	6
5.4. Results .....	6
5.4.1. Relationship operations in computational biology.....	6
5.4.2. PLAN: a technology for integrative analysis .....	6
5.4.3. Solving computational workflows with PLAN.....	6
5.5. Ready for semantic integration?.....	6

5.6. Discussion .....6

Discussion.....6

Conclusions.....6

Conclusiones.....6

Bibliography .....6

Biographical note.....6

Academic and research record.....6

Relevant publications.....6

---

## Figures and tables

---

### List of figures

Figure 1: Biological data analysis is often done in terms of comparative analysis. a) A set of protein sequences is analyzed to discover conserved patterns (e.g. by multiple sequence alignment).....6

Figure 2: Links between Swiss-Prot and PDB using annotated cross-references.....6

Figure 3: The multidimensional data in the BioImage database are generated by various microscopy techniques.....6

Figure 4: Simplified view of the structure of the BioImage data model showing its main entities and relationships (adapted from [31]). .....6

Figure 5: Fitting study of the FMDV-Fab complex.....6

Figure 6: Scope of 3D-EM data. Conceptual representation of the biological and structural scope of 3D Electron Microscopy data, and the relationships of structural information provided. ....6



Figure 7: Some of the entities of the integrated entity-relationship model of the MSD are shown. ....	6
Figure 8: Top elements in the hierarchy of an EMD XML file as defined in the EMD Schema. ....	6
Figure 9: Relationships built among Swiss-Prot, InterPro, ENZYME DB, UniGene, PDB, PQS, CATH and PubMed data sources. ....	6
Figure 10: PLAN System Architecture (adapted from [114]). Oval shapes are processing units, while rectangles contain data structures. ....	6
Figure 11: Illustration of some of the domain signatures (syntax types) corresponding to the InterPro 'Heat shock protein DnaJ, N-terminal' domain. ....	6

## List of tables

Table 1: Overview of approaches and technologies for integration according to the four focuses of sharing data, methods, processes and interfaces. ....	6
Table 2: Some example integrated systems in molecular biology. ....	6
Table 3: Strengths and weaknesses of different integration approaches according to several dimensions (adapted from [27]) ....	6

## **Glossary of acronyms**

---

3D-EM: Three-dimensional electron microscopy

AFM: Atomic Force Microscopy

API: Application programming interface

BLAST: Basic Local Alignment Search Tool

CCP4: Collaboratory Computing Project Number 4

CNB: Centro Nacional de Biotecnología

CORBA: Common Object Request Broker Architecture

DB: Database

DBMS: Data base management system

DDBJ: DNA Data Bank of Japan

DNA: Deoxyribonucleic acid

DTD: Document Type Definition

**EBI:** European Bioinformatics Institute

**EMBL:** European Molecular Biology Laboratory. Sometimes also used to designate EMBL data bank.

**EMD:** Electron Microscopy Database (also known as Electron Microscopy Data Bank, EMDB)

**EST:** Expressed sequence tag

**HUGO:** Human Genome Organisation

**ICTV:** International Committee on Taxonomy of Viruses

**IEEE:** Institute of Electrical and Electronics Engineers, Inc.

**ISO:** International Organization for Standardization

**JDBC:** Java database connectivity

**MSD:** Macromolecular Structure Database

**NCBI:** National Center for Biotechnology Information

**NMR:** Nuclear Magnetic Resonance

**NSF:** National Science Foundation

**OAIS:** Open archival information system

**ODBC:** Open Database Connectivity

**OMG:** Object Management Group

**OODB:** Object-oriented data database

**ORDB:** Object-relational data base

**PDB:** Protein Data Bank

**PSI:** Proteomics Standards Initiative

**RCSB:** Research Collaboratory for Structural Bioinformatics

**RDB:** Relational base

**RNA:** ribonucleic acid

**SOAP:** Simple Object Access Protocol



SQL: Structured Query Language

UML: Unified Modeling Language

URL: Uniform Resource Locator

WAP: Wireless Application Protocol

WWW: World-Wide Web

XML: Extensible Markup Language

XMLDB: XML database





---

*Chapter 1*    **Introduction**

---

**1.1. Scientific data management**

Nowadays, there is a growing number of scientific and technical data that are regularly deposited and stored in electronic databases. Furthermore, these databases are, in a way or another, publicly available online. Most of the first electronic databases, dating from the 1960's, contained bibliographic references to published literature [1]. Following decades gave as results a richer diversity and complexity of their contents. Scientific databases are nowadays essential to the progress of science as they provide means for data sharing and long-term preservation of data to enable further analysis.

Among scientific disciplines, molecular biology is perhaps one of the pioneers where traditional publishing has naturally merged with the use of electronic databases. Many key scientific journals in the field only admit papers reporting accession numbers to corresponding public databases. This means that relevant data should have been submitted by the authors to the appropriate database prior to publication.

The accumulation of experimental data and information, and thus knowledge, in public repositories, has enabled the development of the computational biology discipline which, at the same time, generates new data to the scientific community.

Development of world-wide infrastructures to support the storage of scientific data involves the creation and use of software technologies, together with the understanding, conceptualization and organization of complex data. More specifically, it entails the collaborative work with a given scientific community which might be reluctant to deposit data in public archives. There is no doubt of the needs of data sharing for the advance of science. The scientific community, data requesters and society are, to a large degree the beneficiaries of data sharing, while primary researches and research participants may also realize gains in some circumstances [2]. As noted by Sterling and Weinkam, [3] *“the willingness of one scientist to share data with another continues to be influenced by a number of economic, social, psychological and political factors. While it has always been possible for consenting scientists to collaborate and work with and on each others’ data, actually to do so was extremely laborious. As a further consequence of the automation of scientific data processing, scientists working with computers are forced to store and document data in an extremely precise fashion. Each data item placed on machine-readable media must be clearly defined and recorded according to precise protocol in order to take advantage of data-processing techniques. [...] While conditions exists that encourage data sharing, very often forces appear to be at work that oppose the sharing of data”*.

In order to understand the demands in scientific environments for data organization and management, it is important to analyse some characteristics of scientific datasets and databases. Participants of the National Science Foundation (NSF) Workshop on Scientific Data Management [4] suggest a taxonomy of databases in terms of three basic dimensions:

- **Level of interpretation:** A dataset can consist of a collection of raw data or, to the opposite extreme, a set of highly processed interpretations. The contents of scientific databases can therefore be categorized according to whether they are conceived as recording either facts about the real world or elaborated scientific conclusions.
- **Intended scientific analysis:** It is assumed that all scientific data sets are subject to further analysis; otherwise there is little reason to retain them. The nature of such subsequent analysis frequently determines what particular representational format is most desirable.

- Data source: A broad spectrum of possibilities is found, from single-source data archive to a multi-source collection of data.

Attending to the level of interpretation scientific data can be classified as:

- Raw/sensor data: obtained directly from the measurement device. It is seldom saved.
- Calibrated data: raw physical values corrected with calibration operators. It is normally preserved.
- Validated data: filtered through quality assurance procedures. It is most commonly used for scientific purposes.
- Derived data, frequently in the form of aggregated data, such as gridded or averaged data.
- Interpreted data or derived data that is related to other data sets, or to the literature of the field.

As clearly stated in [4], many issues regarding scientific databases are similar to those found in conventional business environments, but the focus is different. Scientific data can be characterized by large volume, low update frequency, and indefinite retention. Thus, while efficient transaction processing and concurrency control are critical to business operational databases, it is not a major issue for scientific archives. However, flexible and efficient query processing is essential for most scientific applications.

In general, scientific databases should be defined having in mind a more diverse user community than those for a typical business database. There are a number of important issues that should be also taken into account when creating a scientific data base:

- Data annotations: for the data to be meaningfully processed later, annotations associated with the data must be preserved and accessible.
- Standards: although heterogeneity in data and operational environments is a fact of life, it is important to promote data consistency within and across scientific disciplines.
- Appropriate analysis operators: there is a lack of appropriate operators within existing Data Base Management Systems (DBMS) for manipulating the kinds of data

encountered in scientific applications.

- **Data citation:** a standard citation mechanism would allow other researchers to locate and examine precisely the data used in the investigation. It would also give due credit to the data collectors, as well as means for further reference.

In addition to the above, there are some collateral issues that might be of interest to anybody involved in the development and management of a scientific data infrastructure: intellectual property rights and database protection [5, 6]; as well as the financing of public databases [7, 8].

## 1.2. OAIS Reference Model

The importance of the reference model as an aid to human understanding and communication in the early stage of any software project is widely recognized in the software engineering discipline. The OAIS (Open Archival Information System) Reference Model [9] has been recently adopted by the ISO Archiving Standards (ISO 14721:2003 Space data and information transfer systems – Open archival information system – Reference model; Technical committee /subcommittee TC 20/SC 13).

In this context, an Archival Information System is defined as the hardware, software and people who are responsible for the acquisition, preservation and dissemination of the information to a 'Designated community'. This reference model provides a framework for understanding and applying concepts needed for long-term digital information preservation. It does not specify an implementation and, although it was originally created in the context of space data systems, may be applicable to any archive.

Among other aspects, OAIS addresses a full range of archival information preservation functions including ingest, archival storage, data management, access and dissemination. In order to provide a coherent framework, OAIS concepts will be used throughout this work.

### 1.3. Biological data

In order to develop useful systems for data management and analysis it is essential to understand the characteristics of underlying data, as well as the context in which they are used. Although a software engineer does not need to become an expert in the field, he/she should learn the main particularities of the subject at hand that make a given project “unique” from the general case. As noted in [10] *“Biology is in the middle of a major paradigm shift, driven by computing. Although it is already an informational science in many respects, the field is rapidly becoming much more computational and analytical. [...] Bridging the gap between the “real world” of biology and the precise “logical” nature of computers requires an interdisciplinary perspective”* (IEEE Computer Magazine, 1991).

Some authors stress the fact that the main issue in biological information is not the growing size of experimental data, but the complexity of the living systems [11]. Biological data are hierarchical and non-reducible by symmetry or by temporal considerations, showing a wide range of scales. Due to this complexity, the biological sciences make use of simulation and modelling, requiring the development of new computational and information tools to comprehend the observed data and understand how biological systems work.

Therefore, the following characteristics of biological data have to be considered when designing and developing software systems to support molecular biology research:

- Data are complex and heterogeneous (e.g. just at the molecular level there is a big diversity of data: from sequences, to protein structures and molecular pathways).
- The data present a layered organization (from individual molecules to whole populations).
- Experimental data are dynamic and incomplete (they provide a picture of the current technological and scientific state of the art in a particular timeframe).
- The data can provide information that is putative, probabilistic or verified (resulting from interpretations, computational predictions or experimentally confirmed).



- In many cases, the data have a qualitative rather than quantitative nature (they may result from the interpretation of experimental measurements, as well as from comparison of results obtained from different biological systems).

Finally, as any other scientific discipline, biology is continuously evolving and thus not capable of foreseeing what information will be useful in the future.

#### 1.4. Molecular biology databases

There is a huge number of infrastructures for the storage and management of biological data, either publicly available or commercial. These infrastructures correspond to three specific types of database applications [10]:

- Repository databases. They are created as public resources to contain data from many sources.
- Collaborative databases. They combine databases and data from several laboratories working together on a single problem.
- Laboratory databases. They are designed to support the work of a single laboratory.

The design and development of these databases present formidable challenges. For example, repository databases should take into account that changes become painful and expensive once the database is public. On the contrary, laboratory databases must be able to change database structures easily to accommodate constantly evolving data.

Frishman and co-workers [12] describe a wide spectrum of repository databases in molecular biology:

- General biological databanks: e.g. GenBank and EMBL (nucleotide sequences), Swiss-Prot (protein sequences), PDB (atomic structures).
- Species-specific full-genome databases: e.g. SGD (*Saccharomyces cerevisiae*), Flybase (fruit fly).
- Specialized in subject matter: e.g. TRANSFAC (transcription factors and binding sites), REBASE (restriction enzymes).



- Derived databases, containing added descriptive material by providing novel structuring of data based on global data analysis: e.g. PROSITE (protein motifs), SCOP (structural classification).

The data stored and managed in these collections can come from very different sources. Generally speaking, these can be categorised as data obtained directly from experimental techniques, computational analysis, the scientific literature and even other databases. While large primary databases collect and collate “atomic” information from the scientific community, specialized derived data collections integrate, via curatorial expertise, information from a multiplicity of primary sources.

There is an increasing number of data in the major public data repositories, due to the methodological advances for biological data collection and analysis (e.g. high-throughput methods such as DNA sequencing technologies and DNA microarrays). Additionally, molecular biology data is relevant to a wide spectrum of domains and applications: biotechnology, medicine, nanotechnology, etc.

#### 1.4.1. Service organizations

One of the requirements of scientific databases is that they should ensure long-term preservation of data so as to enable further analysis. In order to provide a stable infrastructure to support this archiving functionality, repository databases are usually managed by large service organizations (e.g. the National Center for Biotechnology Information (NCBI) in the USA).

In Europe, the largest biological database institution is the European Bioinformatics Institute (EBI) ([www.ebi.ac.uk](http://www.ebi.ac.uk)). The EBI is a non-profit research organisation that forms part of the European Molecular Biology Laboratory (EMBL). The EBI manages databases of biological data including nucleic acid, protein sequences and macromolecular structures. Its mission is to ensure that the growing body of information from molecular biology and genome research is placed in the public domain and is accessible freely to all facets of the scientific community in ways that promote scientific progress.



#### 1.4.2. Data quality

As in any other scientific application, one of the main concerns for molecular biological database users is data quality. Data quality has in fact several faces [12]: correctness, completeness and timeliness of capture, applied both to the newly measured properties (e.g. a new gene sequence) and their corresponding annotation (e.g. biochemical activity of gene product). Data quality concerns are therefore quite different in two extreme cases, i.e. passive data repositories (archives), or active reference compendia.

The quality of archived data can be no better than the data determined in the contributing laboratories [13]. Nevertheless, careful curation of the data can help to identify errors. The state of the experimental art is the most important determinant of data quality. Quality control procedures provide the second level of protection. Indices of quality, even if they do not permit error correction, can help scientists avoid basing conclusions on questionable data.

In the case of active reference compendia, data annotation is made either automatically or manually. Although manual annotation is time-consuming, it usually provides more accurate results. Nevertheless, some key issues pertaining to manual annotation should be kept in mind [14]:

- **Human factors.** The curators are a central element in the creation of good biological databases. Their effort is crucial for generating accurate data and deleting erroneous data. As in any expert-based analysis, the scalability is a major concern.
- **Difficult.** In addition to being tedious, manual annotation requires training and expertise for bioinformatics based analysis as well as retrieval and correct interpretation of articles from published literature. This is compounded by the fact that there is little recognition for an individual as a good annotator.
- **Error propagation.** Because a large number of biological annotations are based on previous data, errors can be easily propagated in all entries that inherit information from an incorrect entry. Accuracy of the data must, therefore, be checked at several stages.
- **Community effort.** Although manual annotation by experts at a central database works well, no biological repository can do a better job of annotating than

investigators who individually annotate or edit only a subset of entries in the entire database. Ultimately, the entire biological community should be involved in the upkeep of correct information in public databases.

### 1.4.3. Data types

The complex and heterogeneous nature of biological data demands the use of a great diversity of data types (both in permanent storage, as well as data structures in algorithm implementations). Some of them are not well supported by conventional DBMS [15]:

- Sequences (DNA, RNA, amino acid): often stored as text strings, this representation is awkward when using annotations on individual positions (nucleotides or amino acids). Often DNA sequences include not only individual nucleotides, but also gaps, usually with a length (or bounds on length) specification of the gap.
- Graphs: either as directed/undirected labelled graph, nested graph, or hypergraph (e.g. pathways, genetic maps, phylogenies, taxonomies, chemical structures, sequences, protocols and workflows).
- High-dimensional data: as spatial data (e.g. molecular structure data), scalar data (e.g. microarray expression data sets), vector fields (e.g. molecular and cell dynamics) and temporal data.
- Other complex data types: patterns (e.g. regular expressions, Hidden Markov Models and other types of grammars for sequence motif representation), mathematical and statistical models (e.g. cell simulations), constraints (e.g. energy conservation constraints in chemical reactions, torsion angles in macromolecular structures) and texts.

### 1.4.4. Data operations

Users demand particular needs in data access and database searching mechanisms, as biological knowledge is often gained by analogy. Most typical operations performed on biological data are comparison and combination of data from different methodologies and/or biological systems (see Figure 1). These include similarity queries (e.g. for an



introduction on sequence similarity see [16]), pattern matching and pattern discovery queries (see reviews on sequence patterns [17] and structural patterns [18]), spatial and temporal queries, and general computational queries (like those used in protein structure prediction [19], or specialized database searches for mass spectrometry proteomics [20]). Pair-wise comparison of analogous data translates naturally to content-based retrieval searches (using a term coined in multimedia databases) when comparison is performed against data stored in an archive collection.

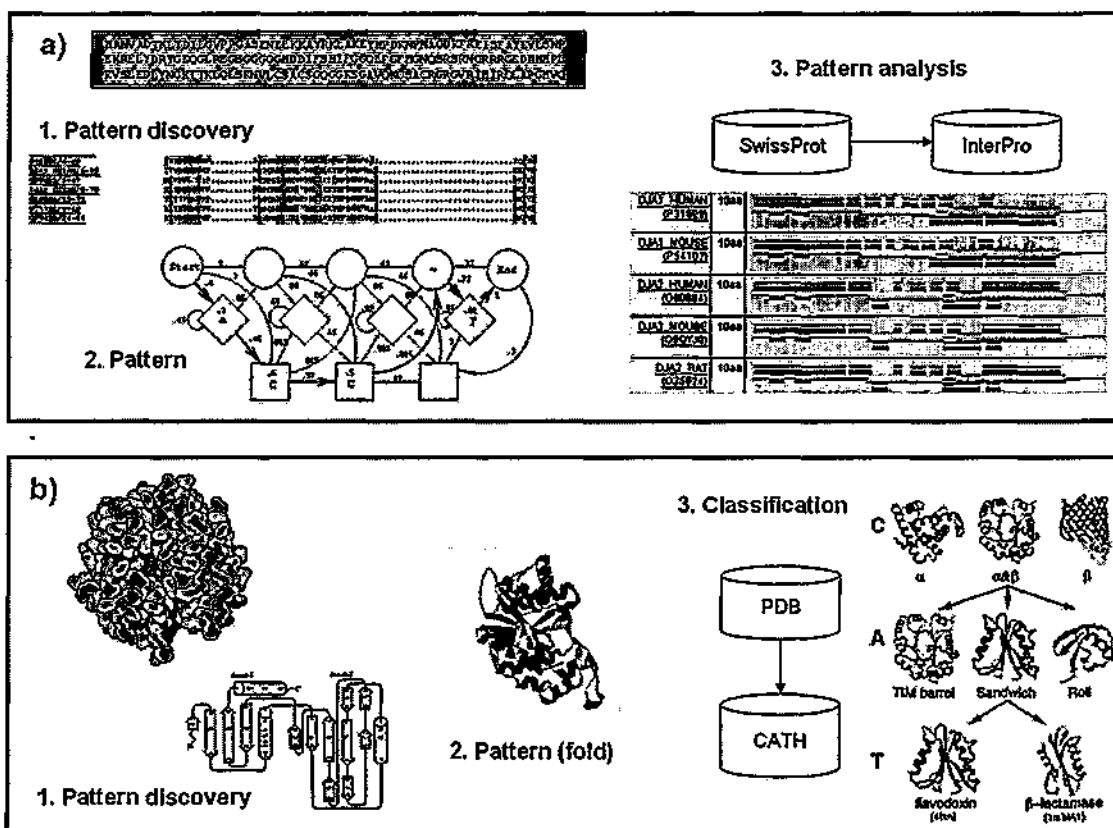


Figure 1: Biological data analysis is often done in terms of comparative analysis. a) A set of protein sequences is analyzed to discover conserved patterns (e.g. by multiple sequence alignment). These patterns are stored for further reference (a Hidden Markov Model is shown as a pattern). Whole database scale analysis can be performed and materialized as a new “derived” database (e.g. family/domain analysis of Swiss-Prot sequences is stored in the InterPro). b) Similar analytical pathway for three-dimensional protein structures.

The increasing awareness of the importance of incorporating common biological query mechanisms into commercial database management systems can be exemplified by the work reported in [21], where a sequence similarity search was implemented as part of the Oracle 8i extended data cartridge. Furthermore, Oracle has incorporated the BLAST sequence alignment algorithm [22] as part of its 10g release.

## 1.5. Biological data integration

The analysis of biological data cannot be performed in isolation. Almost any experimental data obtained should be interpreted and understood in the light of current data available. As an example, the discovery of underlying biological mechanisms that originate co-expression patterns in DNA microarray experiments requires a fairly good characterization of the genes analysed. Thus, integrative data analysis, or the study and interpretation of new experimental data in the context of available state of knowledge, is a must in any molecular biology project.

Difficulties found when integrating biological data are widely recognised and exemplified [23-27]:

1. Data/information is spread over numerous, distributed data sources<sup>1</sup>, both within individual organizations and/or laboratories and across the Internet. The first difficulty found is that biological data might not be accessible, that is, it is kept in laboratory notebooks and/or laboratory computers.
2. Data sources present heterogeneities at various levels:
  - a. System or platform level: these include differences in accessing mechanisms (application programming interfaces (API), protocols and user interfaces) and query/searching languages and capabilities.
  - b. Syntax level: heterogeneous data formats and/or schemas (one for each source) and data models (relational, object-relational, object-oriented, or XML data bases, flat files ...). Another concern at this level is heterogeneity of query languages.
  - c. Semantic level: data duplicated across multiple databases is represented differently in the underlying database schemas [28]. Biological data sources often differ in their representation of key concepts (e.g. for a

<sup>1</sup> In the context of the present work, the term 'data source' is used in a very broad sense: including databases, WWW sites, file collections, software applications, analytical instruments, etc...

given source a gene is an annotation on a sequence, while for another it is a locus which confers phenotype).

3. Limited expressive power of query interfaces to data collections (semi-structured models, web sites, output of analytical programs). Not all relevant web accessible databases provide means for posing *ad-hoc* declarative queries [29], but predefined form-based query interfaces.

4. Unclear and hidden semantics: semantics of sources are not easy to determine or reason about (e.g. incoherent terminology, multiple/informal taxonomies, implicit assumptions, etc.). Database schemas/models are not provided in most cases (reengineering efforts are needed in order to provide a model to be integrated).

5. Semantics at the data level: for some biological entities, there is no standard naming convention or nomenclature. In the absence of such shared terminology, how can 'correct' names be assigned and maintained across data sources?

6. Data dynamics: users might want to access the most up-to-date information. Update of sources occurs frequently (in most cases every day). Changes in underlying structure and syntax can also be expected. Finally, addition of new sources should also be considered.

7. Not all sources represent biological objects optimally for the kinds of queries and/or analysis that investigators typically want to pose (e.g. genes found as annotations to nucleic sequence data, pathways found as diagrams or figures, etc.). Thus, further transformations on data representations should be performed.

8. Data in an integrated system should be retraceable to its original location for further reference. Some authors refer to this matter as *data provenance*, defined as the process of tracing and recording the origins of data and its movement between databases [30].

9. Finally, the different data access policies for public repositories as well as commercial databases have to be taken into consideration when developing a practical implementation.

## 1.6. Contents

This thesis is organized in four chapters that contain an overall description of my work performed on integration of information in molecular biology. Due to the broad spectrum of molecular biology data and the need to define every detail relevant to a given scientific area of expertise, it is necessary to narrow the scope of the biological data integrated. In this line, this work has been centred in the area of structural biology, particularly around three-dimensional image data of biological macromolecules.

In Chapter 2 I present a bibliographic review of the main approaches and systems that enable biological information integration. Systems and technologies developed in a general software engineering environment are included if they have had an impact in the biological domain, together with their corresponding implementations.

Chapter 3 describes my contribution to the design and development of a database to support the organization of multidimensional image data of biological specimens obtained from various microscopy techniques [31]. This database was created together with the European Molecular Biology Laboratory in the context of an international collaborative project [32] which was coordinated by our laboratory in the Centro Nacional de Biotecnología.

Chapter 4 describes the continuation of the work presented in chapter 3 in order to provide public infrastructures capable of integrating data on macromolecular structures. It is focused on the organization of structural data obtained by three-dimensional electron microscopy (3D-EM), and its relationships with atomic coordinate data. In this occasion, I have worked in very close collaboration with the Macromolecular Structure Database group at the European Bioinformatics Institute, which also acts as database service provider.

Finally, chapter 5 contains the work carried out, together with the San Diego Supercomputer Center, towards the development of a programmable integrator suitable for the definition and execution of computational biology workflows.

Except chapter 2, written in the style of a review, the rest of the chapters have been conceived as proper scientific articles, including their own introduction, as well as

methodology, results and discussion sections. An overall discussion and a conclusion section close this thesis.

The inherent collaborative nature of the research projects I have been involved in makes it sometimes difficult to isolate or highlight my role in the overall work performed. This thesis as a whole may provide a better picture of my contributions to the subject of biological data integration. I hope this is interpreted as a sign of a joint and interdisciplinary effort.



## Chapter 2    **The path towards integration**

---

Activities and methodologies for biological data integration can be analysed and undertaken very differently depending on two broad perspectives. In the first one, the integration is a goal *per se* (e.g. building a consolidated resource for protein sequence family/domain information, which resulted in the creation of the InterPro database [33]). In the second, integration is not a goal but a need in order to answer a given question or perform a certain analysis (e.g. finding human ESTs sequences that may correspond to interesting neurological targets [24], assessing the conservation of protein-protein interfaces [34]).

In both situations there is a need of linking and relating diverse pieces of data and information through the establishment of a number of relationships. The task of discovering, building, representing these associations computationally and finally making efficient use of them, is not straightforward.

In this chapter I review and present a picture of the main approaches and developments that facilitate biological data integration, with an emphasis on the impact of software technologies and methodologies in the biological domain.

## 2.1. The basics: establish means of data citation

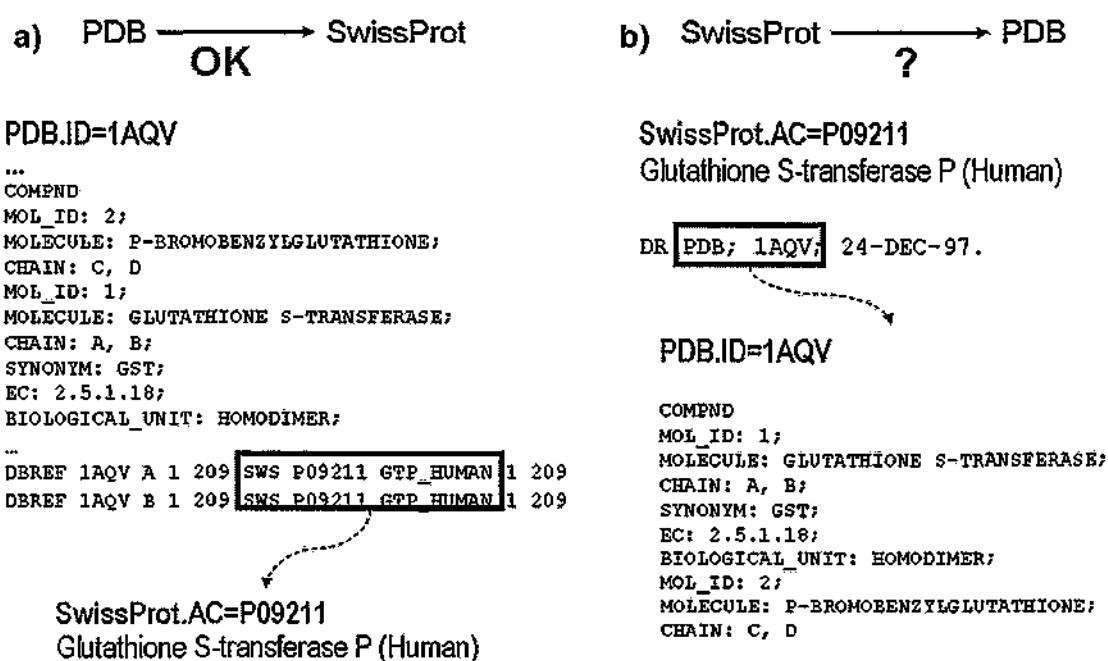
Citation of data stored in public databases is normally done by providing the database name and accession code corresponding to a given database 'entry'. This accession code acts as an external, unique identifier that is kept or maintained consistently during the lifetime of an archive or database. Data integration in molecular biology has benefited from the effort done by data providers towards preserving these accession codes through the lifetime of the data collection.

But what exactly constitutes a database 'entry'? It depends on the database and its contents. Databases containing experimental datasets usually provide an accession code or external identifier referring to a unique submission or deposition (e.g. the Protein Data Bank). In this case, an accession code will be related to an experiment but not to a given biological object. In other cases (e.g. Swiss-Prot database), attempts are made in order to group experimental evidences for a given biological object into the same database entry. The definition of the biological object in terms of database accession codes is therefore needed.

Early attempts towards data integration in molecular biology were based on these cross-database references or database links [35] built upon unique identifiers. A link is defined as a pointer from an entity in one database (the source) to an entity in a second database (the target) through the annotation of the target unique identifier in the source entity. By this definition, all links are unidirectional, meaning that two links are required to encode a bi-directional relationship.

Bi-directional links need to be consistent, available, well-documented, and maintained [36]. However, consistency is difficult to enforce over independent databases and many publicly available links suffer from a lack of characterization. Many of these cross-references are annotated at the time of data insertion, and several are not updated. Attempts to provide automatic solutions for the construction of links have been made (e.g. [37] describe genXref to discover links between the Genome Database (GDB) and GenBank). Exceptions are found in highly curated databases, but even in these cases, the flood of new data doesn't allow the proper timing for annotation of the most up-to-date information.

Additionally, some cross-references are not created with the appropriate level of granularity, i.e. they not represent faithfully the biological entity on which the relationship or connection is built (see Figure 2). Thus integration of information relying solely on unique identifiers and cross-database links works only if dealing with naturally related databases, such in the case of derived databases (e.g. InterPro and Swiss-Prot/TrEMBL databases).



**Figure 2:** Links between Swiss-Prot and PDB using annotated cross-references. a) PDB correctly points to Swiss-Prot sequence data at the level of chain (A, B). b) Swiss-Prot sequence is ambiguously annotated referring to PDB entry (1AQV).

In spite of the limited usefulness of unique identifiers for the purpose of distributed data integration, they provide a very good solution for data citation. As described in [23], there are two main lines of thought among groups that are interested in establishing global unique identifiers:

- Object identifiers should point to the biological objects themselves and use a URL syntax, therefore coupling the identity of a biological object with the location of its representation on the WWW.
- The identifiers should decouple the notion of the location of a resource from its authoritative source. E.g. the Life Sciences Identifier (LSID) proposal from the I3C (Interoperable Informatics Infrastructure Consortium).

The difficulty in adopting common schemes for global unique identifiers for biological data is that the same solution cannot be adopted for, in fact, two different levels: the *scientific object* (the truth or model) and the *experimental evidence* (or measurement) as stored and represented as a *digital object* in a given collection or archive.

An alternative is to provide parallel mechanisms of object identification (OID), corresponding to the object-oriented concepts of classes (biological entities) and instances (experimental evidences). While the scheme for OID of biological entities should not be tied to any particular database (apart from the archive of the corresponding authoritative committee), the OID of experimental evidences is necessarily tied to a given data collection.

Perhaps, the major contribution of the post-genomic biology in terms of biological data integration comes from the general recognition of the main biological molecules (DNA, RNA and proteins) as chemical entities uniquely defined by the exact composition of their sequence in terms of nucleotides or amino acids. Biological sequences are said to be as close to factual data as a major archival database in biology is likely to find [38]. A biological sequence can be defined as a string of chemical residues from a specified alphabet, which can be put in the context of a coordinate system. Relationships among sequences can be expressed as correlated set of location on a given coordinate system (e.g. a chromosome), or an alignment.

But, there are biological entities and data that cannot be directly attached to a known biological sequence. Some examples include complex molecular machines or organelles whose exact composition in terms of biochemical entities is not known. Therefore, the data obtained from these complexes and organelles (e.g. structural data from electron microscopy) cannot be referenced by the use of a set of biological sequences.

Nevertheless, integrative data analysis demands more complex approaches than just cross-referencing data items in distributed databases. The aim is to provide tools that enable information and method sharing in a heterogeneous and distributed environment.

## 2.2. Integration approaches

The need of information and application integration is not only found in molecular biology and biomedicine but spans many other scientific and business domains (manufacturing, electronic commerce, banking, etc.). Consequently it has been addressed from different perspectives, both in the computer science research and practical software business approaches. The main result of earlier work from the computer science community is a set of technological implementations and proposed methodologies and approaches ready to be used as enabling mechanisms towards achieving integration.

These approaches can be classified attending to their focus, that is, the type of object that is meant to be shared: data, interfaces, methods or processes. This classification corresponds to the four general categories of integration approaches described in [39], which were also partially considered in previous categorizations [28, 40, 41] :

- **Data sharing:** the focus is the integrative access to information within databases and applications. Data source and target systems are always entities that produce and consume information: e.g. databases, applications, end user interfaces, embedded devices, etc.
- **Interface sharing:** providing a single-user interface or application to view a multitude of systems. The user interface of each system is adapted to a common user interface (aggregated user interface)—most often a Web browser. As a result, all participating systems are accessed through the browser, although the applications are not directly integrated within or between the providers.
- **Method sharing:** applications can share common business logic and methodologies. This is accomplished either by defining methods that can be shared, and therefore integrated, or by providing the infrastructure for such method sharing such as Web services. Methods may be shared either by being hosted on a central server, by accessing them inter-application (e.g., distributed objects), or through standard Web services mechanisms.



- Process sharing: handles the movement of data, and the invocation of processes in the correct and proper order to support the management and execution of common processes that exist in and between applications.

Table 1: Overview of approaches and technologies for integration according to the four focuses of sharing data, methods, processes and interfaces. Corresponding categories in previous classifications and enabling technologies are also shown.

References	FOCUS				
	Data		Method	Process	Interface
Integration approaches [39]	Information-oriented		Service-oriented	Business-process oriented	Portal-oriented
Data integration [40]	Eager / in-advance (e.g. datawarehouse [42])	Lazy / on-demand (e.g. mediator [43])			
Middleware systems [41]	ETL systems	Data federation systems	Application servers EAI products	Workflow systems	Portal software
Architectures for database interoperability [28]	Integrated read-only views (mediation): materialized	Integrated read-only views (mediation): virtual Federation		Workflow	
Enabling technologies [39]	Call level interfaces Native DB middleware XML ( <i>Component technologies</i> )		TP monitors Application servers RPCs, MOM Component technologies Proprietary APIs	All	WWW interfaces

Integration focusing on data can be further classified attending on the time in which the data are actually extracted from the corresponding sources in relationship to the construction of the integrative relationship [40]:

- Lazy or on-demand approach: corresponds to the mediator architecture [43] where the integrator systems first accept a query, determine the appropriate set of information sources to answer it, and generate the appropriate sub-queries or

commands for each information source. Finally, it obtains the results from the sources, perform appropriate translation, filtering and merging of the information, and return the final answer to the user or application.

- Eager or in-advance approach: is the one followed when building data warehouses [42] where a new repository is created to store the data of interest. In this case, data of interest is extracted in advance from the sources, translated and filtered as appropriate, merged with additional information, and centrally stored. Access to integrated information is done through querying the central warehouse.

A summary of the different approaches and technologies for integration according to their focus is provided in Table 1.

Some authors, like Sheth [44], review the changing focus on information systems interoperability research. Their work provides a new temporal dimension for the classification of integration approaches:

- First generation research and development emphasised data management and structured data, searching for interoperability in heterogeneous DBMS (data models, query languages and system aspects). The predominant architectural framework was the federated database system [45] for the integrated system, and the relational data model for underlying data sources.
- Second generation approaches were influenced by (i) broad variety of data (such as semi-structured formats and multimedia) and (ii) the Internet and WWW revolution. Some of the systems adapted the federated architecture to include more diverse information systems (e.g. object-relational, object-oriented DBMS). However, the mediator architecture [43] was the preferred one. Significant progress towards achieving system and syntactic interoperability was made through the increasing use of middleware solutions.
- Third generation approaches address semantic interoperability. Semantic interoperability will support high-level, context-sensitive information requests over heterogeneous information sources, hiding system, syntax, and structural heterogeneity.

Two additional dimensions for the analysis of integration approaches are provided by the generalization of the taxonomy proposed in [46], originally addressed for the classification of multi-database systems:

- The degree of autonomy of the sources.
- The level of schema or data model integration: from tightly coupled integration (usually created and maintained by the system developer) to loosely coupled integration (created and maintained by the system users).

### 2.2.1. Use of software standards

The use of software standard technology for the management and sharing of molecular biology data was not “so common” until recently. Perhaps, the introduction of the WWW and associated technologies changed completely the way bench biologists, and even bioinformaticians dealt with biological data collections [47], as most public available databases became available through web interfaces. Meanwhile, a parallel revolution was occurring in the main data management institutions (such as the National Center for Biotechnology Information (NCBI) or the European Bioinformatics Institute), and other data providers, where there was an increasing use of commercial database management systems for data organization.

Although most public data in molecular biology are managed using relational database software, very few are directly available through standard SQL interfaces or even distributed as relational tables. There is still a plethora of historic proprietary data formats (some of them are *de facto* standards in particular domains) coexisting with recent XML implementations.

The situation with analytic applications and software packages is similar (if not less ‘standard’). Software developments in research environments first follow scientific goals and necessities; therefore in many cases they lack an engineering perspective in their conception (as algorithms are developed by experts in a domain with programming skills, but not software specialists). Additionally, only some research organizations and laboratories can afford software professionals in order to perform the implementation of novel applications. In some cases, simply the software engineering skills and knowledge are not appreciated and therefore not planned. This results in a number of very useful



and important analytic programs that lack “well-defined” APIs, and with strict invocation and input/output mechanisms.

#### **2.2.1.a. The impact of XML**

One of the recent trends in biological data management is the growing number of sources (both databases and applications) that provide and digest XML data (e.g. XEMBL [48], DDBJ and XML [49], EMDB [50] and FEMME database [51]). As mentioned above, any standard data model recently developed produces the corresponding XML format. Some authors suggest XML as the “lingua franca” for science [52] or for bioinformatics data integration at least [53], while others make use of it indeed as a universal language for data integration [54].

#### **2.2.1.b. Software suites, toolkits and frameworks**

In spite of the above situation, and due to the long history of some computational biology and bioinformatics domains, there are a number of projects and initiatives to enable some sort of integration of methods (algorithms, data access, etc...) through the development of shared code and algorithms. Some examples are open source software suites such as EMBOSS [55, 56] for sequence analysis and BioConductor ([www.bioconductor.org](http://www.bioconductor.org)) for genome analysis; CCP4 program suite [57] for X-ray crystallography and CCPN programs and data model [58] for Nuclear Magnetic Resonance (NMR) structural determination; open source tools developed with particular programming languages (BioPerl, BioJava, BioPython, BioRuby) organised around the Open Bioinformatics Foundation ([www.open-bio.org](http://www.open-bio.org)); and ERATO [59] for systems biology.

There are, as well, some available and reported implementations of data parser such as [60] EMBL/Swiss-Prot Perl parser, [61] Swiss-Prot Perl parser and [62] object-oriented parsing with Python.

#### **2.2.1.c. Providing standard application interfaces**

Programmable interfaces to applications are much more complex than DBMSs (where native SQL and standard call-level interfaces such as JDBC or ODBC can be used), taking into account that they are much more diverse in the ways they consume and

produce data and information. Nevertheless, there are some standard mechanisms for defining application interfaces, such as CORBA or the Component Object Model (COM) as well as the possibility of providing well-defined proprietary APIs using programming languages such as COBOL, Java, C and C++ [39].

Although there have been a number of CORBA implementations in the molecular biology domain, they have not had a great impact as integration facilitators. Some examples include: CORBA-based genome mapping system [63], SRS CORBA interface to SRS [64]; Bio-Objects Project [65]; CORBA interface to EMBL [66]. What is more, a Life Sciences Research group was one of the Domain Task Forces at the OMG [67].

Recently there is move toward standard application interfaces such as J2EE Connectivity Architecture (JCA) and Web services. Such movement is also observed in biology: the bioWidget toolkit [68] is a set of JavaBeans components; Web services implementations such as <sup>m</sup>GRID [69], BioMoby [70], SOAP and Web services in DDBJ [71], even some implementations of WAP accesses have been reported [72].

## **2.2.2. Developing standards**

### **2.2.2.a. Nomenclature, taxonomies and ontologies**

In order to ensure data sharing and information communication, researches in different scientific disciplines have traditionally developed formal nomenclatures and agreed on naming conventions for the objects they study. Some systematic naming schemes have been created providing unambiguous identification based on some aspect of the object described (e.g. in the case of organic chemicals, the property addressed is structure; in the case of enzymes, the reaction they catalyze). When such unambiguous identification is not possible naming conventions can be built, generally emanating from expert committees or taxonomical classifications (e.g. the human gene nomenclature [73] created by the HUGO Gene Nomenclature Committee or the virus naming conventions used in the ICTV virus taxonomy [74]).

Both systematic names and agreed nomenclatures are encouraged to be used as annotations in biological databases. These can be handled as controlled vocabularies in curated databases, as well as naming suggestions for data submitters.

In the same line, recently there has been a growing interest in molecular biology in the development and establishment of different ontologies<sup>2</sup>, such as the Gene Ontology [76]. This type of ontologies corresponds to encyclopaedic efforts in order to establish a common taxonomic hierarchy of terms with the purpose of annotating biological data. Several ongoing research efforts in this direction are being made in different fields of biology and medicine, involving the cooperative work of a great number of experts. The creation of such collections of terms is labour-intensive, demanding various cycles of knowledge acquisition, abstraction and categorisation.

Other ontologies are developed to support data sources and integrated systems such as TaO [77-79], that describes a wide range of bioinformatics concepts, the ontologies underlying the RiboWeb [80] for the ribosome structure, EcoCyc [81] for *Escherichia coli* pathways, BAO (ontology underlying the BACIIS system) [82] and the ontologies to describe, discover and compose services in a bioinformatics setting in myGrid [83].

#### **2.2.2.b. Data and meta-data content (ensuring data exchange)**

In addition to the creation of naming conventions, there are also a great number of efforts in order to define data and meta-data standard models and corresponding file formats for data exchange among different laboratories. Most of the latest initiatives in this direction result in XML's DTD and Schema definitions.

Some examples are: Minimum Information About a Microarray Experiment - MIAME (gene expression data) [84] and its corresponding format MAGE-ML [85]; HUPO Proteomics Standards initiative [86] and the suggested PSI Molecular Interaction

<sup>2</sup> The term ontology was borrowed from Philosophy by the AI community. Ontologies are content theories in AI about the sorts of objects, properties of objects, and relations between objects that are possible in a specified domain of knowledge [75] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, "What are ontologies, and why do we need them?," *IEEE Intelligent Systems*, vol. 14, pp. 20-26, 1999. Therefore, they provide potential terms for describing our knowledge about this domain. Ontologies need not be limited to hierarchical structure of subsumption relationships.

Standard [87] for protein-protein interaction information and PSI-MS Format for mass spectrometry data; BioPAX ([www.biopax.org](http://www.biopax.org)) for pathway information; the Systems Biology markup language [88].

### 2.2.3. Bridging the semantic gap

As long as computational entities (programs, databases, knowledge bases and servers) do not communicate with each other or with human beings, semantic ambiguities are isolated and relatively inconsequential [26]. However, as soon as some sort of sharing is needed, each entity must understand the other's language, that is to know mappings between the symbols (or syntax) of the language and the real-world abstractions they are referring to.

Most of the systems developed in molecular biology do not explicitly declare the semantics of the biological entities they manage and/or analyze. The mapping to real-world objects (or in fact the concepts and abstractions representing those objects) is implicit and hidden in the data structure and syntax. Nevertheless, there are some examples of declarative systems, such as RiboWeb [80], EcoCyc [89], MHCWeb [90] and BACIIS system [91].

In addition, some tools have been specially designed to express database schema mapping in the context of biological data integration: Cheung and co-workers extended the entity-attribute-value (EAV) modelling technique to express interdatabase 'schema' mappings [92], demonstrating query interoperation between two chromosome map databases (DB/12 and GDB) (although it allows schema evolution to be handled gracefully, it does not allow the execution of inter-database joins); Davidson *et al.* [93] proposed a language for expressing database schema transformations and specify integrity constraints, based on the data transformations needs in the integration at the Philadelphia Genome Center for Chromosome 22.

### 2.2.4. Making use of the data

A paradigm shift in access to molecular biology data was driven by the huge increase of users due to the availability of databases through the WWW. Although most of the users experience database access through form-based interfaces, some work has been done towards new ways of interacting with biological data in distributed environments.

In the context of the Kleisli system [94], Buneman and co-workers present a set of techniques for querying and transforming biological data using the CPL (Collection Programming Language), based on a type system that allows arbitrary nesting of the collection types together with record and variant types [95].

Chen *et al.* [96] describe a suite of tools that provide advanced querying mechanisms in the framework of the Object-Protocol Model (OPM). They are schema-driven (thus generic) and allow ad-hoc queries to be constructed using graphical, Web based interfaces. They generate queries in an object-oriented query language, OPM-QL, and processed using OPM query translators. Querying support for complex (application-specific) objects is provided via OPM Application-Specific Data Types (ASDTs) and methods.

Mork *et al.* [97] introduce the PQL query language, which generalizes StruQL, a query language for semi-structured data such as XML, used in the GeneSeek genetic data integration project. PQL generalizes capabilities of other XML query languages (such as XPath, Lorel, XML-QL and XSL) by allowing the user to express assumptions that guide the construction of complex paths. The query contains a collection of rules that are used to instantiate paths that adhere to the rules.

Chen and Jamil [98] propose the Internet Function Definition Language (IFDL), an extension of SQL data definition language to allow the creation of Internet functions (available analysis tools on the Internet) as remote user defined functions. Declarative queries are supported by means of a new query language, the hyper text query language (HTQL).

Labrix and Jakoniene [99] propose a query language containing operators that should be present in any query language for biological databanks.

### 2.3. Integrated systems

In addition to the previously presented work, there are a number of initiatives and systems that are designed to act as biological data integrators. The main characteristics of such systems are: sources to be integrated remain autonomous and users demand read-only access to data sources (write access is not needed).



A categorisation of these systems is provided in table 2, based on the following dimensions:

- Data, interface, method or process sharing (corresponding to information, portal, service or business process oriented approaches as described by Linthicum [39]).
- Materialized versus view integration (corresponding to eager/in-advance or lazy/on-demand as described by Widom [40]).

Most solutions and approaches towards integration in molecular biology can be classified as data focused in the context of the proposed categorization.

**Table 2:** Some example integrated systems in molecular biology

<i>Approach</i>		<i>Generic solutions</i>	<i>Implementations</i>
Data sharing	Syntactic materialized (Datawarehouse)	LIMBO architecture [100] EnsMart [101]	PEDANT [102] BioMolQuest [103]
	Syntactic view (Mediator)	DiscoveryLink [104] Kleisli [94] P/FDM [105]	TINet [24]
	Semantic integration	TAMBIS [106] model-based mediation [107] SEMEDA [108]	
Interface sharing		ENQUIRE [109] SRS [110]	Entrez (NCBI) [38, 111] Integr8 (EBI) [112] DBGET/LinkDB [113]
Method sharing		BioMoby [70] myGrid [69] (*)	
Process sharing		PLAN [114] HyBrow [115]	PRECIS [116]

(\*) Has recently incorporated the workflow paradigm by the use of Taverna [117]

In addition to the above categorisation, some other dimensions can be considered when analysing integration examples, such as if they correspond to generic solution/technologies or particular implementations (hard-coded); the underlying data models and technologies used; the type of interfaces provided (e.g. programmable vs. non-programmable; browsing vs. navigation).

## 2.4. Discussion

### 2.4.1. Choosing the right approach

From the analysis above it is clear that there is not a single winning solution or even methodology to create and integrated system for molecular biology data. Current systems provide different perspectives on a complex mission, providing in some cases overlapping functionalities, and in others complementary views.

Table 3: Strengths and weaknesses of different integration approaches according to several dimensions (adapted from [27])

Access mode	Strength	Weakness
Browsing	Suited for exploration Suited for manual inspection Single-page retrieval Easy to understand	Not suited to handle large data sets Not suited to multi-step workflow with processing of interim results Limited flexibility
Querying	Flexibility Suited to handle large data sets Suited to multi-step workflow with processing of interim results	Not suited to manual inspection Use of query language requires sophistication
<b>Integration type</b>		
Materialized	Permits data cleansing, transformation, filtering Load on operational sources only at data refresh times Referential integrity is designed in Changes in remote sources do not directly affect the warehouse's availability	Heavy maintenance burden Data currency depends on refresh frequency May not scale well May lose specialized search capability of native data source Requires centralized control over data
View	Data is always current Lesser maintenance burden Support for native specialized search capability Can access data that cannot be copied (e.g. Web) Supports autonomy of individual data source providers	Data cleansing on the fly can degrade performance Load on operational sources at query time Referential integrity across sources is difficult to maintain Changes in remote sources need to be dealt with on the fly
<b>Query language</b>		
Procedural	Can be very precisely tuned for a specific task No limitation on expressive power	No limitation on expressive power Ad hoc inquiries can be difficult Extension can be difficult
Declarative	Programming ease, economy, maintainability Flexibility Ease of ad hoc querying	May be more difficult to learn Some tasks require procedural step-by-step access to data
<b>Common data model</b>		
Relational	Well-understood data model (since 1970) Mature technology SQL powerful and widely used	Tabular data model may not fit scientific data well
Non-relational	Hierarchical data models good fit for scientific data	Relatively immature technology Standard database desiderata hard to attain due to increased complexity
<b>Type of approach</b>		
Generic approach	Extensibility Maintainability Easy to understand	Greater up-front cost Sometimes greater complexity
Hard-coded approach	Can be finely tuned to optimize for specific case Can be rapidly prototyped	Not readily extensible May be difficult to maintain



In order to help in the process of evaluating different systems and solutions enabling integrative data analysis, some of the strengths and weakness of a given approach should be taking into account, according to five parameters: access mode, integration type, query language, common data model and general type of approach (see Table 3).

In spite of the difficulty to integrate molecular biology data, it is still possible to work towards this goal from different fronts:

Data providers (both databases and applications) should ensure proper means of data citation, adopt software standards (e.g. XML, UML, relational data models) and provide standard interfaces (e.g. JDBC/ODBC, SOAP, CORBA, Web services). It is also necessary to continue defining, developing and adopting data standards and controlled vocabularies and/or ontologies. In order to facilitate the task of creating new data services and applications it is important to provide and share data models (both at the syntactic and semantic level). Finally, any new data source or application built from previous data should conveniently track and handle data provenance.

Research in data technology can continue their work towards the development of standard languages for workflow definition, efficient query planning over web/XML and semantic integrators, handling change (e.g. schema transformations and mappings) and developing biological relevant native data types and operations.



---

*Chapter 3*    **Organising biological multidimensional image data**

---

### **3.1. Introduction**

Imaging techniques (i.e. microscopy techniques) are widely used in biological research as they provide, in contrast to other experimental techniques (such as spectroscopy), directly interpretable structural information of the object being studied. The three main groups of microscopies in biology are: light (optical) microscopy, electron microscopy and scanning probe microscopy.

The range of nominal resolutions that can be attained is limited in practice by a number of factors: the microscopy technique, the particular microscope setup and, above all the sample preparation techniques needed when analysing biological specimens. All these factors limit the use of a particular technique for just a range of biological structures and processes.

Some microscopic imaging techniques are able to provide biological images that expand the first two spatial dimensions: confocal and video microscopy (among the light microscopy group), three-dimensional electron microscopy (among the electron microscopy group) and atomic force microscopy (among the scanning probe group). The new dimensions added provide data in the third spatial dimension, as well as the temporal dimension. These 3D spatial, 2D temporal and 3D temporal images will be generically termed "multi-dimensional images".

Multi-dimensional light microscopic techniques fall into two distinct classes [118]. The first is video-enhanced contrast microscopy: as a conventional wide-field light microscopic imaging technique, the whole of the field is subjected to simultaneous illumination, and multi-dimensionality is achieved through the addition of a temporal dimension. The second is that of confocal microscopy, a type of scanning light microscopy, in which the observed specimen is interrogated by a small point of light which is moved relative to the specimen in a regular scan in order to generate an image. Confocal microscopy has the advantage of eliminating out-of-focus blur from the resultant images, permitting 3D data to be collected easily and non-invasively by optical sectioning. For a recent review of light microscopy techniques in live cell studies see [119].

While the collection of multi-dimensional data sets with light microscopes has now become routine, the analysis and interpretation of these image generally require significant time and effort [120]: each type of image seems to require a specific set of processing algorithms and parameters and the software tools required for extracting useful information from the resulting multi-dimensional data sets are not completely developed.

Three-dimensional electron microscopy acquires projections of the specimen from different directions, which are later merged computationally to obtain a "reconstruction" of the three-dimensional structure or map [121]. In its wider sense it includes electron crystallography of proteins, helical reconstructions, single-particle methods and electron tomography in which all projections are recorded for the same specimen by physical tilting.

High resolution scanning probe microscopy of biological samples is primarily achieved using Atomic Force Microscopy (AFM) [122]. The AFM is a powerful technique that reveals the surface structure of protein assemblies in their native environment at sub-molecular resolution. AFM is able to acquire surface topographies with a lateral resolution of 0.5-1 nm and a vertical resolution of 0.1-0.2 nm. For a recent review see [123].

Despite the fact that key biological information has been produced by a large number of these different types of microscopies for many years, the information was neither organised nor became generally easily accessed by the scientific community. This situation, clearly unsatisfactory, motivated the launching of an international collaborative project, BioImage, to create a new infrastructure to support the storage and management of biological multi-dimensional images. This project constituted a collaborative effort of a number of scientific laboratories as well as industrial partners, complemented with a network of associated laboratories acting as test users, and was coordinated by our laboratory at the Centro Nacional de Biotecnología (Madrid, Spain).

A number of exploratory studies were carried out between 1993 and 1996 [32], before the BioImage project was officially launched. One was centred on macromolecular structures and developed a 'proof of concept' prototype that outlined some of the general organization principles applicable to complex image data [124]. Another was centred on the organization of experimental data relevant to confocal microscopy [125].

### **3.2. Objective**

Design and develop a new database system to store and manage multidimensional microscopic images of biological specimens.

### **3.3. Methods**

#### **3.3.1. Architecture**

The content data in BioImage is quite homogeneous from the viewpoint of data types, as it addresses the organization of multi-dimensional images. However, it clearly deals

with several distinct levels of biological organisation, ranging from structural biology of macromolecules to three-dimensional images and videos often found in cell biology studies (Figure 3). 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 was organized around two database servers specialising in two broad areas of biological interest:

- structural biology of macromolecules (at the Centro Nacional de Biotecnología (CNB) - Madrid, Spain) and
- cell biology (at the European Molecular Biology Laboratory (EMBL) - Heidelberg, Germany).

The two database server sites acted also as two design and development centres focusing on the information relevant to the multi-dimensional microscopy techniques that provide structural information at both biological levels: three-dimensional electron microscopy (3D-EM) and atomic force microscopy (AFM) for the study of macromolecules, and optical microscopy for the study of cellular structures.

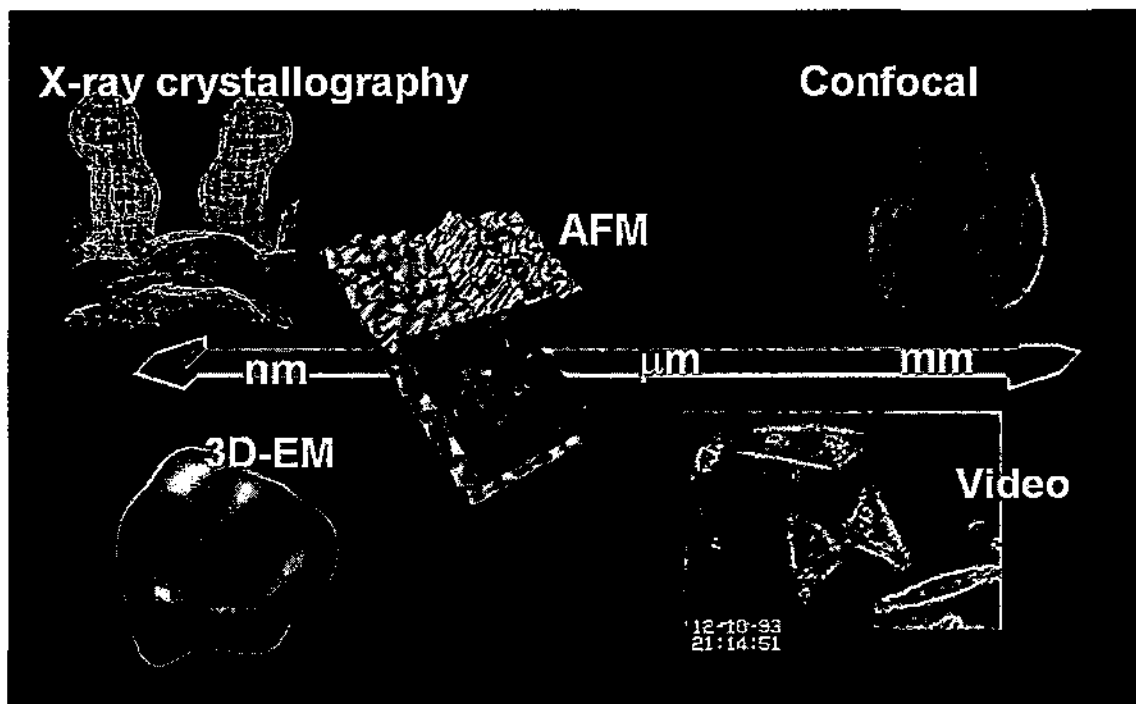


Figure 3: The multidimensional data in the BioImage database are generated by various microscopy techniques that provide different resolutions, and are therefore most suitable to study biological structures at different organizational levels (from macromolecules to cells and tissues).

### 3.3.2. Phases and methodology

The design and development of the BioImage database was accomplished as part of the overall BioImage project. Due to the broad scope of the database contents, as well as the wide range of skills needed to accomplish a completely new conceptualization and design in an evolving research environment, an interdisciplinary development team was organized. In this team I acted as the coordinator of the technological aspects of the database, as well as the principal designer and developer of the CNB server site (in charge of the organization of the data relevant to the multidimensional images of biological macromolecules).

#### 3.3.2.a. Definition

The BioImage database design and development followed an iterative and incremental approach, also known as prototyping approach [126]. It proceeded as a series of iterations that evolved into the final system. Each iteration worked on the results of the previous one, allowing for the revision of user requirements and the cycle of abstractions and instantiations to refine data models. Prototypes created during each phase or iteration varied in the degree of data content, functionality and nature (from paper diagrams to operational software).

The first step was the determination of the scope of the data to be stored, together with their complete description, taking into account the targeted group of users (designated community) and the applications they were interested in. To be useful for scientific purposes, every image must be properly identified and described. To determine both the Content and Preservation Description information, an initial list of relevant data items (descriptors) for the BioImage database was compiled with the help from biologists and microscopy experts (Producers). The input from scientists having various backgrounds was essential to define all aspects of the various microscopy techniques and the biological specimens to be documented in the database.

Most of the work in the database design was spent to identify and structure the properties relevant when describing the circumstances in which a multidimensional image was obtained.

The methodology used to obtain the specifications for database content included:

- **Brainstorming group sessions:** The first specifications for the BioImage database were defined at a meeting of both partners and test users (in January 1997), where first hand contacts were done with the diversity of data types brought by the test user community and the different partners. These specifications were compiled in the form of a “list of descriptors” that was further processed in subsequent months.
- **Interviews with experts:** Additionally, specific studies on a number of topics, such as the report on “interaction with atomic resolution data”, were produced in collaboration with some of the partners and test users.
- **Fake submissions:** Building upon the BioImage list of descriptors, the first data submissions (in paper) from the partners took place.
- **Analysis of literature:** The study of examples of “real world” studies of macromolecular data led us to refine the Content and Preservation Description Information details.

### **3.3.2.b. Implementation decisions**

Due to the broad scope of the database, and the wide data producer community of researches, multi-dimensional images obtained by the various microscopy techniques presented a high degree of heterogeneity in terms of image data formats.

A decision was made to archive and store the data in the format provided by the submitters. This certainly simplified the submission procedure, as data transformations were not performed at ingest. Nevertheless, image data should be made available to consumers in a convenient way. As the database stored image in multiple file formats, appropriate tools had to be created to allow for transformations in dissemination (through the tools developed by project partners described in [127]).

### **3.3.2.c. Database design**

The database design was accomplished following the classic three-step methodology: conceptual design, logical design and physical design.

Independent analysis and organization of the specifications and database contents for the two database servers, led to a first conceptual design. An abstract model of the contents of the database was created using Infomodeler version 3.1, by InfoModelers

Inc. Bellevue, WA, USA. Infomodeler provides the possibility of designing at two abstraction levels: an object-oriented approach (object role modelling, ORM; see [www.orm.net](http://www.orm.net)) and logical modelling (entity-relationship modelling [128]).

Conceptualizations of the data relevant to 3D-EM and AFM studies (for the macromolecular structure server in Madrid) were done using object-role models and formalized in FORML (Formal Object-Role Modeling Language) [129], the language used in Infomodelers to express object-role models. A parallel and independent design was also performed at the cellular server in Heidelberg (centred on light microscopy studies).

Formalization of contents of the two database subsections provided an appropriate first level of abstraction for designers to reason about data, a powerful and most important, common mechanism to communicate design decisions of the two development teams, and a convenient automatic mapping vehicle to obtain a first version for the database logical model. A first demo prototype was developed at the end of December 1997 for each database server, that was presented at the project meeting held in January 1998.

Integration of the independent models created for the macromolecular and cellular servers was accomplished by re-examination and discussion among designing teams of the two servers, with the help of scientific database experts. Although communication and debate on database models was performed using object-role modelling concepts, a decision was made to directly create a single logical model for the BioImage database.

Once the full database model was created the first BioImage Database prototype, was implemented using a commercial DBMS (Informix Universal Server; Informix Software Inc., Menlo Park, CA, USA) that was presented to the test users in June 1998.

#### **3.3.2.d. Development of WWW-based query and submission interfaces**

A first prototype for search interface was presented at a project meeting in January 1998. The presentation of this prototype was the cause of quite vigorous and stimulating discussions among the partners and test users, and as their result a number of modifications were proposed to both the data model and the query interface. Most importantly, the main lines for the submission interface, which was still lacking at the

time of the meeting, were defined. The result of successive rounds of revisions and evaluations that took place during the second semester of 1998 was a revised BioImage Database prototype that was opened to world-wide use.

A technical meeting was held in November 1998 with the aim of launching a new prototype, including a revised database model and the tools developed by the rest of the partners. This prototype was released for the test users and the general public in February/March 1999. Further revisions to this prototype were done during the last months of the project, including a fully functional submission interface for macromolecular studies, a new customizable query interface, and the addition of eight studies to the database.

### **3.4. Results**

The overall result is a software infrastructure to store and manage biological image data obtained from a variety of microscopic techniques [31]. This software infrastructure is designed to accommodate the needs and uses of the Designated community, i.e. those researches that make use of multidimensional imaging techniques, namely Atomic Force Microscopy, Three-dimensional Electron Microscopy, Confocal Microscopy and Video Microscopy. The Designated community is also the community of 'Producers', while the 'Consumers' are the set of biological science researches.

#### **3.4.1. Information model**

The BioImage database model (archive information package) was designed general enough to store any kind of image data, independently of its level of interpretation (e.g. from raw images, to reconstructed 3D maps). The database was populated by direct submissions from the scientific community (the data producers).

The intended scientific analysis had several implications, and was independently studied at different levels:

- direct interpretations by means of visualization [127],
- relationships with atomic coordinate data (further details in following 'Combined studies' section),



- advanced access to macromolecular structural data by means of query-by-content methods [130].

The information model was restrained to contain data relevant to archiving purposes. Advanced image description in terms of structural content (such as that required for query-by-content applications) was not part of the core data model.

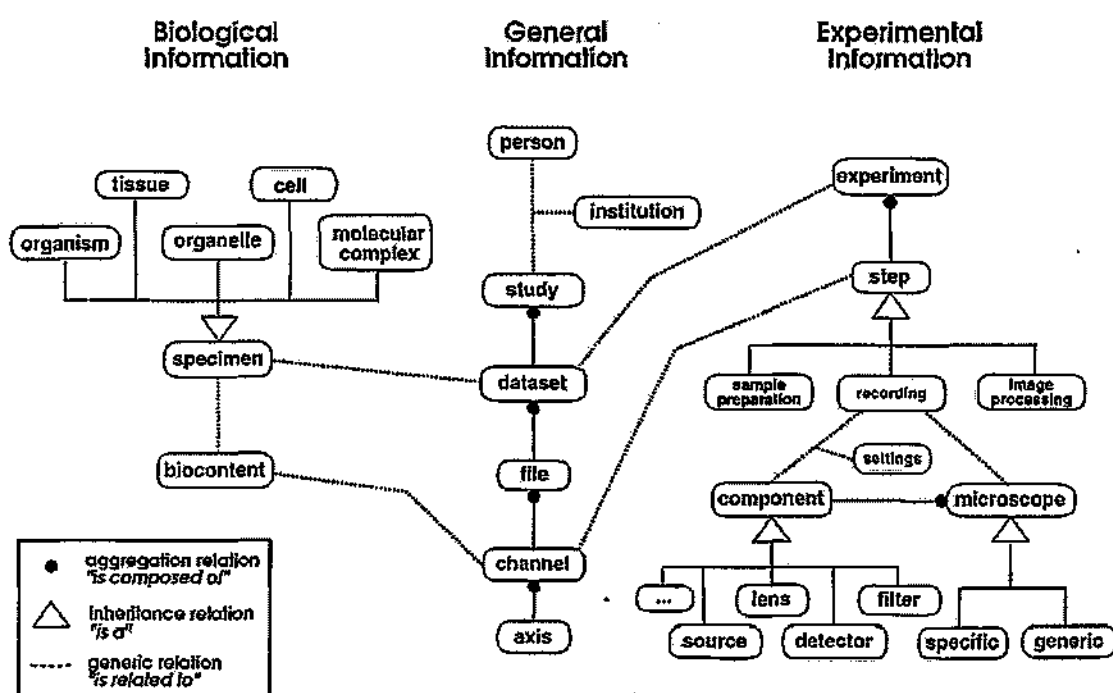


Figure 4: Simplified view of the structure of the BioImage data model showing its main entities and relationships (adapted from [31]).

Content Information on biological multi-dimensional images should contemplate both the image data and its representation in a digital format, as well as the biological object of which structural data is provided. Therefore, each multidimensional image in BioImage is accompanied by a description of the biological specimen being studied, as well as an account of the experimental details involved in the sample preparation, observation, and subsequent data processing. The BioImage information model can be subdivided into the following areas:

- Content Information
  - Microscopic data
  - Biological data
- Preservation description information:



- Administration and organization
- Experimental details:
  - Sample preparation
  - Data acquisition and instrumentation
  - Image processing

### 3.4.2. Database model

The data model of the archive information package has been carefully designed to suit the needs of the different microscopy communities (Figure 4). The central entity in the model is the 'data set' (or result of an individual experiment). Since tightly linked data sets could be produced by different, but related, experiments a grouping entity ('study') was created. A study is thus a collection of data sets. Data sets are further described in terms of three major classes of information:

#### **Biological information**

The biological information is organised around two entities: 'specimen' and 'biological content'. The specimen specifies the biological object used in the experiment. It can be an organism, an anatomical structure/tissue, a cell, an organelle, or a molecular complex. The biological content represents the biological feature observed in the microscope, which can be just a part of the specimen used for the experiment.

#### **General information**

The general information contains data relevant to all kinds of studies, independently of the microscopy technique or biological specimen of the experiment. This includes general information on the study (authors, funding, publications and supporting data), as well as the description of the data set themselves (location, format, size, etc.)

#### **Experimental information**

The experimental part of the model was designed to store the experimental workflow (from the sample preparation steps to the data processing, via the mounting and the data acquisition. The central entity is therefore the 'experiment', composed of a set of 'steps'. Some steps, like 'mounting' and 'recording', require an explicit treatment in the model due to the large number of different parameters, while others are stored

generically and documented only in terms of buffer medium, instrument, and physical and biochemical parameters.

Additionally, special attention was paid to establish links to already existing databases (e.g. taxonomy, literature, sequence and atomic coordinate databases). The content of these databases is either imported into BioImage and used as controlled vocabularies that ensure data consistency (e.g. NCBI taxonomy for naming organisms) or linked dynamically (e.g. bibliographic information in MEDLINE). Other biological databases relevant for molecular structures included are Swiss-Prot and EMBL databank (for biological sequences) and the Protein Data Bank (PDB) (for atomic structures).

### **3.4.3. Functional model**

BioImage was designed from the onset to be accessible through the Internet. This is consistent with mainstream technology and makes it relatively simple to query, download, and visualise single images and complete three-dimensional data sets.

The BioImage database provided WWW interfaces for data ingest, access and visualization. Implemented using the Informix Web Integration Option, a set of HTML template pages are stored within the database. Actual pages visible to the browser are created dynamically based on user requirements and database content.

#### **Ingest: submission interface**

The submission interface is certainly the most complex interface, since a well-designed dynamic submission interface not only queries the database and incorporates the returned information into forms, but also writes the entered information into a database and/or into a tagged flat file. For security reasons, the submitters should not be allowed to interact directly with the production database; therefore an independent submission database will temporarily store all the information entered into the submission forms.

#### **Finding aid: query interface**

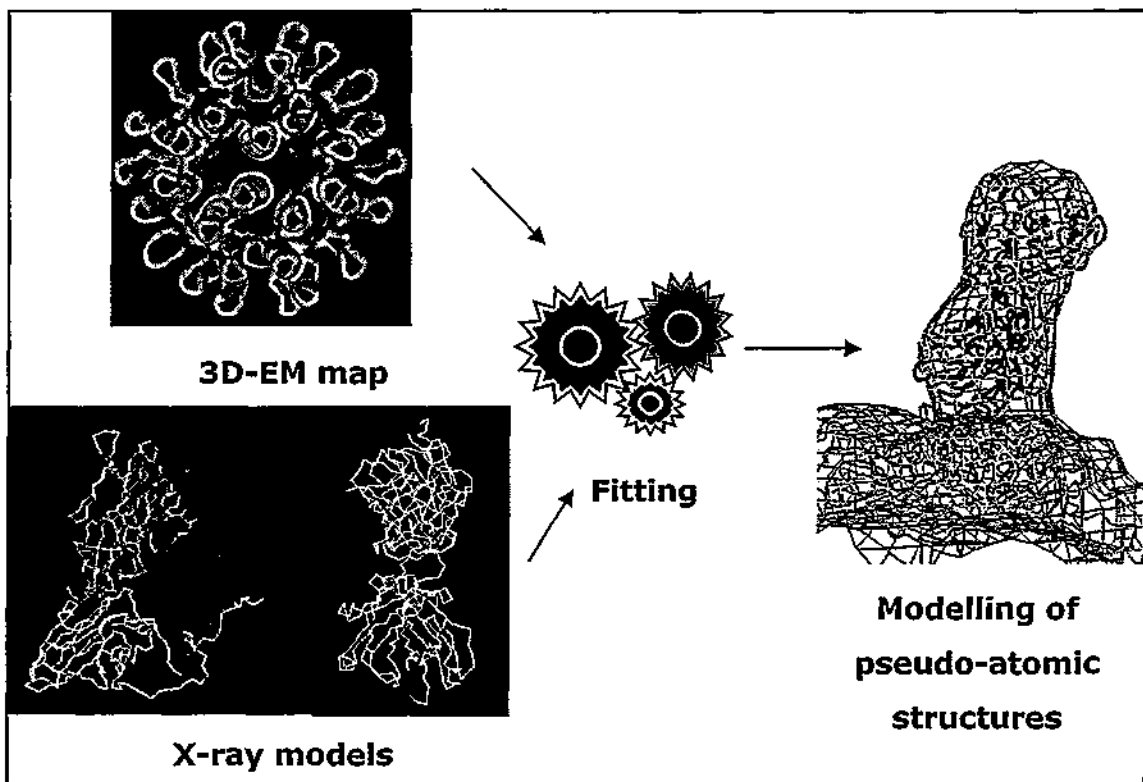
The query interface allows the user to enter search criteria and to query the database for matching studies. Query settings are translated to SQL code and sent to the database, so the user does not have to write any SQL statement.

#### **Visualization interface**

The central element of the visualization interface is the data set with its details. The visualization of and the interaction with the different multi-dimensional microscopic data require the development and integration of new tools, also developed in the context of the BioImage project, namely conversion and rendering of three-dimensional image files [127] and handling of video data [131].

#### 3.4.4. Combined studies

The database BioImage opened the possibility of organizing a diversity of combined studies where multi-dimensional images were correlated with other types of structural information, in particular, with molecular models. Careful consideration to this type of studies was taken at the time of designing the database model [132].



**Figure 5:** Fitting study of the FMDV-Fab complex. Atomic structures of both the FMDV capsid protomer and the antibody complex are fitted to the 3D map of the whole virus-antibody complex. The relative disposition of the Fab in the complex allowed the modeling of residues situated in the hinge of the epitopic loop.

Most of these combined studies related information at high resolution (mainly obtained by X-ray diffraction) with three-dimensional maps obtained by means of 3D-

EM. These studies can be classified into three broad categories, according to the type of data combination:

- Comparison/validation: atomic models are used for the comparison/validation of 3D-EM structural data.
- Fitting: 3D-EM maps are used as spatial/structural context where atomic models are fitted, providing a new atomic model according to the “divide and conquer” strategy. The information obtained ranges from the atomic modelling of a complex molecular assembly to the characterization of local movements in conformational changes within large macromolecules.
- Phasing: 3D-EM maps are used as search models for the crystallographic phase determination.

The implications, in terms of infrastructural needs and uses, of these three categories are quite different. In the case of comparison/validation of 3D maps, microscopists are Consumers of the current archives containing atomic models (such as the PDB), while phasing studies are performed by X-ray crystallographers, therefore becoming Consumers of the database containing 3D-EM maps (like the BioImage database).

Relationships between atomic models and 3D maps are much more complex in fitting studies. The main result of a fitting procedure is a new atomic model (Figure 5) and, as such, it can be deposited in the PDB. Nevertheless, the experimental description of such result should account for the computational methodology behind the fitting procedure, as well as the details of the three-dimensional electron microscopy experiment. Finally, access to initial experimental data (both atomic models and 3D map) should be provided. This is accomplished in the BioImage database by providing cross-references to initial atomic models in the PDB and storing 3D-EM maps and corresponding geometric operations in the form of fitting and assembly matrices.

### 3.5. Discussion

The BioImage database is the first database reported to manage and store biological multidimensional images obtained by a wide range microscopy techniques. It has also shown the needs and possibilities, in terms of data model and data access, of biological image information.

The design and development of the BioImage database was performed in a collaborative environment. The development team was geographically distributed, with diverse backgrounds (in terms of know-how and training), for which it was essential the use of a formal language that helped in the discussion and conceptualization of data models. The database content was very broad in scope, both in terms of experimental techniques and biological content. Except for 'proof of concept' prototypes, providing partial and incomplete abstractions, no other previous data conceptualizations of experiments were available.

The tight links found in the context of the BioImage project, between macromolecular structural data of diverse nature, such as the studies combining atomic models obtained by X-ray diffraction and maps obtained by 3D-EM, demand further work in order to develop public infrastructures to facilitate the integration of macromolecular structural data.

From the pioneer work done in the BioImage project, a number of efforts have been reported in the literature towards the creation of databases for managing biological image data. None of these works addresses the creation of a world-wide archive of multidimensional images, but highlight complementary aspects pertaining to biological image data:

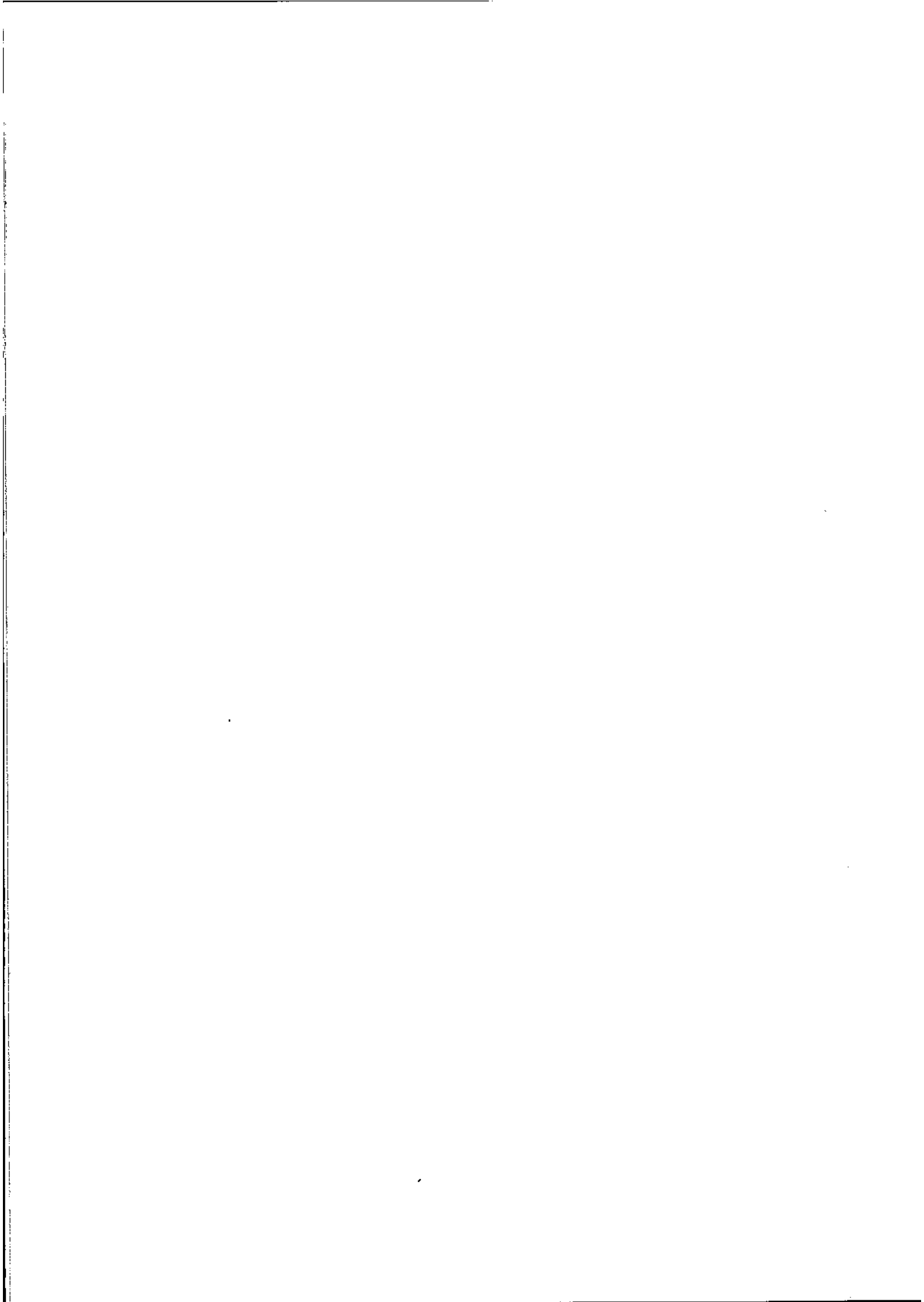
The Global Image Database (GID) [133] is a web-based structured central repository for scientific annotated images developed by GlaxoWellcome. It is designed as a collaborative database to manage images from a wide spectrum of imaging domains ranging from microscopy to automated screening.

The Open Microscopy Environment (OME) [134] is an informatics solution for the storage and analysis of optical microscope image data. The primary goal of OME is

to enable the automatic analysis, modelling, and mining of large image sets with reference to specific biological hypotheses.

Finally, there are also a number of databases attached to particular instruments and software in the context of 3D electron microscopy, such as the three-dimensional reconstruction program suite for biological bundles [135], the IMIRS system [136, 137], the cell-centered database for electron tomographic data [138] and the Electron Microscopy Electronic Notebook [139].







---

*Chapter 4*    **New infrastructures for structural data**

---

**4.1. Introduction**

Elucidation of the three-dimensional structure of biological macromolecules can be done using different experimental techniques. Among these techniques, three-dimensional electron microscopy (3D-EM) offers some advantages for understanding the cell at a molecular level. Although it cannot be used routinely to obtain atomic resolution information on molecules (with remarkable exceptions in the case of some electron diffraction experiments), it provides enough quantitative measurements of the conformation of macromolecules in the range of 8 to 30 Å. Therefore it enables the structural characterization in the gap between atomic resolution methods (basically X-ray diffraction and NMR) and other microscopy techniques.

During the last years, there have been a significant number of advances in the field: more suitable specimen preparation procedures, instrument enhancements and better algorithms for data processing. All these efforts have resulted in different complementary directions of progress: better resolutions obtained which allow the location of folds and even secondary structure elements; advances in tomography that

allow to get structural information of complex machineries in the cell; quantitative integration with atomic resolution information; and successful application to the study of conformation changes. Some recent reviews highlight different aspects and domains of application for this powerful technology: structural genomics [140], molecular medicine [141], virology [142] and cell biology [143].

#### 4.1.1. 3D-EM data

Images taken with the electron microscope can be considered as two-dimensional projections of the specimen being studied. After combination of projections at different angles a complete three-dimensional reconstruction of the sample can be obtained. The exact steps taken for the reconstruction from projection images vary according to the nature and symmetry of the specimen. Thus, the main structural result obtained in a 3D-EM study is a three-dimensional image, or map, in which each voxel is related to the Coulomb potential of the biological sample at that position.

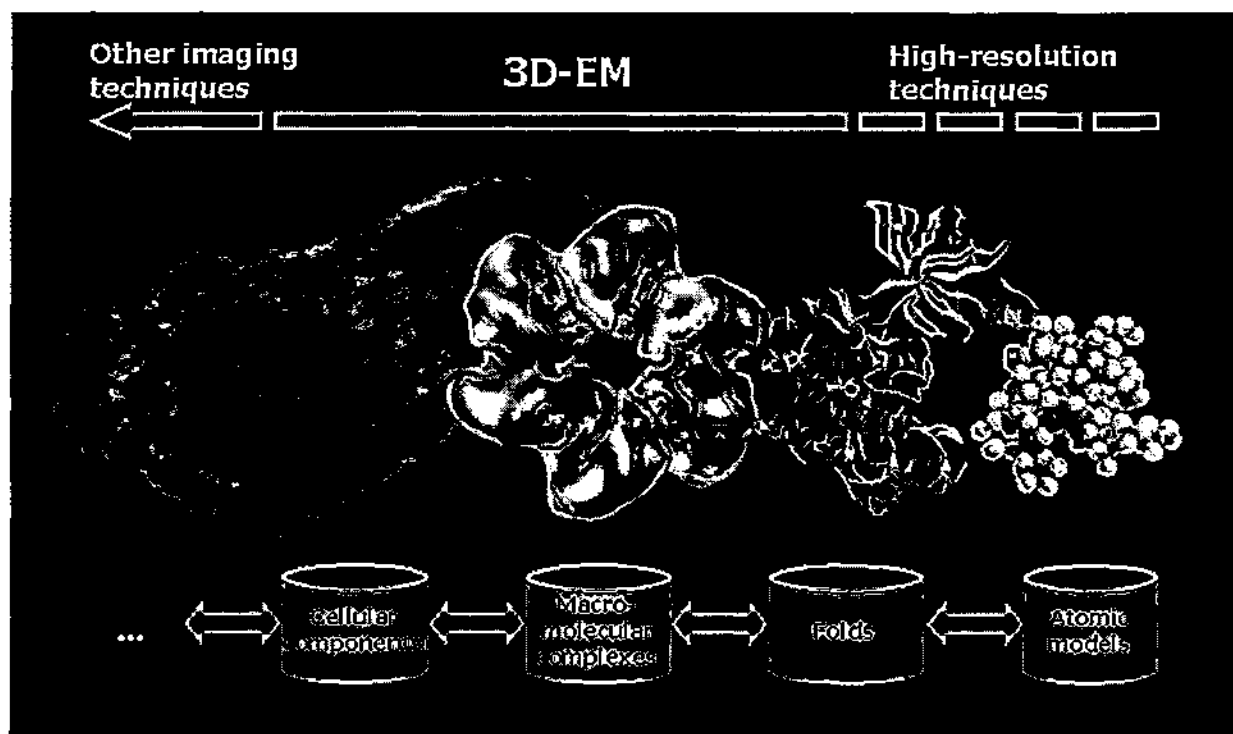


Figure 6: Scope of 3D-EM data. Conceptual representation of the biological and structural scope of 3D Electron Microscopy data, and the relationships of structural information provided. (Image of the mitochondria, corresponding to a section of a tomographic reconstruction, was kindly provided by Drs. G. Perkins and M. Ellisman).

3D-EM has been used to study macromolecular complexes (the smallest being reconstructed is around 200 kDa [144]), icosahedral viruses [142], complex machines [145] and whole sub-cellular elements using tomographic approaches [143].

There are different methodological areas of application within the field of 3D-EM, which translates into certain differences in the experimental approaches, as well as differences in the level of resolution that can be expected from the actual 3D reconstructions. For instance, the field of electron crystallography has already solved a number of macromolecules at atomic resolution as well as a larger number of them at an intermediate resolution in the range of 0.6 nm. In fact, there are as today a few PDB entries that actually hold these atomic-resolution 3D-EM reconstructions.

The current state of the art allows undertaking the following projects:

- Obtaining atomic coordinates of a relatively small number of “difficult-to-solve” proteins (like membrane proteins).
- Structural information, at medium resolution (8-20 Å), of a number of macromolecular complexes which cannot be directly studied by X-ray diffraction or NMR approaches.
- Study of conformational changes [146], under different conditions (pH, ionic strength, cofactors, etc.) or different life cycle states.
- Three-dimensional visualization and characterization of organelles and sub-cellular components, with increasing resolutions achieved (5- 20 nm).
- Modelling atomic coordinates of a whole assembly or sub-assembly, using the 3D-EM volume as an experimentally determined structural constraint where atomic models are fitted.

An increasing number of 3D-EM fitting studies have been reported, where information of atomic models and 3D-EM data is integrated. The motivation of these experiments is clear: nowadays it is still not possible to obtain the structure, at atomic resolution, of very large macromolecular machines or complexes. Early qualitative approaches to combine and compare 3D-EM and atomic resolution data have evolved into quantitative methods [147-149]. First attempts to detect and identify

macromolecular complexes in sub-cellular tomographic reconstructions have been also reported [150].

#### 4.1.2. Organising the information

Despite the complementary nature of atomic models (obtained mainly from X-ray crystallography and NMR techniques) and low/medium resolution maps (obtained by three-dimensional microscopy), the level of access to such experimentally determined data was indeed very different. Atomic models are routinely deposited and can be retrieved efficiently from a number of databases (specially the Protein Data Bank, PDB). In contrast, despite the importance and relevance of 3D-EM structural information, access to the data has been, until very recently, restricted to the form of direct contact with the authors. The situation started to change with the work done during the BioImage project, in which the first database for handling biological multi-dimensional images was developed (see previous chapter).

Until very recently, the major database centres in the world have focused attention on integrating atomic resolution data and providing flexible tools to search such data. Knowledge of the three-dimensional structure of a number of macromolecules has been organized as a work of many years within the framework of the Protein Data Bank (PDB), which was established in 1971 at Brookhaven National Laboratory.

The Protein Data Bank (PDB) is the single consistent world archive of 3-D macromolecular structure coordinate data. The Research Collaboratory for Structural Bioinformatics (RCSB) in the USA, the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (EBI) and the Protein Data Bank Japan (PDBj) at the Institute for Protein Research in Osaka University serve as custodians of the so called world-wide PDB (wwPDB), with the goal of maintaining a single archive of macromolecular structural data that is freely and publicly available to the global community [151]. The wwPDB members serve as deposition, data processing, and distributions sites.

The Macromolecular Structure Database (MSD) [152, 153] is the European Bioinformatics Institute division for the collection, management and distribution of data about macromolecular structures. MSD is working closely with the RCSB to ensure that the core data that make up the PDB is maintained in a consistent and uniform

manner. EBI-MSD staff has processed all PDB depositions made at EBI since June 15th, 1999. They have developed a database schema for the representation of macromolecules and experimental data from X-ray and NMR techniques (providing access to PDB contents and added value data).

The PDB contains three-dimensional atomic coordinates of macromolecules, obtained mainly by experimental techniques (such as X-ray diffraction, NMR, electron diffraction, fibre diffraction, neutron diffraction) and theoretical models. It contains also the results of fitting atomic models into 3D-EM maps. The PDB does not contain any structural information in the form of 3D-EM maps, nor structural information of medium or low resolution.

While problems for X-ray and NMR data are far from solved, the issue of integrating data from 3D-EM was not addressed by the major structural database providers until the launch of an international collaborative project, the IIMS (Integrating Information about Macromolecular Structure). The strategic goal of the IIMS project was to provide a public infrastructure for the storage and management of structural information on biological macromolecules, integrating the 3D-EM data with already available collections of atomic coordinates structural data.

The flow of 3D-EM data publication is in the order of 120 structures per year which, while being an order of magnitude lower than the number of structures determined by X-ray crystallography, it corresponds to the number of X-ray structures deposited per year no more than a decade before. Considering that the standardisation of 3D-EM techniques is quickly making possible the access to this form of analysis to a much larger community, these numbers are expected to grow substantially over the coming years, demanding an urgent solution in order to enable their organization in a public infrastructure.

## 4.2. Objectives

1. Design and development of a public infrastructure for the management, organisation and dissemination of data on the structures of biological macromolecules solved by three-dimensional electron microscopy (3D-EM): the Electron Microscopy Database (EMD).

2. Development of a prototype system to integrate the results of 3D-EM with models from X-ray and NMR methods into a single standardised data base at the European Bioinformatics Institute through the incorporation of 3D-EM data to the current Macromolecular Structure Database (MSD).

### **4.3. Methods**

#### **4.3.1. Establishing relationships with atomic models**

The complementary nature of atomic data and 3D-EM maps was realised during the creation of the BioImage database, where we first created a data model for the organization of X-ray/EM combined studies [132]. The result of the fitting experiments is, in most cases, a new model (at “pseudo-atomic” resolution) that can be now deposited in the Protein Data Bank (PDB). But the detailed characterization of the whole experimental procedure will only be accomplished if the 3D-EM data, as well as the atomic models are fully described. This will be handled by the appropriate cross-references between the PDB and the EMD, the new infrastructure for the deposition of 3D-EM data.

#### **4.3.2. Architecture**

The Electron Microscopy Database (EMD) is being designed and developed from the very beginning to be fully integrated and compatible with the structural data in the PDB, enabling future tools and services to provide, when possible, structural information and knowledge regardless of the resolution level achieved by the experimental method.

The policy for new macromolecular structural data submissions to PDB and EMD will be the following:

- Atomic models obtained by high-resolution 3D-EM should be deposited in the PDB.
- Atomic models obtained by fitting atomic coordinate data into 3D-EM maps should be deposited in the PDB.

- 3D-EM maps used in fitting experiments should be deposited in the EMD database. Appropriate links between PDB models and 3D-EM maps in the EMD will be provided.

Therefore, the MSD data model will enable access to macromolecular structural data, regardless of its nature.

### 4.3.3. Phases and methodology

Working in collaboration with the MSD group at the European Bioinformatics Institute, and building on our previous results obtained during the design and development of the BioImage database (see chapter 3), we redesigned, extended and developed a new infrastructure for the organisation of 3D-EM structural data to be managed and publicly accessible at the EBI.

#### 4.3.3.a. Definition

The first stage of the development of the new database has been devoted to the definition of the relevant data and complementary information to be archived. Working on previous information compiled for the BioImage database, the Content and Preservation Description information for 3D-EM studies were compiled.

The main purpose of the EMD is to provide a central repository for 3D-EM maps, i.e. structural data reconstructed by 3D-EM, plus additional descriptive information (or meta-descriptors) and additional data files. One 3D-EM map corresponds to one EMD entry (i.e. a single accession code). Apart from 3D-EM maps, other complementary information will be stored:

1. Textual descriptors: Together with the 3D-EM maps, a set of textual annotations covering all aspects of the experimental procedure (from sample preparation, image acquisition and processing) and detailed description of the biological specimen being studied, as well as reference data (authors, bibliographic references, etc.) have been defined. Appropriate links to other biological databases have also been identified (e.g. NCBI taxonomy, Gene Ontology, InterPro). Meta-descriptors cover all areas needed to characterise the results of a 3D-EM experiment: the biological sample being studied, the experimental conditions (sample preparation, data acquisition and data processing), and the structural results in terms of a 3D-EM

map, as well as any administrative and reference data (such as bibliographic references). These descriptors have been categorised as “mandatory”, i.e. those that should always be provided by the author, or “optional”, i.e. those that would be desirable to be stored but the author may decide not to provide. This categorisation will give the authors a chance to choose the level of detail to describe the results of their experiments, while ensuring a common minimum description of the data in the database homogeneously.

2. Complementary data files: additional data files that might provide supplementary and relevant information on the experiment performed. These can be further classified as:
  - Supplementary figures, for illustrating important aspects of the resulting structures or the experimental intermediate data.
  - 3D surface data (masks), for iso-surface rendering purposes, provided as a binary map format.
  - Structure factors (only in crystallographic experiments) and layer line data (only in helical reconstructions). Sending these data to the EMD is optional, although we strongly encourage depositing them.

#### 4.3.3.b. Database design and integration

An entity-relationship model was created for 3D electron microscopy data using Oracle Designer 2000. This model was subsequently integrated with the existing model of atomic coordinate data in the MSD (containing over 400 tables that describe the results of experiments in NMR and X-ray crystallography). MSD data is maintained and managed at the EBI as a relational database implemented using Oracle database (Oracle Corporation [www.oracle.com](http://www.oracle.com)).

Appropriate relationships have been carefully analysed for those entities representing biological information, both in terms of integration with the MSD, as well as in the context of other relevant biological databases. This integration with already existing biological databases is essential in order to provide cross-references, and it is a valuable resource for establishing a common nomenclature by the adoption of widely used controlled vocabularies and ontologies.



Due to the special characteristics of 3D-EM structural data (i.e. maps), further analysis of requirements for map representation and storage were performed. Currently, there is not a standard format for 3D maps in the 3D-EM community: several proprietary volume formats are used by the different software packages for three-dimensional reconstruction by electron microscopy (e.g.: MRC, Brandeis, Duchy, Synu, EM, IVE, IMAGIC, BMD, PIC, SUPRIM, Semper, Spider, etc...). Nevertheless, a single 3D-EM map format has been adopted by the EMD: the CCP4 (Collaboratory Computing Project Number 4 for Protein Crystallography, Daresbury UK) map format [154] used in X-ray crystallography and electron microscopy domains.

#### 4.3.3.c. Development of EMD interfaces

An important aspect for the success of this kind of initiative, which is often neglected, is the interaction and collaboration with the scientific community that produces the data. It is essential to avoid any potential obstacle (either technical or sociological) in the way of the data from the author's laboratory to the database. At the end of the day, the value of a database is the value of the data it contains.

##### Data ingest

Appropriate tools for data conversion should be used during ingest in order to store and manage the 3D-EM maps homogeneously in the archive. The submission system converts uploaded map format to CCP4 by using Image Science's EM2EM map conversion utility (see <http://www.ImageScience.de/em2em/>).

##### Data dissemination

EMD data will be disseminated as a set of files: 3D-EM maps and complementary data files (e.g. CCP4 for 3D-EM files), while textual descriptors need the development of an XML file format. This XML file format is intended for data distribution and download, not for data management process. The EMD XML file format is described in terms of its corresponding XML Schema.

A release lock-in period can be placed on the 3D-EM map (up to 4 years) by the author, while the descriptive information will be immediately released (after it has been



reviewed by the authors). 3D-EM maps should be sent to the 3D-EM MSD for getting an accession code.

## **4.4. Results**

### **4.4.1. Electron Microscopy Database (EMD)**

In order to manage, organize and disseminate the data on the structure of macromolecules solved by 3D electron microscopy, the Electron Microscopy Database (EMD) has been set up at the European Bioinformatics Institute. The new public infrastructure provides a facility for storing 3D maps, relevant textual descriptor and complementary data files (supplementary figures, masks, structure factors and layer line data). Where applicable, the database also contains layer-line data and structure factor files. The deposition system has been active since June 2002 [155]. The EMD database can be accessed at [www.ebi.ac.uk/msd](http://www.ebi.ac.uk/msd)

#### **4.4.1.a. Integrated data model**

A first version of a fully integrated entity-relationship database model was formalised, containing the information on structural data and their experimental conditions. This model incorporates electron microscopy data to the previously existing X-ray diffraction and NMR data in the MSD. This unique data model (figure 7) is now fully part of the developments of the MSD (Macromolecule Structure Database) at the EBI.

#### **4.4.1.b. EMDep**

The EMDep is the web-based submission interface that provides the facility to deposit information to the Electron Microscopy Database (EMD) [50]. EMDep is a flexible and portable system, following a dictionary driven design that provides total separation of presentation and content. The page layout is defined in an interface definition XML dictionary. Parameter names are defined in the data XML schema. Help text is also defined in the dictionary. For those data items that require validation, the validation is also defined in the XML.

The content, or data entered by the submitter, is stored in a structured XML format, allowing the data to be accessed, read and modified. Submission data is stored

in XML until final annotation is completed and author approval is obtained. It is only at this stage that all relationships are known and entry information can be loaded into the MSD database.

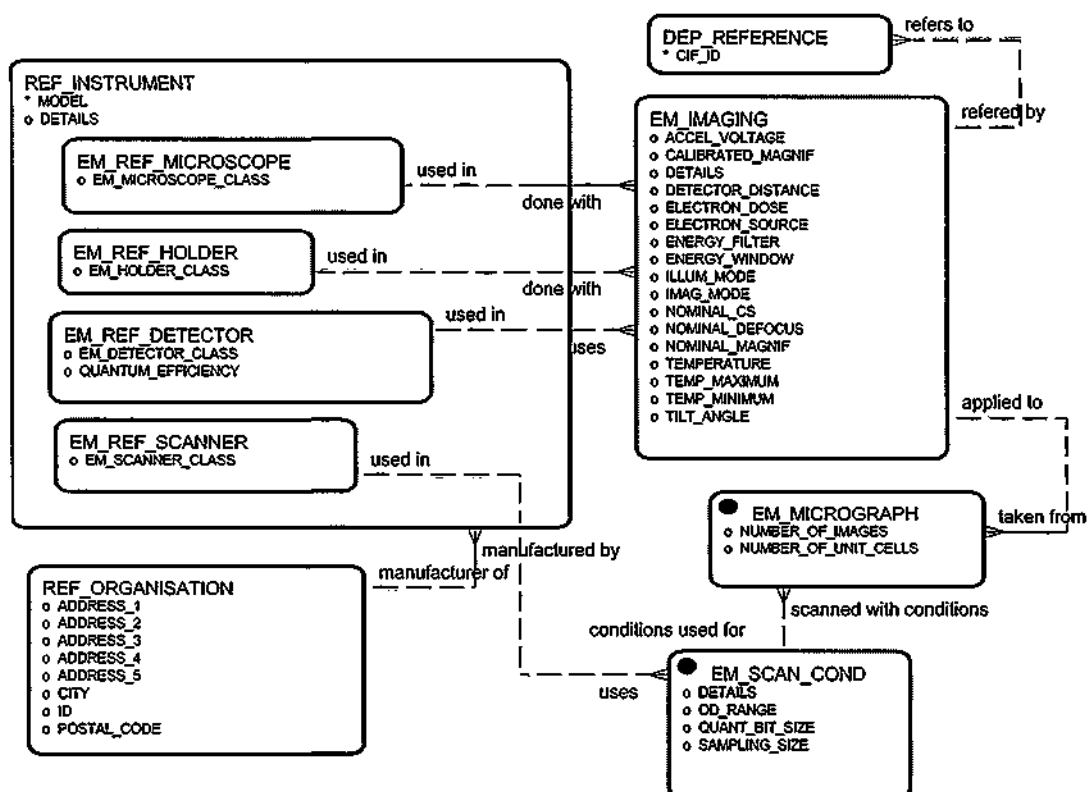


Figure 7: Some of the entities of the integrated entity-relationship model of the MSD are shown. Red lines correspond to EMD entities, while blue lines correspond to previously existing MSD entities.

EMDep was designed to be easily extensible. As EMDep was used in production, frequently encountered user errors and other user-requested extensions were simply addressed and corrected by editing the corresponding dictionaries.

The technology developed to create EMDep will be the base for the new version of AutoDep (the submission tool for PDB at the MSD).

#### 4.4.1.c. EMD XML file format

An EMD XML Schema was designed using XML Spy version 4.3 (Altova, [www.altova.com](http://www.altova.com)) for the definition of the EMD XML file format. XML format was chosen as it is becoming the *de facto* standard for the exchange of data on the World Wide Web.

Further details of the EMD XML file format can be found at <ftp://ftp.ebi.ac.uk/pub/databases/emdb/doc/XML-schema/>

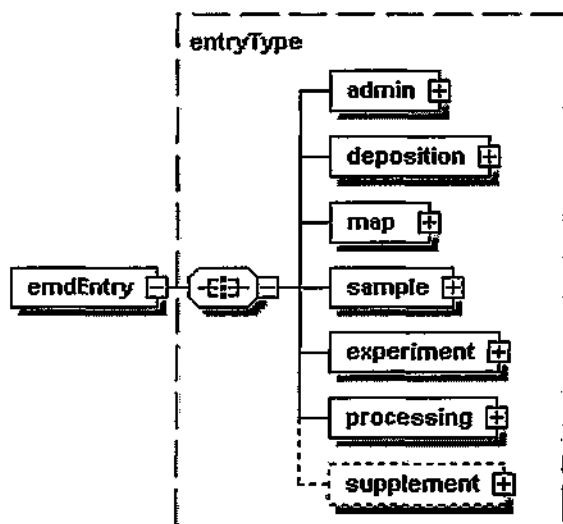


Figure 8: Top elements in the hierarchy of an EMD XML file as defined in the EMD Schema. Image produced with XML Spy (Altova).

#### 4.4.2. Additional results

In addition to the main objectives established in the current chapter, it is also relevant to mention some satellite work around the EMD that produced the following results.

##### 4.4.2.a. FEMME database

FEMME (Feature extraction in a multi-resolution macromolecular environment) is an infrastructure designed to collect topological and geometric information obtained from macromolecular structures solved by 3D-EM [51]. It is being populated with the analysis of data stored in the EMD using a novel implementation of the alpha-shape theory applied to image data [156, 157].

##### 4.4.2.b. Standardization of 3D-EM conventions

The development of the 3D-EM computational methodology in independent groups, as in many other disciplines, naturally implied the adoption of dissimilar data conventions in different software packages. This certainly is an obstacle found by those researches that make use of more than a software package, or even want to relate 3D-EM data with other structural information. The birth of a public repository of research results such as the EMD reinforces the need to define and adopt a consistent set of conventions in the field. Jointly with two other collaborators, we propose a set of common conventions

named the "3DEM Image Conventions" [158], designed as a standardized approach to image interpretation and presentation. The conventions would serve as a vehicle for data exchange among software packages and for long-term preservation of 3D-EM data in the EMD.

#### 4.4.2.c. EM at the PDB

Atomic models obtained by fitting atomic coordinate data into 3D-EM maps are deposited in the PDB under the "electron microscopy" technique category. In collaboration with the Research Collaboratory for Structural Bioinformatics (RCSB) a set of meta-descriptors for 3D-EM information relevant to PDB contents has been fully defined as an mmCIF dictionary<sup>3</sup>. New templates for the PDB were designed and with agreement with the RCSB have been adopted.

### 4.5. Discussion

The European Bioinformatics Institute (EBI) starts accepting 3D-EM data submissions from the scientific community in June 2002.

Participants in the "Workshop promoting software development in the field of high resolution electron microscopy" organized by the EBI in November 2002, provided their support to EMD as the public infrastructure to store structural information obtained by 3D-EM. *"We note that the European Bioinformatics Institute (EBI) through the Macromolecular Structure Database (MSD) now provides a permanent resource for the deposition of three-dimensional maps derived by electron microscopy [...]. In addition, coordinate data derived from these maps are deposited in*

<sup>3</sup> The mmCIF dictionary [159] P. E. Bourne, H. M. Berman, B. McMahon, K. D. Watenpaugh, J. D. Westbrook, and P. M. D. Fitzgerald, "Macromolecular crystallographic information file," *Method Enzymol*, vol. 277, pp. 571-590, 1997., based on the Self-defining Text Archive and Retrieval (STAR) format [160] S. R. Hall, "The Star File - a New Format for Electronic Data Transfer and Archiving," *Journal of Chemical Information and Computer Sciences*, vol. 31, pp. 326-333, 1991., manages crystallographic data on biological macromolecules. mmCIF is used by the wwPDB consortium to manage and exchange information on PDB data.

*the PDB archive for macromolecular structural data. We intend to use these facilities for the routine deposition of maps and coordinate data produced by our work. These databases are open to the international community and will become part of a family of linked databases in biomedical research.” [161]*

Additional support to ensure EMD data population is given by the open-access data policy of some scientific publications. In this line, the editorial *Nature Structural Biology* published in May 2003 (vol. 10, num. 5) stated: “*Nature Structural Biology is strongly supportive of the general principle that scientific data should be professionally maintained and freely accessible and so its editors will from now on encourage scientists to deposit their work in EMDB when papers describing EM structures are published in the journal.*”

Although having an appropriate tool such as EMDep for data deposition to the EMD is essential, it involves the existence of one person manually typing textual information in a Web form, which is a tedious and error-prone task to do. Taking into account that the final steps of the whole 3D-EM experiment are mostly driven by software, we envision that appropriate parameters and descriptors could be directly exported by programs, to be further imported into the EMD, as it is currently done in data-harvesting approaches with other experimental techniques (currently X-ray crystallography and NMR). This needs first some sort of standardization of the processing steps required to achieve the reconstruction of a 3D structural map from projection images, which at the current state-of-the-art in the 3D-EM is still not feasible, although might be achievable in the near future.

## *Chapter 5*    **Developing tools for biological data integration**

---

### **5.1. Introduction**

New experimental techniques and the advances on computational biology make possible to undertake genome-wide projects, as well as the study of complete molecular networks and whole protein families. The analysis of data coming from these techniques can only be done in the light of integrated information from diverse data sources, which can be described as heterogeneous, distributed and rapidly evolving.

In this context, it is difficult to develop a global bioinformatics infrastructure to assist scientists in their research. Each area of molecular biology generates its own public data repositories and a wide range of specialized query and analysis tools are commonly used over these resources. In addition, the wide use of high-throughput

technologies forces experimentalists to handle huge amounts of data within a single laboratory, and bioinformatics tools are also commonly used in a typical experimental setup.

This network of information services and sources available to the scientist forms a worldwide federation of autonomous, distributed, heterogeneous data repositories, which clearly demands information integration. By integrating data from so many sources (from the in-house experimental data to the accumulated knowledge in public archives, incorporating computational models and predictions), scientists will be able to identify correlations across the biological data spectrum from genomics to proteomics to drug design.

Existing technologies and practices in data management are new to a number of experimental laboratories and communities which did not face the need of efficient data handling until very recently. Meanwhile, several efforts towards the development of efficient tools for biological data integration have been done in the past, while other are yet in progress.

#### 5.1.1. Computational biology workflows

As in many different fields of science, a computational biology method of problem solving needs to interleave *information access* and *algorithm execution* by putting them together in a problem-specific “workflow”. In a complex domain like molecular biology, such a workflow often involves executing a number of algorithms where each algorithm may require access to multiple information sources (or multiple times to the same information source) to provide parameters at different steps of its execution.

The Workflow Reference Model [162] defines a workflow as the computerised facilitation or automation of a business process, in whole or part. A Workflow Management System (WMS) is a system that completely defines, manages and executes “workflows” through the execution of software whose order of execution is driven by a computer representation of the workflow logic.

Alimaki and coworkers [163] describes workflows as a set of tasks involved in a procedure along with their interdependencies and their inputs and outputs. In the traditional conceptualization of workflows, the focal point is the action, i.e., the



processes that take place during workflow execution. In this sense, workflows are considered as transactions, with the information that they manipulate playing a subordinate role. But in most scientific applications, the focal point is the information. In this sense workflows should be considered as graphs of objects, with the processes that created them having a secondary role.

Scientific applications can therefore be explained in terms of analytical workflows, and therefore are data-oriented, in contrast to traditional production workflows which can be defined as action-oriented. In this way, analytical workflows become then the 'business' processes in computational biology and bioinformatics domains.

### 5.1.2. Scientific workflow framework

We use the framework for defining scientific workflows presented in [164]. This framework handles workflows at two conceptual levels: abstract and executable. An abstract workflow is a network of abstract tasks and data which are semantically typed using concepts from an underlying application domain ontology. An executable workflow is a network of executable tasks and corresponding data.

A workflow graph (either abstract or executable) is a directed graph with the following types of nodes [164]:

- Task nodes: represent abstract functions or executable applications. The function signature of a task is determined by its data-in and data-out nodes.
- Data-in and data-out nodes: used to represent the input and output data of tasks. With each node we associate a semantic type and a syntactic type. When connecting tasks, semantic type checking and automatic data type conversion is performed, provided an appropriate type theory for the former, and conversion rules for the latter are given.
- Parameter nodes: represent parameters of tasks. A concrete function is obtained by instantiating the parameters.

## 5.2. Objectives

To design and develop a general purpose “programmable integrator” to create biological computational workflows that intersperse data access and algorithm execution.

## 5.3. Methods

Requirements for the programmable integrator were analysed by selecting a number of representative “study cases” in computational biology. This study cases helped us to define the characteristics of the data sources, as well as the type of relationship operations we wanted to perform. The architectural design and development of the programmable integrator was done by our collaborators in the San Diego Supercomputer Center. Once the programmable integrator engine was implemented, it was tested and used in the context of different computational workflows. Additional analysis in order to assess the possibilities of creating abstract workflows was finally accomplished.

## 5.4. Results

Our goal was to develop the technology to grant a user access to multiple information systems as though they were a single one with a uniform way to retrieve information and perform computations. The first complexity in achieving this goal is that the information sources are often independent and autonomous, have completely different schema structures and use different data formats. To provide uniform access, an integration system must therefore surmount the problem of data heterogeneity at the system, syntax and structural level.

### 5.4.1. Relationship operations in computational biology

Integrating data sources is all about discovering the associations and bridges among the different pieces of information, from the relationships between any two data sources (either archives, programs or ontologies) to the links between instances or subsets of instances in these data sources (Figure 9). We will define the following types of relationships:

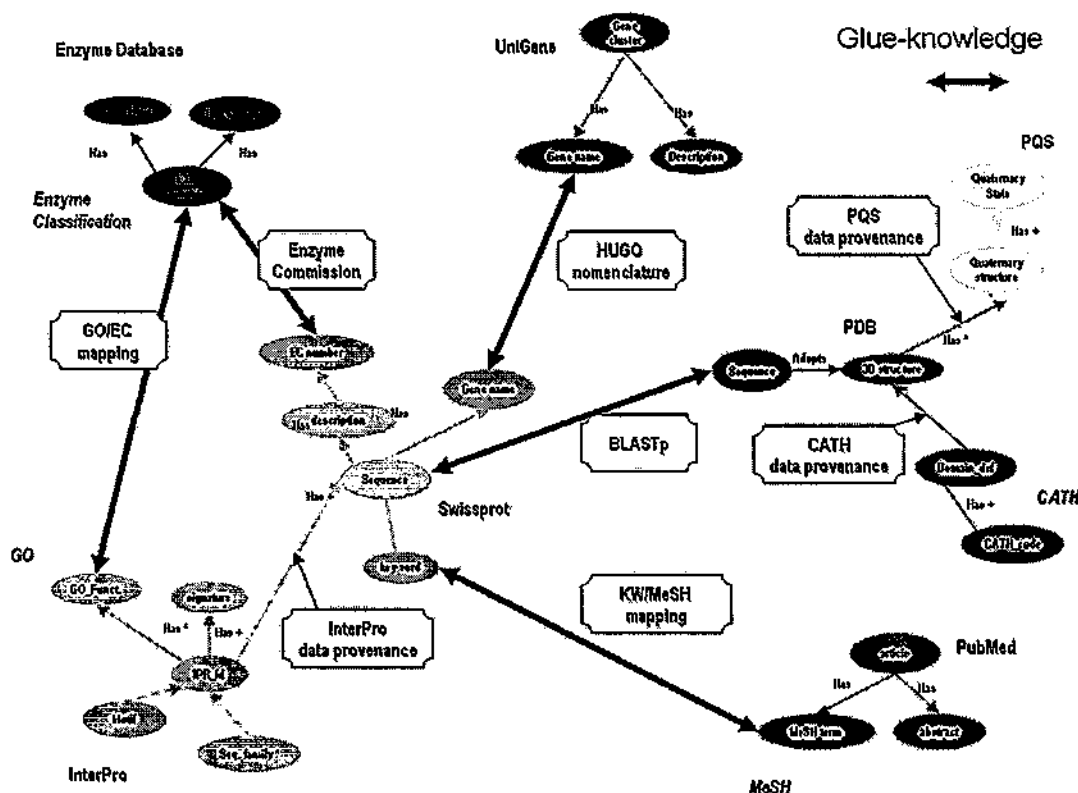


Figure 9: Relationships built among Swiss-Prot, InterPro, ENZYME DB, UniGene, PDB, PQS, CATH and PubMed data sources. Cross-references are used in associations of derived databases (namely InterPro from Swiss-Prot data analysis, PQS and CATH from PDB data analysis). Shared controlled vocabularies correspond to Enzyme Commission numbers annotated in both Swiss-Prot and ENZYME, and HUGO nomenclature in Swiss-Prot and UniGene. Ontology mappings are used between Gene Ontology and Enzyme Commission as well as in Swiss-Prot keywords and MESH terminology. Special joins are represented by the relationships built between Swiss-Prot and PDB using BLASTp sequence alignment.

- Cross-references: Joins created between two data sources making use of the foreign unique identifiers annotated. These are equivalent to the use of keys and foreign keys for creating associations between two tables in a relational data model. In the case of derived data sources, they can be resolved at the syntactic layer. Incorporation of additional information on how these sources relate to each other can be used during query evaluation and optimization (e.g. CATH contains only structures from PDB solved by X-ray diffraction whose resolution is better than 3.0 angstroms, together with structures solved by NMR).
- Special joins: Joins established by the use of a domain application (e.g.: relationships discovered between Swiss-Prot and PDB through a BLASTp search).
- Common nomenclature and shared vocabularies: There is a great awareness in the

Molecular Biology community on the need of creating standards to facilitate, among others, data sharing and interoperability. The adoption of these standards and controlled vocabularies by the data providers allows its use for data integration (e.g. HUGO nomenclature for human genes, and the Gene Ontology vocabulary for cellular components, biological processes and functions).

- Mapping vocabularies and ontologies: In some cases, common standards and/or vocabularies are not shared by two data sources, but some correspondences can be built between the annotations of the two (e.g. mapping of Gene Ontology “enzyme activity” GO:0003824 sub-tree to the Enzyme Classification).
- Joining by a third data source: Sometimes it is not possible to create a direct link between two pieces of information. Nevertheless, one of the advantages of incorporating more data sources to the integrated system is that transitive relationships can be used to relate two pieces for which no join exists (e.g. ENZYME database to PDB through the corresponding associations to Swiss-Prot).

#### 5.4.2. PLAN: a technology for integrative analysis

PLAN is a procedural programmable integrator suitable for the creation of bioinformatics workflows that resolves the obstacle of data heterogeneity among data sources. It allows the definition and execution of a cascade of data access, querying/filtering and algorithm invocation events (i.e. executable workflows).

PLAN is a simple XML-based language for the definition of workflows that simplifies data search and analysis by providing a uniform XML view on both data sources and analytical applications. The use of internal XML data structures is very reasonable, since an increasing number of data providers in the molecular biology domain offer the possibility of downloading information in XML format. Furthermore, many programs in the field also provide input/output XML-based mechanisms. In the case in which the data are not provided in XML format, a wrapper mechanism should be implemented to provide the necessary translation.

The overall PLAN architecture is shown in Figure 10. Briefly, the *resource catalogue* contains relevant information on registered sources; data retrieved from sources is translated to XML (if needed) by the appropriate wrapper, and temporary

stored in the *global buffer* for further filtering/querying; resulting data from each step are managed in the *global data table*; all data access/computation instructions are handled in the *execution stack*; data transfer between consecutive steps in a workflow is performed by invocation of the desired *named global data table* structures into the *global buffer*.

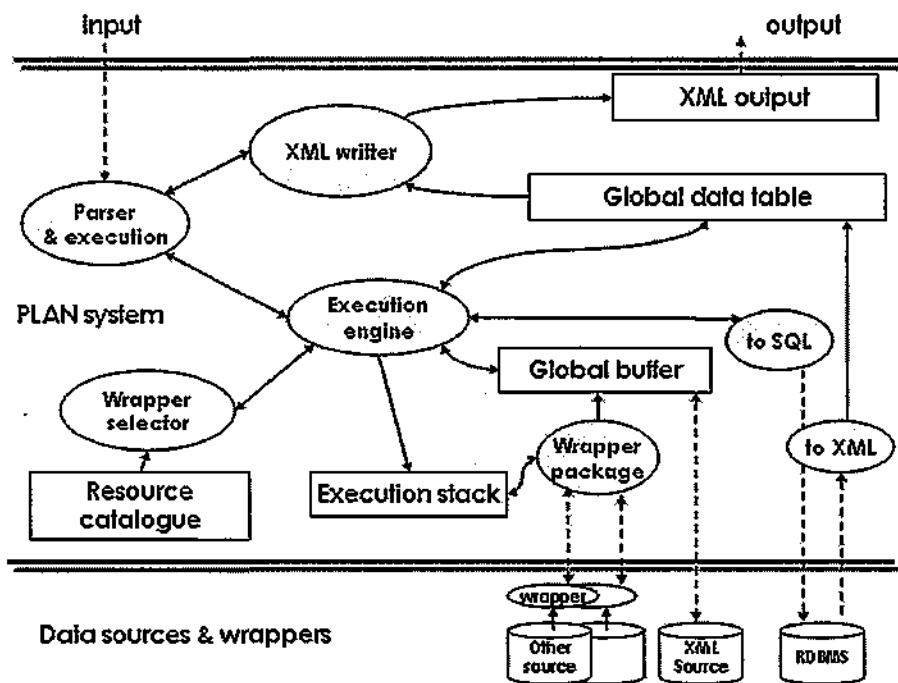


Figure 10: PLAN System Architecture (adapted from [114]). Oval shapes are processing units, while rectangles contain data structures.

A generic wrapper utility [165] automatically translates resulting data from relational database searches into XML format. In case the data source is neither XML nor relational, an external wrapper is needed. We used the Minerva wrapper toolkit as a freeware technology for accessing and transforming web pages [166]. In any case, the choice of a particular wrapper to transform sources does not affect either the PLAN language or the PLAN execution model. Full details of PLAN can be found in [114].

### 5.4.3. Solving computational workflows with PLAN

As the survey on bioinformatics tasks performed by [167] highlighted, moving data between repositories and analysis tools is of great importance when building

complex queries. Both data repositories and analysis tools are treated as data sources in our computational workflow paradigm.

Creation and execution of a workflow using PLAN requires:

- (1) Declaration of new data sources that are not already registered in the *resource catalogue*.
- (2) Definition of the process workflow using PLAN syntax and commands.
- (3) Execution of the workflow.

PLAN is highly flexible due to its modular design and programmable interface. It is designed to easily handle heterogeneous data sources (facilitated by the use of the resource catalogue and wrapping mechanisms for non-XML data sources), while providing powerful mechanisms for data integration and filtering (through the use of an internal XML data structure, a declarative query language and a procedural instruction set). These mechanisms allow the user to have full control over relevant parameters on which associations and filters are built.

## 5.5. Ready for semantic integration?

The current implementation of PLAN handles integration of data at the pure syntactic level, through the definition of executable workflows. Nevertheless it has been designed so as to work in a global infrastructure to support scientific workflows, whose architecture is described in [164]. In this section some of the requirements and difficulties in order to create such system are highlighted.

### 5.5.1.a. Data semantics

As explained before, scientific workflows can be defined as data-oriented. Therefore we should carefully consider the semantics of the information being handled in any analytical flow. Each data node should have a semantic type and a syntactic type.

Semantic types can be defined at two different levels: the basic semantic type of a data node is the type defined in the context of its data source (either database or application), i.e. the local semantic type. This local semantic type should therefore be independent of any particular use of the data. A second layer of semantics can be established at the abstract workflow level, i.e. at the level of the application ontology.

This application ontology is usually built around particular objectives (intended user analysis), and can consequently vary in different workflows.

How are the mappings of local semantic types to global semantic types (in the application domain ontology) done?

Local semantic types can be mapped to corresponding categories in the application ontology. This mapping is referred as contextualization in [168]. E.g. Swiss-Prot keywords (syntax: swissprot-keyword, data type: string) can be semantically typed using corresponding 'categories' as defined by the Uniprot Knowledgebase [169]. In this way 'helicase' and 'hydrolase' are categorised as "molecular function, enzyme"; 'hereditary haemolytic anemia' and 'Alzheimer's disease' are categorised as "disease"; 'SH3 domain' and 'Zinc-finger' are categorised as "domain"; etc. The Swiss-Prot keyword categories provide a local semantic characterization that can be imported as such or mapped to corresponding terms in the application ontology.

It is worth noticing that local semantic types should be used for constraining connections (by semantic type checking), while global semantics will be normally used for building application domain rules that may not be present in the underlying data sources (databases or methods).

Local semantic types should be linked to their corresponding syntactic type(s) (a process known as ontological grounding). Some data sources can provide more than one syntactic type for each semantic class. Usually this will be the case of integrated data collections, such as InterPro. E.g. (see Figure 11): The unified InterPro (IPR001623) 'Heat shock protein DnaJ, N-terminal' domain is syntactically expressed through five signatures as: (PF00226) DnaJ in Pfam; (PS00636) DNAJ\_1 and (PS50076) DNAJ\_2 in Prosite; (SM00271) DnaJ in Smart; (SSF46565) DnaJ\_N in Superfamily.

Although a sequence search to InterPro (through InterProScan service) will return a list of matched InterPro family/domains, the recognition of such family/domains is done through the mapping of underlying sequence signatures (InterProScan is in fact a query to a multidatabase using local query mechanisms, namely: BlastProDom, FPrintScan, HMMPfIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, ScanRegExp, SumexFamily). While InterPro offers an integrated consolidated view on protein families and domains and a whole analysis of Swiss-Prot/TrEMBL sequence databases,



it does not provide a unified syntax for non-instantiated sequence patterns (nor even a 'local' copy of them). The syntax and data of sequence patterns (or signatures) should be obtained from the underlying databases.

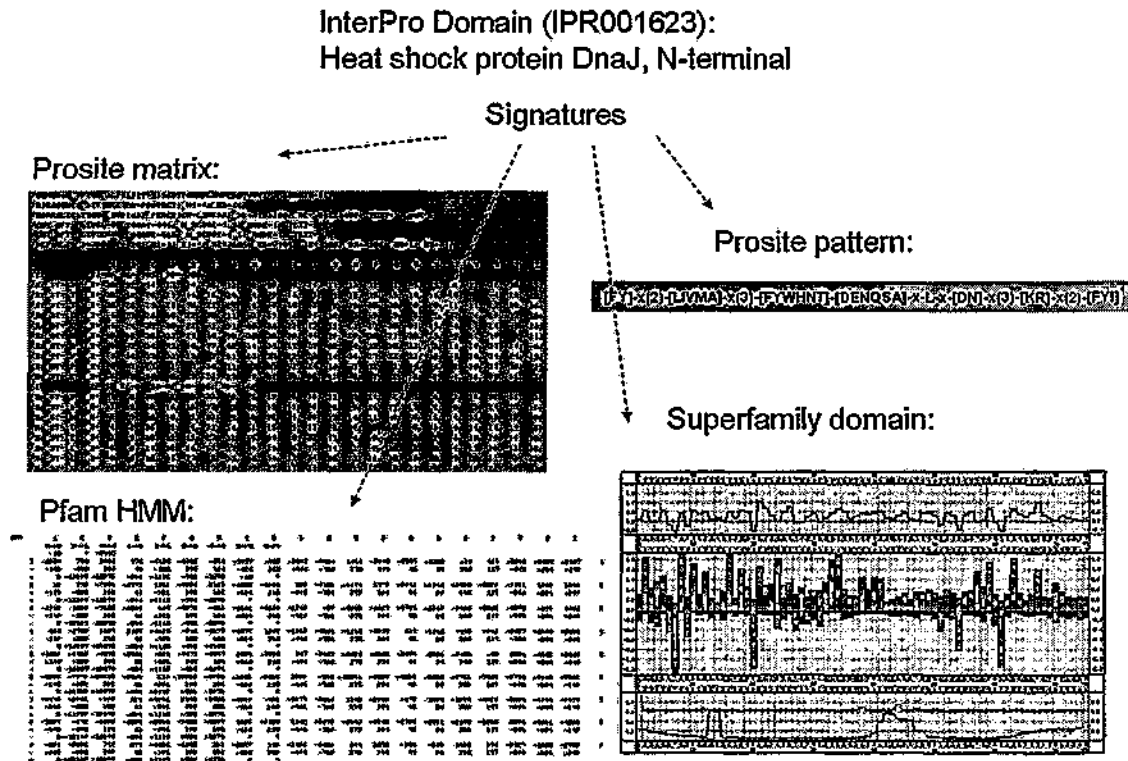


Figure 11: Illustration of some of the domain signatures (syntax types) corresponding to the InterPro 'Heat shock protein DnaJ, N-terminal' domain.

### 5.5.1.b. Task semantics

Even if there is a great diversity and still growing number of specialised applications in molecular biology, I will analyse just a single and widely used bioinformatics application in order to show the complexity of handling and designing a semantic layer around tasks. The analysis is done in the context of the two example workflows used as motivating examples in [114].

The executable task to consider is the BLAST search against a protein sequence collection [22] (it can be invoked as a local application, an HTTP request or even a Web service). The abstract task can be described as "perform a protein-protein sequence similarity search". Such abstraction allows the definition of more than one semantic-to-syntax (or abstract-to-executable) mapping for the "protein-protein similarity search".



E.g. instead of using the BLAST algorithm, a FASTA search could be used as an alternative method.

What is more, the semantics of an abstract task are in fact modulated by its parameter and data-in nodes as well as the conditions established upon data-out nodes. In the two example workflows, the “protein-protein similarity search” was in fact used for at least four different purposes:

- Feature mapping: transfer of sequence annotations (in this case a sequence domain) to three-dimensional structures, through the evaluation of corresponding aligned segments. Search sequence collection: PDB (contains structures).
- Group sequences by ‘protein’: sequences corresponding to the same protein are recognised. Search sequence collection: Swiss-Prot (representing the known non-redundant protein sequence space).
- Find numerous distant homologues: distant homologues are defined as those with sequence identity in the range 30-70%. Numerous, at least 10. Search sequence collection: Swiss-Prot.
- Does it correspond to a full wild-type protein?: full (non-fragment) protein, wild-type (non-mutant) protein are evaluated through a search in Swiss-Prot (as containing full, wild-type sequences). Full is assessed by comparing query and hit sequence lengths; wild-type as having 100% sequence identity.

Therefore, a “canonical” abstract task such as ‘protein-protein similarity search’ can in fact represent a great number of higher-level tasks, depending on the context in which it is used (being its context the sum of data-in and parameters nodes and data-out evaluation).

The creation of hierarchical task structures from executable workflows to the definition of more abstract task should therefore consider not only the canonical semantics of executable workflows, but its intended use. This contextual use of a particular executable task results in different abstract functions due to:

- Fixation of parameters (e.g. perform a search in a particular database). The fixed parameter is not lifted to the higher-level task.

- Constraining data-in semantic types (e.g. query sequence of proteins with known three-dimensional structure). In this case, syntactic types remain, although the semantic type attached to data-in node changes.
- Evaluation of data-out nodes (e.g. filtering output information upon a condition, such 100% similarity search).
- Selection of subsets of data-out information (e.g. keeping only hit unique identifiers).

For the evaluation and selection of subsets in data-out nodes some declarative query mechanism on syntactic data is needed (such as the one provided by PLAN). This will involve changes in data-out syntactic as well as semantic types.

Therefore any abstract-as-view definition (i.e. the definition of an abstract task in terms of underlying executable tasks) will generally imply a redefinition of the syntax and semantics of data-in and data-out, and a subset of underlying parameters.

## 5.6. Discussion

Our approximation to the integrative data analysis in molecular biology can be classified as “process-oriented” (or business process integration as described by Linthicum [39]), involving both information (data) and methods (application or services). In process-oriented integration, relationships between data sources are built on-demand. Thus, there is no need to design and provide a universal integrated view on the component sources.

In most cases, users know which are the relevant information sources and applications that should be used in a particular analysis. Although they might not be aware of the exact schema of every data source relevant to their analysis, they are experts in the content of the data source, as well as the nature and semantics of the data, quality, etc... They also know the connections they want to build between data, as well as the applications for data transformation required in order to establish those connections.

Some other characteristics of workflow-based solutions include:

- it is possible to support multiple semantic mappings (one per user or use), that might not be anticipated by the system integrator,
- users can exploit their knowledge on data sources to specify workflows achieving a good performance,
- in some cases, transparency of data location (at the level of data collection) might not be desirable, due to trust and data quality issues,
- procedural languages are easier to learn (and closer to the way analysis are made),
- there is no need to maintain a pre-defined integrated schema over the available sources. Several integrations are possible corresponding to user applications and demands.

The main strength of PLAN is to significantly reduce the complexity of information integration from multiple sources. To do so, PLAN combines a declarative query language with the additional power of a procedural instruction set using a uniform and easy to manipulate XML format. Information can be kept in its original location and accessed only during run time. Only a resource catalogue defining access mechanisms and properties of the data is required. PLAN can be easily extended allowing the incorporation of additional data sources by registration to the catalogue. Available sources in the catalogue can seamlessly be used together in a computational workflow. The use of a declarative query language allows filtering operations on data, as well as any other complex queries provided by XQuery. Custom user defined functions can be easily added to be used in the query language.

Workflow approaches are not the single paradigm for biological data integration having a process-oriented focus. For example, an alternative approach is the one followed by HyBrow (Hypothesis Browser) [115], a tool for designing hypotheses and evaluating them for consistency with existing knowledge. In this case, the processes modelled, instead of analytical workflows, correspond to the traditional scientific method of working around hypotheses.



Todo aquello había sido una forma de sintaxis, un modo de ordenación de la realidad quizá no menos arbitraria que la alfabética.

Juan José Millás, *"El orden alfabético"*

## Discussion

---

Software projects are, in many cases, only a part of a wider project. I have worked in scientific research environments, where creativity and innovation are essential and impregnate other aspects of the project. In such context, software design and development is difficult to control and restrict to strict methodologies. This is partially due to the evolving nature of research activities that forces technological developments to change with the definition of the project itself. From my experience, best development methodologies for research projects involving the creation of databases or software applications are those following an incremental approach, and very important, working with prototypes.

Development of systems to support the preservation and organization of scientific archives is not just a technological endeavour. It is a research mission involving inventiveness as well as group dynamics and culture. Close interactions with data producers in the designing phases are essential, but also difficult. Furthermore,

considerable knowledge and understanding of the specific research field is needed to design and implement an appropriate software system.

In addition to integration challenges found in many business environments, integration in molecular biology poses an additional level of complexity due to the nature of biological data. These particularities have to be taken into account in order to avoid the development of elegant systems from the technical point of view, but lacking appropriate functionality to be used in practical applications. Attention should also be taken to solutions originally created to fulfil short-term goals for specific purposes, which may not be scalable or maintainable in the future as they have been designed as crafted products.

The complexity of the subject at hand clearly demands an interdisciplinary work. As noticed in [15]: *“Orchestrating fruitful interdisciplinary research across biology and data management is not easy. Lack of sufficient interaction between biologists and data management researchers can easily lead to attempts to reinvent well-known data management technologies by bioinformaticists, or sterile pursuits of irrelevant (or misunderstood) problems by data management researchers. For fastest progress in the biological sciences, we must encourage both the development of content for biological databases as well as data management technology for managing this content”*

Previous and ongoing initiatives in molecular biology to facilitate integrative data analysis can be grouped as those aiming at providing better means of data source interoperability, and those developing generic software systems. Among the last, it is worth noticing two trend directions. First, there is an increasing awareness of sharing not only information, but also applications with recent developments around service-oriented approaches (motivated by Web service technologies). Second, the semantic paradigm is also gaining acceptance (once again parallel to the creation of the “semantic Web”).

At this time it is worth questioning if these two directions are appropriate. Service-oriented approaches are suitable to share methods and/or algorithms, but they fall short if they do not take into account the need to also share information. Consequently, they should be complemented with means of data integration in the case that data standardization is not guaranteed, as in many applications in molecular

biology. Thus, it seems that process-oriented solutions are more appropriate, as they consider both data and method integration. Among these, workflow and hypothesis-building paradigms seem to fit smoothly with applications in molecular biology.

The second question to answer is whether molecular biology (and related scientific domains) is ready for transparent semantic interoperability. My answer is “not yet”. Data semantics in molecular biology are either not well known and/or not properly specified. For semantic interoperability to be real and ubiquitous in biology, clear specification of semantics must happen. Although some formal ontologies are emerging, they are normally used for annotation purposes, not for describing data models (with exceptions as noted in chapter 2).

In addition, complexity of molecular biology data makes that granularity of data stored in databases may not be the appropriate granularity to represent biological information in a given application. Thus, the process of defining and establishing semantic correspondences among data sources and application domains will require the intensive use of semantic transformations. Furthermore, most computational biology applications are designed for the discovery of new information out of biological data. While applications in an “operational mode” can be formalized and created around predefined semantic models, applications working in a “discovery mode” are less suitable to express and share semantic conceptualizations.





---

## Conclusions

---

After some years of research in this field, I have reached the following conclusions. A general insight is that, in spite of the conceptual complexity of the biological data integration problem, the main bottlenecks to achieve it are still found at very practical and technical levels. There are three reasons for this. First, there is a poor selectivity of searching mechanisms in many databases. Second, an important number of data is represented using complex data types. These complex data usually lack appropriate native operations and standard interfaces which are normally available for more common data types. Third, there is a need of using a wide range of data operations and transformations due to the lack of standards.

As of today, there is not a unique best solution for the task of providing systems to facilitate integrative biological data analysis. The work presented in this thesis has illustrated that:

- As a first step it is essential to create publicly accessible biological databases. Our contribution to this aim has been the establishment of the Electron Microscopy



Database (EMD) as the world-wide public archive to store structural data obtained by 3D-EM.

- There is a need to develop new databases providing infrastructures to support access to heterogeneous data. This is the case of work on the BioImage database, designed to store and manage multidimensional images of biological specimens obtained from various classes of microscopy techniques.
- It is necessary to create federated infrastructures to relate data collections. In the case of macromolecular structural data this has been achieved by establishing correspondences between the atomic models stored in the Protein Data Bank (PDB) and three-dimensional maps in the EMD.
- Consolidated access to biological information can be accomplished through the creation of integrated data models underlying data warehouses. In this line, we extended the Macromolecular Structure Database (MSD) to contain electron microscopy data.
- An important aspect to ensure better means of interoperability is to supply appropriate means of data citation, helping to provide reliable mechanisms for data provenance tracking. These mechanisms are essential when creating derived data infrastructures as the FEMME database built on the analysis of EMD data.
- Integrative data analysis will benefit from the developments of generic systems to share processes suitable for molecular biology research, as our proposal for the construction of computational workflows with PLAN.
- Finally, standardization in well-delimited areas of research will enhance the interoperability of software platforms, as well as the exchange of data. In this line, we have launched the initiative towards the establishment of common conventions enabling data interchange among the 3D-EM field.

## Conclusiones

---

Después de algunos años de investigación en este campo, se pueden extraer una serie de conclusiones. Quizá la más evidente es que, a pesar de la complejidad conceptual del problema de la integración de datos biológicos, los principales cuellos de botella se encuentran a niveles prácticos. En primer lugar, muchas bases de datos proporcionan mecanismos de búsqueda con selectividad limitada. En segundo lugar, un importante número de datos se representa con tipos de datos complejos, que carecen de las operaciones nativas e interfaces estándares normalmente disponibles para tipos de datos más comunes. Finalmente, es necesaria la utilización de un abanico amplio de operaciones y transformaciones sobre los datos debido, en gran parte, a la escasez de estándares.

A día de hoy, no existe una única solución tecnológica ganadora en el conjunto de soluciones que posibilitan el análisis integrado de datos biológicos. El trabajo realizado en ésta tesis ilustra las siguientes conclusiones:

- Como punto de partida, es esencial crear bases de datos biológicas accesibles públicamente. Nuestra contribución hacia dicho objetivo ha sido el

establecimiento de "Electron Microscopy Database" (EMD) como archivo público de ámbito mundial para almacenar datos estructurales obtenidos por microscopía electrónica tridimensional.

- Es preciso el desarrollo de nuevas bases de datos que proporcionen infraestructuras para el acceso a datos heterogéneos. Es el caso de BioImage, diseñada para el almacenamiento y la gestión de imágenes multi-dimensionales de especímenes biológicos obtenidas mediante diversas técnicas de microscopía.
- Es necesaria la creación de infraestructuras federadas que permitan relacionar distintas colecciones de datos. En el caso de datos estructurales de macromoléculas biológicas ésta se ha llevado a cabo mediante el establecimiento de las correspondencias necesarias entre los modelos atómicos almacenados en la "Protein Data Bank (PDB)" y los mapas tridimensionales de EMD.
- El acceso consolidado a información biológica puede llevarse a cabo mediante la creación de modelos de datos integrados que soporten "data warehouses". En este sentido, la "Macromolecular Structure Database (MSD)" se ha extendido para gestionar también datos de EMD.
- Un aspecto importante para asegurar mejores formas de interoperabilidad es proporcionar medios adecuados para la cita de datos, que ayuden a su vez a proporcionar mecanismos fiables para la procedencia de datos ("data provenance"). Estos mecanismos son esenciales cuando se crean infraestructuras de datos derivadas como la base de datos FEMME construida a partir del análisis de los datos almacenados en EMD.
- El análisis integrado de datos puede beneficiarse de los sistemas genéricos desarrollados para compartir procesos computacionales apropiados para soportar la investigación en biología molecular. Es el caso PLAN, nuestra propuesta para la creación de flujos de trabajo computacionales.
- Por último, la estandarización en áreas de investigación bien delimitadas mejorarán la interoperabilidad de las aplicaciones software. En esta dirección se ha lanzado una iniciativa para establecer convenciones comunes que

faciliten el intercambio de datos en las aplicaciones de microscopía electrónica tridimensional.



## Bibliography

---

- [1] M. E. Williams, "Electronic databases," *Science*, vol. 228, pp. 445-50, 455-6, 1985.
- [2] S. E. Fienberg, Martin, M.E., Straf, M.L., "Sharing research data," The National Academies Press, 1985, pp. 240.
- [3] T. D. Sterling and J. J. Weinkam, "Sharing scientific data," *Communications of the ACM*, vol. 33, pp. 112-119, 1990.
- [4] J. C. French, A. K. Jones, and J. L. Pfaltz, "Scientific Database Management (Final Report)," University of Virginia CS-90-21, August 1990.
- [5] S. M. Maurer and S. Scotchmer, "Database protection: is it broken and should we fix it?," *Science*, vol. 284, pp. 1129-30, 1999.
- [6] G. M. Ramsey and E. A. Howard, "Databases in the biological sciences. A user's guide to the current copyright landscape," *Plant Physiol*, vol. 132, pp. 1131-4, 2003.
- [7] L. B. Ellis and D. Kalumbi, "The demise of public data on the web?," *Nat Biotechnol*, vol. 16, pp. 1323-4, 1998.
- [8] L. B. Ellis and D. Kalumbi, "Financing a future for public biological data," *Bioinformatics*, vol. 15, pp. 717-22, 1999.

- [9] CCDS, "Reference Model for an Open Archival Information System (OAIS)," Consultative Committee for Space Data Systems (CCSDS) CCSDS 650.0-B-1, January 2002 2002.
- [10] E. S. Lander, R. Langridge, and D. M. Saccocio, "Computing in molecular biology: mapping and interpreting biological information," *Computer*, vol. 24, pp. 6-13, 1991.
- [11] J. C. Wooley, "Trends in computational biology: a summary based on a RECOMB plenary lecture, 1999," *J Comput Biol*, vol. 6, pp. 459-74, 1999.
- [12] D. Frishman, K. Heumann, A. Lesk, and H. W. Mewes, "Comprehensive, comprehensible, distributed and intelligent databases: current status," *Bioinformatics*, vol. 14, pp. 551-61, 1998.
- [13] CODATA, "Quality control in databanks for molecular biology," *Bioessays*, vol. 22, pp. 1024-34, 2000.
- [14] J. D. Navarro, V. Niranjana, S. Peri, C. K. Jonnalagadda, and A. Pandey, "From biological databases to platforms for biomedical discovery," *Trends Biotechnol*, vol. 21, pp. 263-8, 2003.
- [15] H. V. Jagadish and F. Olken, "Report of the NSF/NLM Workshop on Data Management for Molecular and Cell Biology," LBNL-52767, November 4, 2003 2003.
- [16] C. Sansom, "Database searching with DNA and protein sequences: an introduction," *Brief Bioinform*, vol. 1, pp. 22-32, 2000.
- [17] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert, "Approaches to the automatic discovery of patterns in biosequences," *J Comput Biol*, vol. 5, pp. 279-305, 1998.
- [18] I. Eidhammer, I. Jonassen, and W. R. Taylor, "Structure comparison and structure patterns," *J Comput Biol*, vol. 7, pp. 685-716, 2000.
- [19] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science*, vol. 294, pp. 93-6, 2001.
- [20] P. R. Graves and T. A. Haystead, "Molecular biologist's guide to proteomics," *Microbiol Mol Biol Rev*, vol. 66, pp. 39-63; table of contents, 2002.
- [21] J. Y. Chen and J. V. Carlis, "Similar\_Join: extending DBMS with a bio-specific operator," presented at 2003 ACM Symposium on Applied Computing, Mailborune, Florida, 2003.
- [22] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, pp. 3389-402, 1997.



- [23] L. D. Stein, "Integrating biological databases," *Nat Rev Genet*, vol. 4, pp. 337-45, 2003.
- [24] B. A. Eckman, A. S. Kosky, and L. A. Laroco, Jr., "Extending traditional query-based integration approaches for functional characterization of post-genomic data," *Bioinformatics*, vol. 17, pp. 587-601, 2001.
- [25] U. Leser, H. Lehrach, and H. Roest Crolius, "Issues in developing integrated genomic databases and application to the human X chromosome," *Bioinformatics*, vol. 14, pp. 583-90, 1998.
- [26] T. Kazic, "Semiotics: a semantics for sharing," *Bioinformatics*, vol. 16, pp. 1129-44, 2000.
- [27] B. Eckman, W. Rice, and W. Swope, "Heterogeneous Data and Algorithm Integration in Bioinformatics," in *ISMB 2002 Tutorial*, 2002.
- [28] R. Hull, "Managing semantic heterogeneity in databases: a theoretical prospective," presented at 17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of Database Systems, Tucson, Arizona, USA, 1997.
- [29] V. M. Markowitz, I. M. Chen, A. S. Kosky, and E. Szeto, "Facilities for exploring molecular biology databases on the Web: a comparative study," *Pac Symp Biocomput*, pp. 256-67, 1997.
- [30] P. Buneman, S. Khanna, and W. C. Tan, "Data provenance: some basic issues," presented at 20th Conf. Foundations of Software Technology and Theoretical Computer Science, New Delhi, India, 2000.
- [31] S. Lindek, R. Fritsch, J. Machtynger, P. A. de Alarcon, and M. Chagoyen, "Design and realization of an on-line database for multidimensional microscopic images of biological specimens," *J Struct Biol*, vol. 125, pp. 103-11, 1999.
- [32] J. M. Carazo and E. H. K. Stelzer, "The BioImage database project: organizing multidimensional biological images in an object-relational database," *J Struct Biol*, vol. 125, pp. 97-102, 1999.
- [33] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, and E. M. Zdobnov, "The InterPro database, an integrated documentation resource for protein families, domains and functional sites," *Nucleic Acids Res*, vol. 29, pp. 37-40, 2001.
- [34] W. S. Valdar and J. M. Thornton, "Protein-protein interfaces: analysis of amino acid conservation in homodimers," *Proteins*, vol. 42, pp. 108-24, 2001.
- [35] P. D. Karp, "Database links are a foundation for interoperability," *Trends Biotechnol*, vol. 14, pp. 273-9, 1996.

- [36] F. Achard, C. Cussat-Blanc, E. Viara, and E. Barillot, "The new Virgil database: a service of rich links," *Bioinformatics*, vol. 14, pp. 342-8, 1998.
- [37] F. Achard and P. Dessen, "GenXref. VI: Automatic generation of links between two heterogeneous databases," *Bioinformatics*, vol. 14, pp. 20-4, 1998.
- [38] J. M. Ostell, "Integrated access to heterogeneous data from NCBI," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 14, pp. 730-736, 1995.
- [39] D. S. Linthicum, *Next Generation Application Integration: From Simple Information to Web Services*, August 15, 2003 ed: Addison Wesley, 2003.
- [40] J. Widom, "Integrating Heterogeneous Databases: Lazy or Eager?," *ACM Computing Surveys*, vol. 28, pp. article 91, 1996.
- [41] M. Stonebraker, "Too much middleware," *ACM SIGMOD Record*, vol. 31, pp. 97-106, 2002.
- [42] W. H. Inmon, *Building the Data Warehouse*: John Wiley and Sons, 1992.
- [43] G. Wiederhold, "Mediators in the architecture of future information systems," *Computer*, vol. 25, pp. 38-49, 1992.
- [44] A. Sheth, "Changing focus on interoperability in information systems: from system, syntax, structure to semantics," in *Interoperating Geographic Information Systems*, M. F. Goodchild, M. J. Egenhofer, R. Fegeas, and C. A. Kowwman, Eds.: Kluwer, 1998, pp. 25.
- [45] D. Heimbigner and D. McLeod, "A federated architecture for information management," *ACM Transactions on Information Systems (TOIS)*, vol. 3, pp. 253-278, 1985.
- [46] A. P. Sheth and J. A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Computing Surveys*, vol. 22, pp. 183-236, 1990.
- [47] H. Recipon and W. Makalowski, "The biologist and the World Wide Web: an overview of the search engines technology, current status and future perspectives," *Curr Opin Biotechnol*, vol. 8, pp. 115-8, 1997.
- [48] L. Wang, J. J. Riethoven, and A. Robinson, "XEMBL: distributing EMBL data in XML format," *Bioinformatics*, vol. 18, pp. 1147-8, 2002.
- [49] S. Miyazaki, H. Sugawara, T. Gojobori, and Y. Tateno, "DNA Data Bank of Japan (DDBJ) in XML," *Nucleic Acids Res*, vol. 31, pp. 13-6, 2003.
- [50] K. Henrick, R. Newman, M. Tagari, and M. Chagoyen, "EMDep: a web-based system for the deposition and validation of high-resolution electron microscopy macromolecular structural information," *J Struct Biol*, vol. 144, pp. 228-37, 2003.

- [51] N. Jimenez-Lozano, M. Chagoyen, J. Cuenca-Alba, and J. M. Carazo, "FEMME database: topologic and geometric information of macromolecules," *J Struct Biol*, vol. 144, pp. 104-13, 2003.
- [52] E. Barillot and F. Achard, "XML: a lingua franca for science?," *Trends Biotechnol*, vol. 18, pp. 331-3, 2000.
- [53] F. Achard, G. Vaysseix, and E. Barillot, "XML, bioinformatics and data integration," *Bioinformatics*, vol. 17, pp. 115-25, 2001.
- [54] Y. Huang, T. Ni, L. Zhou, and S. Su, "JXP4BIGI: a generalized, Java XML-based approach for biological information gathering and integration," *Bioinformatics*, vol. 19, pp. 2351-8, 2003.
- [55] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: the European Molecular Biology Open Software Suite," *Trends Genet*, vol. 16, pp. 276-7, 2000.
- [56] T. Carver and A. Bleasby, "The design of Jemboss: a graphical user interface to EMBOSS," *Bioinformatics*, vol. 19, pp. 1837-43, 2003.
- [57] CCP4, "The CCP4 suite: programs for protein crystallography," *Acta Crystallographica Section D*, vol. 50, pp. 760-763, 1994.
- [58] R. Fogh, J. Ionides, E. Ulrich, W. Boucher, W. Vranken, J. P. Linge, M. Habeck, W. Rieping, T. N. Bhat, J. Westbrook, K. Henrick, G. Gilliland, H. Berman, J. Thornton, M. Nilges, J. Markley, and E. Laue, "The CCPN project: an interim report on a data model for the NMR community," *Nat Struct Biol*, vol. 9, pp. 416-8, 2002.
- [59] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. Doyle, and H. Kitano, "The ERATO Systems Biology Workbench: enabling interaction and exchange between software tools for computational biology," *Pac Symp Biocomput*, pp. 450-61, 2002.
- [60] M. R. Pocock, T. Hubbard, and E. Birney, "SPEM: a parser for EMBL style flat file database entries," *Bioinformatics*, vol. 14, pp. 823-4, 1998.
- [61] H. Hermjakob, W. Fleischmann, and R. Apweiler, "Swissknife - 'lazy parsing' of SWISS-PROT entries," *Bioinformatics*, vol. 15, pp. 771-2, 1999.
- [62] C. Ramu, C. Gemund, and T. J. Gibson, "Object-oriented parsing of biological databases with Python," *Bioinformatics*, vol. 16, pp. 628-38, 2000.
- [63] J. Hu, C. Mungall, D. Nicholson, and A. L. Archibald, "Design and implementation of a CORBA-based genome mapping system prototype," *Bioinformatics*, vol. 14, pp. 112-20, 1998.
- [64] T. Coupaye, "Wrapping SRS with CORBA: from textual data to distributed objects," *Bioinformatics*, vol. 15, pp. 333-8, 1999.



- [65] S. Z. Maltchenko, "The bio-objects project. Part I: the object data model core elements," *Bioinformatics*, vol. 14, pp. 479-85, 1998.
- [66] L. Wang, P. Rodriguez-Tome, N. Redaschi, P. McNeil, A. Robinson, and P. Lijnzaad, "Accessing and distributing EMBL data using CORBA (common object request broker architecture)," *Genome Biol*, vol. 1, pp. RESEARCH0010, 2000.
- [67] R. Stevens and C. Miller, "Wrapping and interoperating bioinformatics resources using CORBA," *Brief Bioinform*, vol. 1, pp. 9-21, 2000.
- [68] S. Fischer, J. Crabtree, B. Brunk, M. Gibson, and G. C. Overton, "bioWidgets: data interaction components for genomics," *Bioinformatics*, vol. 15, pp. 837-46, 1999.
- [69] R. D. Stevens, A. J. Robinson, and C. A. Goble, "myGrid: personalised bioinformatics on the information grid," *Bioinformatics*, vol. 19 Suppl 1, pp. i302-4, 2003.
- [70] M. D. Wilkinson and M. Links, "BioMOBY: an open source biological web services proposal," *Brief Bioinform*, vol. 3, pp. 331-41, 2002.
- [71] H. Sugawara and S. Miyazaki, "Biological SOAP servers and web services provided by the public sequence data bank," *Nucleic Acids Res*, vol. 31, pp. 3836-9, 2003.
- [72] P. Riikonen, J. Boberg, T. Salakoski, and M. Vihinen, "Mobile access to biological databases on the Internet," *Biomedical Engineering, IEEE Transactions on*, vol. 49, pp. 1477-1479, 2002.
- [73] H. M. Wain, E. A. Bruford, R. C. Lovering, M. J. Lush, M. W. Wright, and S. Povey, "Guidelines for human gene nomenclature," *Genomics*, vol. 79, pp. 464-70, 2002.
- [74] M. H. V. van Regenmortel, C. M. Fauquet, D. H. L. Bishop, E. G. Carstens, M. K. Estes, S. M. Lemon, J. Mniloff, M. A. Mayo, D. J. McGeoch, C. R. Pringle, and W. R.B., "Virus Taxonomy. Seventh Report of the International Committee on Taxonomy of Viruses." San Diego: Academic Press, 2000, pp. 1167.
- [75] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, "What are ontologies, and why do we need them?," *IEEE Intelligent Systems*, vol. 14, pp. 20-26, 1999.
- [76] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, 2000.

- [77] P. G. Baker, C. A. Goble, S. Bechhofer, N. W. Paton, R. Stevens, and A. Brass, "An ontology for bioinformatics applications," *Bioinformatics*, vol. 15, pp. 510-20, 1999.
- [78] R. Stevens, C. Goble, I. Horrocks, and S. Bechhofer, "Building a bioinformatics ontology using OIL," *IEEE Trans Inf Technol Biomed*, vol. 6, pp. 135-41, 2002.
- [79] R. Stevens, C. Goble, I. Horrocks, and S. Bechhofer, "OILing the way to machine understandable bioinformatics resources," *IEEE Trans Inf Technol Biomed*, vol. 6, pp. 129-34, 2002.
- [80] R. B. Altman, M. Buda, X. J. Chai, M. W. Carillo, R. O. Chen, and N. F. Abernethy, "RiboWeb: an ontology-based system for collaborative molecular biology," *Intelligent Systems, IEEE [see also IEEE Expert]*, vol. 14, pp. 68-76, 1999.
- [81] P. D. Karp, "An ontology for biological function based on molecular interactions," *Bioinformatics*, vol. 16, pp. 269-85, 2000.
- [82] Z. B. Miled, Y. W. Webster, Y. Liu, and N. Li, "An ontology for semantic integration of life science web databases," *International Journal of Cooperative Information Systems*, vol. 12, pp. 275-294, 2003.
- [83] C. Wroe, R. Stevens, C. Goble, A. Roberts, and M. Greenwood, "A suite of DAML+OIL ontologies to describe bioinformatics web services and data," *International Journal of Cooperative Information Systems*, vol. 12, pp. 197-224, 2003.
- [84] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," *Nat Genet*, vol. 29, pp. 365-71, 2001.
- [85] P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Jordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, C. J. Stoeckert, Jr., and A. Brazma, "Design and implementation of microarray gene expression markup language (MAGE-ML)," *Genome Biol*, vol. 3, pp. RESEARCH0046, 2002.
- [86] S. Orchard, H. Hermjakob, and R. Apweiler, "The proteomics standards initiative," *Proteomics*, vol. 3, pp. 1374-6, 2003.
- [87] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski,

- H. Husi, C. Brun, K. Shanker, S. G. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, and R. Apweiler, "The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data," *Nat Biotechnol*, vol. 22, pp. 177-83, 2004.
- [88] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, Goryanin, II, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang, "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, pp. 524-31, 2003.
- [89] P. D. Karp, "Pathway databases: a case study in computational symbolic theories," *Science*, vol. 293, pp. 2040-4, 2001.
- [90] L. Hon, N. F. Abernethy, V. Brusica, J. Chai, and R. B. Altman, "MHCWeb: converting a WWW database into a knowledge-based collaborative environment," *Proc AMIA Symp*, pp. 947-51, 1998.
- [91] Z. B. Miled, O. Bukhres, Y. Wang, N. Li, M. Baumgartner, and B. Sipes, "Biological and Chemical Information Integration System," presented at Network Tools and Applications in Biology, Genoa, Italy, 2001.
- [92] K. H. Cheung, P. M. Nadkarni, and D. G. Shin, "A metadata approach to query interoperation between molecular biology databases," *Bioinformatics*, vol. 14, pp. 486-97, 1998.
- [93] S. B. Davidson, A. S. Kosky, and B. Eckman, "Facilitating transformations in a human genome project database," presented at 3rd Int. Conf. Information and Knowledge Management, Gaithersburg, Maryland, USA, 1994.
- [94] S. Y. Chung and L. Wong, "Kleisli: a new tool for data integration in biology," *Trends Biotechnol*, vol. 17, pp. 351-5, 1999.
- [95] P. Buneman, S. B. Davidson, K. Hart, and G. C. Overton, "A data transformation system for biological data sources," presented at VLDB'95 21th Int. Conf. on Very Large Data Bases, Zurich, Switzerland, 1995.
- [96] I. M. Chen, A. S. Kosky, V. M. Markowitz, E. Szeto, and T. Topaloglou, "Advanced query mechanisms for biological databases," *Proc Int Conf Intell Syst Mol Biol*, vol. 6, pp. 43-51, 1998.

- [97] P. Mork, R. Shaker, A. Halevy, and P. Tarczy-Hornoch, "PQL: a declarative query language over dynamic biological schemata," *Proc AMIA Symp*, pp. 533-7, 2002.
- [98] L. Chen and H. M. Jamil, "On using remote user defined functions as wrappers for biological database interoperability," *International Journal of Cooperative Information Systems*, vol. 12, pp. 161-195, 2003.
- [99] P. Lambrix and V. Jakoniene, "Towards transparent access to multiple biological databanks," presented at First Asia-Pacific Bioinformatics Conference (APBC2003), Adelaide, Australia, 2003.
- [100] S. Philippi, "Light-weight integration of molecular biological databases," *Bioinformatics*, vol. 20, pp. 51-7, 2004.
- [101] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney, "EnSMart: a generic system for fast and flexible access to biological data," *Genome Res*, vol. 14, pp. 160-9, 2004.
- [102] D. Frishman, K. Albermann, J. Hani, K. Heumann, A. Metanomski, A. Zollner, and H. W. Mewes, "Functional and structural genomics using PEDANT," *Bioinformatics*, vol. 17, pp. 44-57, 2001.
- [103] Y. V. Bukhman and J. Skolnick, "BioMolQuest: integrated database-based retrieval of protein structural and functional information," *Bioinformatics*, vol. 17, pp. 468-78, 2001.
- [104] L. M. Hass, P. M. Schwarz, P. Kodali, E. Kotlar, J. E. Rice, and W. C. Swope, "DiscoveryLink: A system for integrated access to life sciences data sources," *IBM Systems Journal*, vol. 40, pp. 489-511, 2001.
- [105] G. J. L. Kemp, N. Angelopoulos, and P. M. D. Gray, "Architecture of a mediator for a bioinformatics database federation," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 6, pp. 116-122, 2002.
- [106] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. W. Paton, C. A. Goble, and A. Brass, "TAMBIS: transparent access to multiple bioinformatics information sources," *Bioinformatics*, vol. 16, pp. 184-5, 2000.
- [107] B. Ludascher, A. Gupta, and M. E. Martone, "Model-based mediation with domain maps," presented at Data Engineering, 2001. Proceedings. 17th International Conference on, 2001.
- [108] J. Kohler, S. Philippi, and M. Lange, "SEMEDA: ontology based semantic integration of biological databases," *Bioinformatics*, vol. 19, pp. 2420-7, 2003.
- [109] D. C. Jamison, B. Mills, and B. Schatz, "An extensible network query unification system for biological databases," *Comput Appl Biosci*, vol. 12, pp. 145-50, 1996.

- [110] T. Etzold and P. Argos, "SRS--an indexing and retrieval tool for flat file data libraries," *Comput Appl Biosci*, vol. 9, pp. 49-57, 1993.
- [111] J. Ostell, "The Entrez Search and Retrieval System," in *The NCBI Handbook*, 2002.
- [112] P. J. Kersey, L. Morris, H. Hermjakob, and R. Apweiler, "Integr8: enhanced inter-operability of European molecular biology databases," *Methods Inf Med*, vol. 42, pp. 154-60, 2003.
- [113] W. Fujibuchi, S. Goto, H. Migimatsu, I. Uchiyama, A. Ogiwara, Y. Akiyama, and M. Kanehisa, "DBGET/LinkDB: an integrated database retrieval system," *Pac Symp Biocomput*, pp. 683-94, 1998.
- [114] M. Chagoyen, M. E. Kurul, P. A. De-Alarcon, J. M. Carazo, and A. Gupta, "Designing and executing scientific workflows with a programmable integrator," *Bioinformatics*, vol. 20, pp. 2092-2100, 2004.
- [115] S. A. Racunas, N. H. Shah, I. Albert, and N. V. Fedoroff, "HyBrow: a prototype system for computer-aided hypothesis evaluation," *Bioinformatics*, vol. 20 Suppl 1, pp. I257-I264, 2004.
- [116] P. W. Lord, J. R. Reich, A. Mitchell, R. D. Stevens, and C. A. Goble, "PRECIS: an automated pipeline for producing concise reports about proteins," presented at Bioinformatics and Bioengineering Conference, 2001. Proceedings of the IEEE 2nd International Symposium on, 2001.
- [117] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li, "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics*, vol. 20, pp. 3045-54, 2004.
- [118] D. M. Shotton, "Electronic light microscopy: present capabilities and future prospects," *Histochem Cell Biol*, vol. 104, pp. 91-137, 1995.
- [119] D. J. Stephens and V. J. Allan, "Light microscopy techniques for live cell imaging," *Science*, vol. 300, pp. 82-6, 2003.
- [120] H. Chen, J. R. Swedlow, M. Grote, J. W. Sedat, and D. A. Agard, "The collection, processing, and display of digital three-dimensional images of biological specimens," in *Handbook of biological confocal microscopy*, J. Pawley, Ed.: Plenum Press, 1995, pp. 197-210.
- [121] A. J. Koster, R. Grimm, D. Typke, R. Hegerl, A. Stoschek, J. Walz, and W. Baumeister, "Perspectives of molecular and cellular electron tomography," *J Struct Biol*, vol. 120, pp. 276-308, 1997.
- [122] A. Engel, C. A. Schoenenberger, and D. J. Mueller, "High resolution imaging of native biological sample surfaces using scanning probe microscopy," *Curr Opin Struct Biol*, vol. 7, pp. 279-284, 1997.



- [123] J. K. Horber and M. J. Miles, "Scanning probe evolution in biology," *Science*, vol. 302, pp. 1002-5, 2003.
- [124] R. Marabini, C. Vaquerizo, J. J. Fernandez, J. M. Carazo, A. Engel, and J. Frank, "Proposal for a new distributed database of macromolecular and subcellular structures from different areas of microscopy," *J Struct Biol*, vol. 116, pp. 161-5, 1996.
- [125] N. Salmon, S. Lindek, and E. H. K. Stelzer, "Databases for microscopies and microscopical images," in *Handbook of computer vision and applications*, vol. 2, H. H. B. Jahne, P. Geissler, Ed.: Academic Press, 1999, pp. 907-926.
- [126] R. S. Pressman, *Software engineering: a practitioner's approach*: McGraw-Hill, 1997.
- [127] J. J. Pittet, C. Henn, A. Engel, and J. B. Heymann, "Visualizing 3D data obtained from microscopy on the Internet," *J Struct Biol*, vol. 125, pp. 123-32, 1999.
- [128] P. P.-S. Chen, "The entity-relationship model - toward a unified view of data," *ACM Transactions on Database Systems (TODS)*, vol. 1, pp. 9-36, 1976.
- [129] *Guide to FORML*: Infomodelers, Inc., 1997.
- [130] P. A. de Alarcon, A. Gupta, and J. M. Carazo, "A framework for querying a database for structural information on 3D images of macromolecules: A web-based query-by-content prototype on the BioImage macromolecular server," *J Struct Biol*, vol. 125, pp. 112-22, 1999.
- [131] T. Boudier and D. M. Shotton, "Video on the Internet: An introduction to the digital encoding, compression, and transmission of moving image data," *J Struct Biol*, vol. 125, pp. 133-55, 1999.
- [132] S. G. Kalko, M. Chagoyen, N. Jimenez-Lozano, N. Verdaguer, I. Fita, and J. M. Carazo, "The need for a shared database infrastructure: combining X-ray crystallography and electron microscopy," *Eur Biophys J*, vol. 29, pp. 457-62, 2000.
- [133] E. Gonzalez-Couto, B. Hayes, and A. Danckaert, "The life sciences Global Image Database (GID)," *Nucleic Acids Res*, vol. 29, pp. 336-9, 2001.
- [134] J. R. Swedlow, I. Goldberg, E. Brauner, and P. K. Sorger, "Informatics and quantitative analysis in biological imaging," *Science*, vol. 300, pp. 100-2, 2003.
- [135] F. Metoz, M. B. Sherman, and M. F. Schmid, "Adopting a database as a solution to managing electron image data," *J Struct Biol*, vol. 133, pp. 170-5, 2001.
- [136] Y. Liang, E. Y. Ke, and Z. H. Zhou, "IMIRS: a high-resolution 3D reconstruction package integrated with a relational image database," *J Struct Biol*, vol. 137, pp. 292-304, 2002.

- [137] W. Dai, Y. Liang, and Z. H. Zhou, "Web portal to an image database for high-resolution three-dimensional reconstruction," *J Struct Biol*, vol. 144, pp. 238-45, 2003.
- [138] M. E. Martone, A. Gupta, M. Wong, X. Qian, G. Sosinsky, B. Ludascher, and M. H. Ellisman, "A cell-centered database for electron tomographic data," *J Struct Biol*, vol. 138, pp. 145-55, 2002.
- [139] S. J. Ludtke, L. Nason, H. Tu, L. Peng, and W. Chiu, "Object oriented database and electronic notebook for transmission electron microscopy," *Microsc Microanal*, vol. 9, pp. 556-65, 2003.
- [140] W. Baumeister and A. C. Steven, "Macromolecular electron microscopy in the era of structural genomics," *Trends Biochem Sci*, vol. 25, pp. 624-31, 2000.
- [141] M. Auer, "Three-dimensional electron cryo-microscopy as a powerful structural tool in molecular medicine," *J Mol Med*, vol. 78, pp. 191-202, 2000.
- [142] T. S. Baker, N. H. Olson, and S. D. Fuller, "Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs," *Microbiol Mol Biol Rev*, vol. 63, pp. 862-922, table of contents, 1999.
- [143] B. F. McEwen and M. Marko, "The emergence of electron tomography as an important tool for investigating cellular ultrastructure," *J Histochem Cytochem*, vol. 49, pp. 553-64, 2001.
- [144] N. Volkmann, K. J. Amann, S. Stoilova-McPhie, C. Egile, D. C. Winter, L. Hazelwood, J. E. Heuser, R. Li, T. D. Pollard, and D. Hanein, "Structure of Arp2/3 complex in its activated state and in actin filament branch junctions," *Science*, vol. 293, pp. 2456-9, 2001.
- [145] E. Nogales and N. Grigorieff, "Molecular Machines: putting the pieces together," *J Cell Biol*, vol. 152, pp. F1-10, 2001.
- [146] H. R. Saibil, "Conformational changes studied by cryo-electron microscopy," *Nat Struct Biol*, vol. 7, pp. 711-4, 2000.
- [147] N. Volkmann and D. Hanein, "Quantitative fitting of atomic models into observed densities derived by electron microscopy," *J Struct Biol*, vol. 125, pp. 176-84, 1999.
- [148] W. Wriggers and S. Birmanns, "Using situs for flexible and rigid-body fitting of multiresolution single-molecule data," *J Struct Biol*, vol. 133, pp. 193-202, 2001.
- [149] M. G. Rossmann, "Fitting atomic models into electron-microscopy maps," *Acta Crystallogr D Biol Crystallogr*, vol. 56 (Pt 10), pp. 1341-9, 2000.
- [150] J. Bohm, A. S. Frangakis, R. Hegerl, S. Nickell, D. Typke, and W. Baumeister, "Toward detecting and identifying macromolecules in a cellular context:

- template matching applied to electron tomograms," *Proc Natl Acad Sci U S A*, vol. 97, pp. 14245-50, 2000.
- [151] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat Struct Biol*, vol. 10, pp. 980, 2003.
- [152] H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Golovin, K. Henrick, A. Hussain, J. Ionides, M. John, P. A. Keller, E. Krissinel, P. McNeil, A. Naim, R. Newman, T. Oldfield, J. Pineda, A. Rachedi, J. Copeland, A. Sitnov, S. Sobhany, A. Suarez-Uruena, J. Swaminathan, M. Tagari, J. Tate, S. Tromm, S. Velankar, and W. Vranken, "E-MSD: the European Bioinformatics Institute Macromolecular Structure Database," *Nucleic Acids Res*, vol. 31, pp. 458-62, 2003.
- [153] A. Golovin, T. J. Oldfield, J. G. Tate, S. Velankar, G. J. Barton, H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Hussain, J. M. Ionides, M. John, P. A. Keller, E. Krissinel, P. McNeil, A. Naim, R. Newman, A. Pajon, J. Pineda, A. Rachedi, J. Copeland, A. Sitnov, S. Sobhany, A. Suarez-Uruena, G. J. Swaminathan, M. Tagari, S. Tromm, W. Vranken, and K. Henrick, "E-MSD: an integrated data resource for bioinformatics," *Nucleic Acids Res*, vol. 32 Database issue, pp. D211-6, 2004.
- [154] N. Collaborative Computational Project, "The CCP4 suite: programs for protein crystallography," *Acta Crystallogr D Biol Crystallogr*, vol. 50, pp. 760-3, 1994.
- [155] M. Tagari, R. Newman, M. Chagoyen, J. M. Carazo, and K. Henrick, "New electron microscopy database and deposition system," *Trends Biochem Sci*, vol. 27, pp. 589, 2002.
- [156] P. A. De-Alarcon, A. Pascual-Montano, A. Gupta, and J. M. Carazo, "Modeling shape and topology of low-resolution density maps of biological macromolecules," *Biophys J*, vol. 83, pp. 619-32, 2002.
- [157] P. A. de Alarcon, "Representación, análisis geométrico y recuperación por contenido de imágenes tridimensionales. Aplicaciones bioinformáticas en microscopía electrónica," in *Escuela Politécnica Superior*. Madrid: Universidad Autónoma de Madrid, 2002, pp. 168.
- [158] J. B. Heymann, M. Chagoyen, and D. M. Belnap, "Common conventions for the interchange and archiving of three-dimensional electron microscopy information in structural biology," *J Struct Biol*, (submitted).
- [159] P. E. Bourne, H. M. Berman, B. McMahon, K. D. Watenpaugh, J. D. Westbrook, and P. M. D. Fitzgerald, "Macromolecular crystallographic information file," *Method Enzymol*, vol. 277, pp. 571-590, 1997.
- [160] S. R. Hall, "The Star File - a New Format for Electronic Data Transfer and Archiving," *Journal of Chemical Information and Computer Sciences*, vol. 31, pp. 326-333, 1991.

- [161] S. D. Fuller, "Depositing electron microscopy maps," *Structure (Camb)*, vol. 11, pp. 11-2, 2003.
- [162] D. Hollingsworth, "The Workflow Reference Model," Workflow Management Coalition TC00-1003, 19 Jan. 1995 1995.
- [163] A. Ailimaki, Y. E. Ioannidis, and M. Livny, "Scientific workflow management by database management," presented at 10th Int Conf on Scientific and Statistical Database Management, Capri, Italy, 1998.
- [164] B. Ludascher, I. Altintas, and A. Gupta, "Compiling abstract scientific workflows into Web service workflows," presented at 15th International Conference on Scientific and Statistical Database Management, 2003., Cambridge, MA, USA, 2003.
- [165] A. Gupta, B. Ludascher, M. E. Martone, A. Rajasekar, E. Ross, X. Qian, S. Santini, H. He, and I. Zaslavsky, "BIRN-M: a semantic mediator for solving real-world neuroscience problems," presented at 2003 ACM SIGMOD Int Conf Management of Data, San Diego, CA, USA, 2003.
- [166] V. Crescenzi and G. Mecca, "Grammars have exceptions," *Information Systems*, vol. 23, pp. 539-565, 1998.
- [167] R. Stevens, C. Goble, P. Baker, and A. Brass, "A classification of tasks in bioinformatics," *Bioinformatics*, vol. 17, pp. 180-8, 2001.
- [168] A. Gupta, B. Ludascher, and M. E. Martone, "Registering scientific information sources for semantic mediation," presented at 21st Int Conf Conceptual Modeling (ER), Tampere, Finland, 2002.
- [169] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Res*, vol. 32 Database issue, pp. D115-9, 2004.

## Biographical note

---

### Academic and research record

Telecommunication engineer (E.T.S. Ingenieros de Telecomunicación, Universidad Politécnica de Madrid), 1996.

Master Thesis (Instituto de Óptica Daza de Valdés, Consejo Superior de Investigaciones Científicas).

Since October 1995, Biocomputing Unit of the National Center for Biotechnology (Consejo Superior de Investigaciones Científicas), Madrid, Spain.

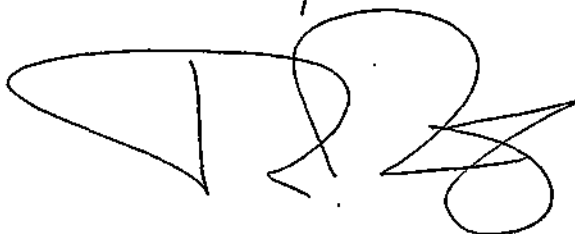
### Relevant publications

- Carazo JM, Stelzer E, Engel A, fita I, Henn C, Machtynger J, McNeil P, Shotton DM, Chagoyen M, de Alarcón PA, Lindek S, Fritsh R, Heymann B, Kalko S, Pittet JJ, Rodriguez-Tome P and Boudier T, "Organising multidimensional biological image information: The BioImage database", *Nucleic Acids Research* (1999) 27: 280-283



- Lindek S, Fritsch R, Machtynger J, de Alarcón PA and **Chagoyen M**, "Design and realization of an on-line database for multidimensional microscopic images of biological specimens", *Journal of Structural Biology* (1999) 125:103-111
- Kalko S, **Chagoyen M**, Jiménez-Lozano N, Verdaguer N, Fita I and Carazo JM, "The need for a shared database infrastructure: combining X-ray crystallography and electron microscopy", *European Biophysics Journal* (2000) 29: 457-462
- Tagari M, Newman R, **Chagoyen M**, Carazo JM and Henrick K, "New electron microscopy database and deposition system", *TRENDS in Biochemical Sciences* (2002) 27: 589
- Jiménez-Lozano N, **Chagoyen M**, Cuenca-Alba J and Carazo JM, "FEMME (Feature Extraction in a Multiresolution Macromolecular Environment) database: new insight for 3D-EM analysis", *Journal of Structural Biology* (2003) 144: 104-113
- Henrick K, Newman R, Tagari M, and **Chagoyen M**, "EMDep: a web-based system for the deposition and validation of high-resolution Electron Microscopy macromolecular structural information", *Journal of Structural Biology* (2003) 144: 228-237
- Jiménez-Lozano N, **Chagoyen M**, de-Alarcón PA and Carazo JM, "Extracting and searching for structural information: a multiresolution approach", *Methods in Protein and Proteome Analysis* (2004) 341-358 Springer-Verlag
- **Chagoyen M**, Kurul ME, de-Alarcón PA, Carazo JM and Gupta A, "Designing and executing scientific workflows with a programmable integrator", *Bioinformatics* (2004) 20: 2092-2100
- Heymann JB, **Chagoyen M** and Belnap DM, "Common conventions for interchange and archiving of three-dimensional electron microscopy information in structural biology", (enviado)

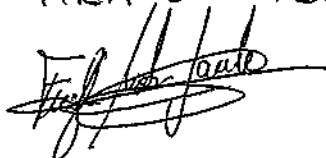
Reunido el tribunal que suscribe en el día  
de la fecha, acordó calificar la presente Tesis  
doctoral con APTO CUM LAUDE  
Madrid, 19 de Mayo de 2005



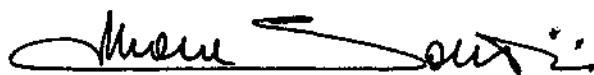
FDO: ROBERTO MORIYÓN SALOMÓN



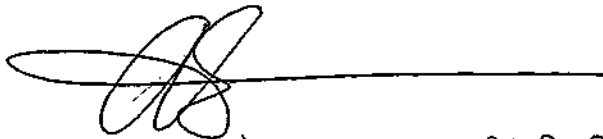
FDO: FRANCISCO TIRADO FERNÁNDEZ



FDO: FERNANDO MARTÍN SÁNCHEZ



FDO: SIMONE SAUTINI



FDO: CARLOS ORTIZ DE SOLORZANO AURUSA

