Digital Human Ontology

European Commission -
US National Science Foundation

Strategic Research Workshop

Digital Human Ontology

Bethesda, USA, 25-26 July 2002

Workshop Report and
Recommendations

Digital Human Ontology

A EC/NSF Workshop of Scientists

under the auspices of the Future and Emerging Technologies Activity of the
Information Society Technologies Programme of DG XII of the European Commission
and the National Science Foundation of the United States of America

US Scientific Coordinator: Dr Henry Kelly

Head of European Delegation: Dr Martin Hofmann

Coordinators: ERCIM and the Federation of American Scientists

National Institutes of Health
Bethesda, Marryland, 20894 USA
July 2002

Disclaimer
The content of this report is the sole responsibility of the EU-NSF Workshops Project Consortium and does not necessarily represent the views of the European Commission, the US National Science Foundation or of their respective services.

Executive Summary

The Digital Human is a project to create a software simulation of the most complex system known: the human body. Using 21st century information technology and tools, scientific communities disseminated across the world (mainly Europe, United states and Japan) are striving to present the human body's full processes, from DNA molecules and proteins to cells, tissues, organs, and gross anatomy**.**

Several research teams have already successfully implemented software simulation for organs, limbs, cells, neuromuscular and cardivascular biomechanics, etc…

Getting to study and simulate human body components or movements usually requires experimental and computational approaches such as biomechanical models developed to analyze muscle function, study movement, design new medical products, and to ultimately guide surgery.

The Digital Human project is now at a crossroad where a large number of simulations of organs and body processes have been validated independently. The objective of the Digital Human project is to unite all those simulation into one full scale digital human.

The road towards completion is however paved with a major obstacle: the lack of interoperability among all the models. The underlying problem is the absence of unified ontology, hence making hard for experts to address common issues or to couple models together.

In this perspective, the EC/NSF meeting was held in the National Institutes of Health Bethesda, Marryland, on the 25 and 26th of July 2002. This workshop gathering both European and American experts had the ambition to be a kick-off the harmonisation and unification of ontology in the field. The workshop Chairman, Dr Henry Kelly from the Federation of American Scientists, was assisted in his task by Dr Martin Hofmann (FhG) head of the European delegation. The list of all participants is available in Annex 2.

Table of Contents

Workshop Structure and Objectives

The goal of this workshop is to develop an approach for unifying ontologies at multiple scales of biological structure and function. First iteration should span cell and above, including tissue, organ and system levels.

In parallel with discussions about the development of a broader ontology framework for biology, discussions will also be placed in context of specific applications including the development of a framework for medical simulation-based training.

In this respect, the meeting focused on producing:

- Outline of a planning document (‘straw-man’), including identification of challenges and obstacles to development of a unified ontology (integration, correlation, evaluation).

- Process for extending current discussion to engage other developers, including molecular level researchers (*e.g.*, Gene Ontology).

- Planning of subsequent, follow-up workshops including meeting objectives and dates.

- Specific recommendations for use of existing ontology (organ, tissue and cell levels) for medical simulation.

To carry out its objectives, the workshop was subdivided in **ontology framework** and **medical simulation** break out groups to address specific question:

Ontology framework – Discussion of an approach for unification of ontologies for broader Digital Human, including planning document and subsequent workshops and meetings.

<u>A. What are the features of candidate ontologies that will facilitate integration and correlation into a larger unified framework?</u>

i) Semantic structure of core ontologies
ii) Concept representation environments:
   (1) Database
   (2) Formalisms:
      - Frame-based
      - Description logic
      - First-order logic
iii) Reference to other terminologies
iv) Query processing
v) Inference engine

B. What are the methods for correlating ontologies?

- i) Comparison of concept domain
  - (1) Correlation of concepts and terms
- ii) Identification of gaps, overlaps and redundancies
- iii) Mapping ontologies to one another
- iv) Automation of the above processes

C. What are the methods for evaluating ontologies?

- i) Expressivity
  - (1) Ability to represent different layers of knowledge
  - (2) What can you say about the domain?
- ii) Scalability
  - (1) Enhance granularity
  - (2) Extend to other domains and subdomains
    - (a) (*e.g.*, embryological development, other species)
- *iii)* Generalizable for multiple applications
- *iv)* Support diverse models and simulations

D. Where do we go from here?

- i. Planning document
- ii. Process
- iii. Subsequent workshops

Applications (Medical Simulation) – Discussion of applications such as medical simulation training that can derive immediate benefit from the use of a unified ontology. This break out group was to address the extent to which the same ontology can be used for software designed for research, training, and medical practice. It was also expected to bring and insight on the difficulty to convert existing code to reflect a common hierarchy (wrappers etc…)

A. Definition of the problem/issues

- i) Is it useful for us to develop a common approach for medical simulation?
- ii) If so, how de we begin the process of development, collaboration, interaction?
- iii) Define requirements: Surgical training versus other medical training applications
- iv) How to include other training issues such as performance assessment, etc…

B. Can we develop a standard ontology for biomedical simulation?

i) Evaluation of existing approaches - What existing ontologies have proven useful to encourage interoperability?

ii) Integration of medical/disease orientation with biological research relevance

C. How should we proceed?

i) What practical steps should be taken to encourage development and use of a common ontology and build interoperable objects?

ii) Immediate goals – How can we work together? What can we use today?

iii) Next steps

Workshop Discussions

The meeting aimed at setting up an international consortium of scientists who want to develop basic concepts and who want to implement a system that ultimately leads to the “digital human”, a computer representation of the human body that allows for data management, information management and simulation of biological processes in the context of human anatomy. The discussions focused mainly on the following aspects:

- Using the foundational model for anatomy as suggested by Rosse et al. as the start point for a unified ontology for human anatomy
- Using the digital human for knowledge management and data integration
- Need for a unified concept based on the agreement of a large and strong group of highly respected international scientists; the need for unified concepts is seen mostly from the side of funding agencies that do not want to continue wasting money on too many parallel activities that do not make it to a *de facto* standard
- Funding politics and need to move forward even if funding is not available at present

Reasons for the success of consortia like the Gene Ontology consortium or the MGED consortium

An overview of the different **on-going projects** related to Digital Human was also made. In particular it addressed the initiatives listed in **Annex 2**.

Workshop Recommendations

1) *Using anatomy for knowledge management.*

Over the last 30 years, significant amounts of information on biomolecules including genes and their products have been accumulated in the eras of molecular biology and genomics. This information, however, is scattered and distributed over several databases. Most biomedical information is not well integrated in the sense that unique links to genes and unique links to higher order structures, especially to tissues and organs, are recorded and maintained. This is partly due to the fact, that when experiments (e.g. gene expression studies) are being described scientists could not refer to a standardised system for the description of anatomy. Such "ontology" [the term ontology is used in this document as a "system to describe the nature and the relationship of objects in a structured way using controlled vocabularies] for anatomy would be one way (amongst others) to structure information and to allow researchers from different fields to refer to a common standard when describing new findings. It should be noted at this point that the lack of such common standard resulted in a massive loss of information in the past; estimations on the financial side are difficult to assess, however, it is probably not unjust to assume that more than 50% of all relevant information that lead to a publication is lost; which in turn means that more than 50% of the money spent for biomedical research is lost.

As a "natural system" to organise biomedical data and knowledge, anatomy provides us with an order system that due to its physical, direct and "natural" appearance makes it inherently "user friendly". This central role of anatomy as the top level structural principle in biology and medicine is also highlighted by the fact that education of medical personal and medical doctors starts with basic courses in this discipline.

2) *Role of ontologies.*

Ontologies are basically systems that reflect the nature and the relationship of objects in a structured way. [for a very condensed as well as an extensive definition of an ontology please see: http://www.cbil.upenn.edu/Ontology/#ontology.whatis] Ontological descriptions follow simple rules and thus can be checked for consistency. Inference from ontological objects is possible because the definition of relationships in an ontology follows rules that allow defining constrains. Object representations in ontologies are frequently used as basic concepts for object models used in database design. The building of an ontology is therefore a first step towards a real understanding of a process or a system of objects and their relationships. It is also the first step towards the handling of a process or a system in the computer. In the life sciences, building an ontology means to generate awareness for the need to agree on controlled vocabularies and is usually intended to provide a standard for the description of observations and analyses.

3) ***Importance of standardization***

With the dawn of genomics - that is the description and analysis of biological phenomena at a genome – wide scale – the data produced in an experiment are becoming relevant for a much wider community of researchers including researchers that have no direct link to the field of research of the laboratory where the physical experiment (e.g. gene expression studies) has been done initially. The use of data from genomics experiments in a different "domain" (e.g. the use of gene expression data generated by a developmental biology group in a physiological model of heart function by a theoretical biology group) requires a sort of "interoperability" between knowledge domains and easy transformation of decomposed (→ fragmented) models or data sets.

Currently, most of the biomedical data cannot be subjected to different types of analyses or analyses outside of the domain where the data have been generated. Systems that allow to fit new data into a framework with well defined fix points (such as anatomy as a very fundamental principle of biology) will help to record and to publish data in a way that enables free exchange of decomposed data, models and information. As a prerequisite, major parts of the research communities have to agree on a common understanding of the basic principles. The digital human project aims at creating a forum to establish this common understanding for a large part of human biology.

4) ***Scope of the international consortium***

The "digital human consortium" will establish itself as an international group that will "produce" a general framework for communities to store, manage, model and simulate biomedical data. The framework will be based on a continuously developed and improved ontology for human anatomy; thus providing a "common anchor" for the allocation of information in the anatomical context. The "digital human consortium" will be set up in a similar way as many other successful consortia in related fields before: it is based on the free participation of international researchers and it will have a strong core team of people driving the program of the consortium. The core mission of the Digital Human Consortium is helping different research communities to fit their knowledge systems (e.g. their own ontologies and controlled vocabularies) into the large framework of the digital human; to provide the computer scientist and developer communities with standards (e.g. XML) and tools via the opensource project and to harvest feedback from users (biologists and medical doctors doing experimental and / or theoretical research). Most important, it will have an "educational" function for both, theoretical simulation scientists and experimental scientists as it will provide a blueprint to record data and to publish results. Moreover, by making it accessible to everybody, the digital human will be THE knowledge environment for human biology (including human pathology, human anatomy, human physiology, human genetics and many more disciplines).

The dimension of the attempt to generate an electronic representation of the human which ultimately allows for knowledge retrieval and knowledge management of all biomedical sciences can only be compared to projects like the first flight of man to the moon, the attempts to prove the existence of quarks or the human genome project. With respect to its direct applicability and its scientific and social impact it reaches much further than the huge projects listed above.

5) ***Participants from the European side***

On the European side the following laboratories and groups are likely to be interested in participating in the digital human consortium:

1. Steffen Schulze-Kremer, RZPD, Berlin, Germany
2. Paul van der Vet, University of Twente, Twente, The Netherlands
3. Ralf Zimmer, University of Munich (LMU), Munich, Germany
4. Eric Minch, LION bioscience AG, Heidelberg, Germany
5. Martijn Huynen, Nijmegen University, Nijmegen, The Netherlands
6. Thure Etzold, European Bioinformatics Institute, Hinxton, England
7. others …..(list to be extended)

6) ***Funding opportunities***

The workshop held at NIH on July 25 / 26 came to the conclusion that it would take too long to wait for funding by the "normal" funding agencies given that this would result in a delay of several month (up to one and a half year) for the start of the project. On a mid-term perspective, however, funding will be necessary. A joint ESF or FP6 (EU) and NSF (USA) funding should be evaluated. Besides efforts going into the direction of joint funding it should also be possible to establish the consortium on independent funding on both sides of the Atlantic. The experience with MGED shows that this works.

Even in the case that there will be two independent lines of funding (one in the US, one in Europe) there should be a bank account for the consortium itself. MGED turned into a "society" or even something like a company ("Ltd." ?) to establish itself as an independent institution that is allowed to take money from sponsors. Both, the gene ontology consortium and the MGED consortium had funding from industry (Astra Zeneca and Incyte in the case of GO; money from the Industry EBI Program in case of MGED).

Actions and Conclusions

- A manuscript / short communication will be written (most likely by Kelly, Higgins and Rosse) describing the formation of the digital human consortium. It was suggested to submit the short communication to a high ranking journal (NATURE or SCIENCE). In addition, a monography is in the making that should summarise the different aspects discussed at the meeting (e.g. need for unified ontologies; anatomy as the basic principle to manage data and information; interoperability of data and methods (especially simulation); future aspects of a public knowledge environment covering human biology)

- The foundational model of Dr Rosse will be used as the basic ontology for human anatomy. Currently, it comprises of about 18,000 terms (? → no website with this ontology so far available) in a controlled vocabulary. Cornelius Rosse is the scientific director and initiator of the "digital anatomist" project at Washington University (see also http://sig.biostr.washington.edu/projects/da/).

- In the discussions it became apparent that there is hardly anything like a comprehensive overview on existing ontologies and projects aiming at "unified ontologies". Even though the majority of participants of the workshop were really experts in the field of computer sciences and bio-medical informatics, none would have been able to write a comprehensive overview on existing concepts (GALEN would be an example for an existing concept) for ontologies in biology and medicine. Thus, getting this overview together is one of the premier tasks for the new consortium and as we have to do a lot of organisational, overview-type work in Europe anyway it would be a good idea if we would start with collecting bio-ontologies and getting everything together on visualisation of organs (e.g. voxels).

- Concerning funding of early activities we all agreed that it takes too much time to wait for funding now. Following the example of GO and MGED we have to get started immediately and to do some initial work based on our "inhouse resources". Later on we will apply for large grants as the "digital human" project will be THE major task for biomedical informatics in the near and mid-term future.

- One of the major scientific problems is the question, whether the "foundational model" should be expanded (from the definition ions (e.g. Ca ++) and their relationships via genes and proteins and protein complexes to higher order structures such as cells and tissues) or whether it would not be better to define the interfaces to or overlaps with existing ontologies such as gene ontology or e.g. tumour classifications. This would leave the task of updating and perfecting of different ontologies within the expert domains. As a side effect, the digital human consortium would not have to handle the complete width of human biology.

- One strong aspect in the discussions was the demand for interoperability of data and annotations in a sense that data as well as simulation algorithms and packages should be made exchangeable. This is a demanding task as currently there is no such interoperability at all. However, using the digital human ontologies as a means to organise data retrieval and data management (e.g. experimental recordings) is a first step towards interoperability of higher order analysis and simulation applications.

Annex 1 - Workshop Presentations

M.C. Çavusoglu, T. Göktekin, Ph.D.
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Open Source / Open Architecture Software Development Framework for Surgical Simulation

Henry Kelly, Ph.D.
President of FAS , PI of Digital Human Project
Digital Human - Unified Ontology

Mark A. Musen, M.D., Ph.D.
Medical Informatics Stanford University
Building Ontologies with Protégé-2000

Alan Rector, Ph.D.
Medical Informatics Group
University of Manchester
Linking levels of granularity and expressing contexts & views using formal ontologies: Experience with the Digital Anatomist FMA & other health & bio ontologies

Brian Athey, Ph.D.
University of Michigan
Linking levels of granularity and expressing contexts & views using formal ontologies: Experience with the Digital Anatomist FMA & other health & bio ontologies

Cornelius Rosse M.D., D.Sc.
Structural Informatics Group
University of Washington
The Potential of the Digital Anatomist Foundational Model for “Unifying” Biomedical Ontologies

Bruce Porter, Art Souther
Department of Computer Science
University of Texas at Austin

Vinay Chaudhri
AI Center
Stanford Research Institute

Peter Clark
Department of Computer Science Math and Computing Research Center, Boeing
Technologies to Enable Biologists to Build Large Knowledge Bases on Human Anatomy and Physiology

Presentations are available at: http://fas.org/dh/conferences/paper.php (July 2003)

Annex 2 - Workshop Participants

- Licinio Angelini, M.D. *University of Roma, Italy*
- Brian Athey, Ph.D., *MCBI, University of Michigan*
- Peter Bird, Ph.D., *MCBI, University of Michigan*
- Christyne Bilton, *MCBI, University of Michigan*
- Fred Bookstein, Ph.D., *University of Michigan*
- Giovanni Bortolan, Ph.D., *Institute of Biomedical Engineering, LADSEB-CNR, Italy*
- Jeff Canceko, *Georgetown University*
- M. Cenk Cavusoglu, Ph.D., *University of Californa, Berkeley; Case Western Reserve University*
- Pavarti Dev, Ph.D., *Director SUMMIT, Stanford University*
- John Gennari, Ph.D., *University of Washington*
- Gerald Higgins, Ph.D., *Director of Digital Human Project, Federation of American Scientists*
- Martin, Hofmann, Ph.D., *SCAI, Germany*
- Don Jenkins, Ph.D., *National Library Medicine*
- Henery Kelly, Ph.D., *President, Federation of American Scientists*
- Steven Koslow, Ph.D., *Director, NIMH, National Institutes of Health*
- Micheal Liebman, Ph.D., *Director, Computational Biology and Biomedical Informatics, Univ.of Pennsylvania*
- Patrick Lincoln, Ph.D., *Director, Computer Science Laboratory, SRI International*
- Andrea Mattasoglio, Ph.D., *CILEA, Italy*
- Jose Mejino, M.D., *University of Washington*
- Ion Moraru, Ph.D., *NRCAM, University of Connecticut*
- Gerry Moses, Ph.D., *TATRC*

Mark Musen, M.D., Ph.D., *Stanford Medical Informatics*

- Heinz-Otto Peitgen, Ph.D., *Director of MeVis, University of Bremen, Germany*
- Francesco Pinciroli, Ph.D., Chair of Medical Informatics, Dipartimento di Bioingegneria, Politecnico di Milano, Italy
- Bruce Porter, Ph.D., *University of Texas, Austin*

Alan Rector, M.D., Ph.D., *University of Manchester, UK*

- Remi Ronchaud, Ph.D., *Project Manager, ERCIM, France*
- Cornelius Rosse, M.D., *University of Washington*
- Shankar Sastry, Ph.D., *University of California, Berkeley*
- Richard Satava, Ph.D., *Yale University, University of Washington*
- Rainer Schubert, Ph.D., *University for Health Informatics and Technology, Austria*
- Art Souther, Ph.D., *University of Texas, Austin*
- Manolis Tsiknakis, Ph.D., *CMI/HTA, Institute of Computer Science, Greece*
- Richard Ward, Ph.D., *Oak Ridge National Lab*
- Ron White, Ph.D., *NASA*

Annex 3 - On-going Projects

Berkeley Drosophila Geneome Project (BDGP)

Project

The Berkeley Drosophila Genome Project (BDGP) is a consortium of the Drosophila Genome Center, funded by the National Human Genome Research Institute, National Cancer Institute, and Howard Hughes Medical Institute.

The goals of the Drosophila Genome Center are to finish the sequence of the euchromatic genome of Drosophila melanogaster and to generate and maintain biological annotations of this sequence. In addition to genomic sequencing, the BDGP is 1) producing gene disruptions using P element-mediated mutagenesis on a scale unprecedented in metazoans; 2) characterizing the sequence and expression of cDNAs; and 3) developing informatics tools that support the experimental process, identify features of DNA sequence, and allow us to present up-to-date information about the annotated sequence to the research community. The BDGP Informatics Group is a member of the FlyBase consortium.

Ontology

The Gene Ontology is a controlled vocabulary "for the description of the molecular function, biological process and cellular component of gene products." The Gene Ontology is always being revised by the ever-vigilant Gene Ontology consortium, a group composed of academic and commercial genetic researchers.

Contact: BDGP Informatics Group

- Suzanna Lewis	suzi@fly2.berkeley.edu
- Brad Marshall	bradmars@yahoo.com
- Chris Mungall	cjm@fruitfly.bdgp.berkeley.edu
- John Richter	jmrichter@lbl.gov

CBIL – Computational Biology and Informatics Laboratory at the U. of Pennsylvania

Project

Houses the following databases: EPConDB, PlasmodiumDB, StemCellDB, RAD, EpoDB, MTIR, and ParaDBs. As well as a "Controlled Vocabulary," which contains a hierarchical controlled vocabulary of anatomy terms used in CBIL's databases. The controlled vocabulary is based on a table of anatomy terms taken from the Mouse Gene Expression Database at the Jackson Laboratory (specifically GXD mouse stage 28 - adult.) However, it has been extended to incorporate human anatomy and also revised in a number of areas, particularly the haemato-lymphoid system, based on the 37th edition of Gray's Anatomy, and the brain, thanks to the contributions of Dr. Jonathan Nissanov of Drexel University. Further value has been added to the vocabulary by mapping each anatomy term onto the relevant set of EST libraries in dbEST.

Ontology

Refer to ProDom and CDD which lists proteins associated with Gene Ontology defined "molecular functions." A heuristic algorithm was implemented. "The utility of these associations is that novel sequences can be assigned a putative function if sufficient similarity exists to a ProDom or CDD domain for which one or more GO functions has been associated."

Also you can access MGED (Microarray Gene Expression Data) ontology at http://www.mged.org/. The Microarray Gene Expression Data (MGED) group is a grass-root movement to promote the adoption of standards in microarray experiments and data. More specific goals are to facilitate the establishment of gene expression databases, comparability of microarray data from different sources, interoperability of different functional genomics databases and data analysis software. See http://www.cbil.upenn.edu/Ontology/biomaterial_1.4.jpg for graphical representation.

Contacts

- Christian Stoeckert, PhD. PI
  Research Associate Professor, Genetics.
  stoeckrt@pcbi.upenn.edu
- Brian Brunk. IT Project Leader, SR.
  brunkb@pcbi.upenn.edu.

COHSE – the Conceptual Open Hypermedia Project

Project

The goal is to research methods to improve "quality, consistency and breadth" of linking of WWW documents at retrieval time (as readers browse the documents) and authoring time (as authors create the documents). The COHSE (Conceptual Open Hypermedia Services Environment) implements three leading-edge technologies:

- "an ontological reasoning service which is used to represent a sophisticated conceptual model of document terms and their relationships;
- a Web-based open hypermedia link service that can offer a range of different link-providing facilities in a scalable and non-intrusive fashion;
- the integration of the ontology service and the open hypermedia link service to form a conceptual hypermedia system to enable documents to be linked via metadata describing their contents."

Ontology

Based on DAML+OIL

They are building ontologies and thesauri for document metadata. Metadata falls into three broad categories:

- "Catalogue information: *e.g.* the artist or author, the title, a picture's dimensions, a document's revision history;
- Structural content: *e.g.* headings, titles, links; for a picture its shapes, colours and textures;
- Semantic content: *e.g.* what the document/picture is about *e.g.* football, sport, person holding trophy, hope, joy."

The aim is to development an ontology service is capable of (semi-)automatic production, deployment and maintenance, and the semi-automatic cataloguing of documents with that ontology such that the cataloguing is discriminatory and sufficiently expressive.

Contacts

- Leslie Carr, Ph.D.
  +44 (0)23 8059 4479
  lac@ecs.soton.ac.uk

- Carole Goble, Ph.D.
  +44 (0)23 8059 4479
  cag@cs.man.ac.uk

EcoCyc

Project

EcoCyc is a Pathway/Genome Database for E. coli. It descrives the metabolic and signal-transductions pathways of E. coli, its enzymes, its transport protein, and its mechanisms of transcriptional controls.

EBI – European Bioinformatics Institute

Project

GOA is a project run by the European Bioinformatics Institute that aims to provide assignments of gene products to the Gene Ontology (GO) resource.

The goal of the Gene Ontology Consortium is to produce a dynamic controlled vocabulary that can be applied to all organisms, even while knowledge of the gene and roles of proteins in the cells are still be elucidated. In the GOA project, this vocabulary will be applied to a non-redundant set of proteins described in the SWISS-PROT, TrEMBL and Ensembl databases that collectively provide complete proteomes for Homo sapiens and other organisms.

The Gene Ontology resource contains a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. The Gene Ontology Consortium has developed three separate ontologies, molecular function, biological process and cellular component, to describe gene products and these allow for the annotation of molecular characteristics across species. Each vocabulary is structured as directed acyclic graphs (DAGs), wherein any term may have more than one parent as well as zero, one, or more children. This makes attempts to describe biology much richer than would be possible with a hierarchical graph.

Currently the GO vocabulary consists of more than 11,000 terms, which will. SWISS-PROT has joined the Gene Ontology (GO) Consortium and has adopted its standard vocabulary to characterize the activities of proteins in the SWISS-PROT, TrEMBL and InterPro (Apweiler et al., 2001) databases. It has initiated the Gene Ontology Annotation (GOA) project to provide assignments of GO terms to gene products for all organisms with completely sequenced genomes by a combination of electronic assignment and manual annotation. By annotating all characterised proteins with GO terms the SWISS-PROT group anticipates a valued contribution to biotechnological research through a better understanding of all proteomes.

InterPro

InterPro is a useful resource for whole genome analysis and has already been used for the proteome analysis of a number of completely sequenced organisms including *preliminary* analyses of the mouse and human genomes.

SWISS-Prot

The SWISS-PROT Protein Knowledgebase is a curated protein sequence database that provides a high level of annotation (such as the description of protein function, domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.

Contacts

- Rolf Apweiler, PhD. SWISS-PROT Coordinator
  apweiler@ebi.ac.uk

- Evelyn Camon, PhD. GO Curator
  camon@ebi.ac.uk

- Nicola Mulder, InterPro Coordinator
  mulder@ebi.ac.uk

- Wolfgang Fleischmann, Automation Coordinator
  wlf@ebi.ac.uk

FLYBASE

Project

FlyBase is a database of genetic and molecular data for Drosophila. FlyBase includes data on all species from the family Drosophilidae; the primary species represented is Drosophila melanogaster.

FlyBase is produced by a consortium of researchers funded by the National Institutes of Health, U.S.A., and the Medical Research Council, London. This consortium includes both Drosophila biologists and computer scientists at Harvard University, University of Cambridge (UK), Indiana University, University of California, Berkeley, and the European Bioinformatics Institute.

FlyBase includes the following:

-Information on more than 46,000 mutant alleles of more than 24,500 genes
-Information about the expression and properties of over 6,600transcripts and 3,000 proteins
-Information on the functions of gene products
-Over 14,400 nucleic acid accession numbers linked to genes
-Over 6,200 protein sequence accession numbers

The FlyBase project is carried out by a consortium of Drosophila researchers and computer scientists at Harvard University, University of Cambridge (UK), Indiana University, University of California, and the European Bioinformatics Institute. A complete list of consortium members is included below.
FlyBase is supported by the National Human Genome Research Institute (U.S.A.), grant P41-HG00739. Additional support for FlyBase is provided by the Medical Research Council (U.K.), grant G9535792MB.

Contact

- Michael Ashburner, PhD.
m.ashburner@gen.cam.ac.uk

Fm -Foundational Model of Anatomy

Project

The Fm is a symbolic representation of anatomical relations with inheritance of structural attributes. The Fm was initiated as a anatomical enhancement of the UMLS Semantic Network. It provides explicit descriptions of structural knowledge and makes this knowledge available to humans and machine interfaces for clinical, educational, and research applications.

Ontology

Based on anatomical relationships. The Foundational Model utilizes four different classes of relationships. It can be summed up as *Fm = (AO, ASA, ATA, MK)*. There is the AO - the Anatomy Onotology (hierarchy of *is-a* relationsps); the ASA - Anatomical Structural Abstraction (structural relationships); the ATA - Anatomical Transformation Abstraction (relationships that describe pre- and postnatal development); and the Mk - Metaknowledge (rules for representing relations in the other three areas).

Fm is represented in Protégé-2000, an ontology tool that is based on the Open Knowledge-Base Connectivity (OKBC) protocol. Protégé implements classes, instances of said classes, slots (representing attributes of classes and instances), and facets (to define slots). Also utilizes a metaclass hierarchy of templates.

Contact

- Jose Leonardo Villaraza Mejino Jr., M.D.
  mejino@u.washington.edu

- Cornelius Rosse, M.D.
  rosse@u.washington.edu

GALEN

Project

GALEN technology is utilized in the management of clinical information. The goal of GALEN is to organize/manage data so that clinicians can store and access appropriate and specific clinical information; and to enable computers to "manipulate" the stored information for perform a variety of tasks such: organizing, comparing, or presenting.

Ontology

"Partonomies" are utilized to represent part-whole relationships. The GRAIL description logic/common logic reference is implemented in the GALEN. Semantic links (or attributes) can be declared as transitive and inheritable. Semantic links and concepts can be organized in a is-kind-of hierarchy.

Contact

- Alan Rector, Ph.D.
  Medical Informatics Group
  Department of Computer Science, University of Manchester
  Tel +44-161-275-6133, FAX +44-161-275-6204
  rector@cs.man.ac.uk
  www.cs.man.ac.uk/mig/

GO - Gene Ontology Consortium

Project

To produce a dynamic controlled vocabulary that can be applied to genomes of all eukaryotes even as current research and discovery is producing vast amounts of novel data. There are several organizations currently contributing to GO. The first three to lay the foundation were FlyBase (database for the fruitfly *Drosophila melanogaster*), *Saccharomyces* Genome Database (SGD - database for the budding yeast), and Mouse Genome Database and Gene Expression Database (MGD and GXD - databases for the mouse *Mus musculus*).

Ontology

GO is actually three ontologies. These are controlled vocabularies for the description of the Molecular function, the Biological process, and the Cellular component of gene products. The terms are used as attributes. These are used to annotate collaborating databases, thus allowing for uniform queries between various groups.

Function, process and component are represented as directed acyclic graphs (DAGs) or networks. There are also several browsers available to search the GO (e.g. MGI GO browser, AmiGo browser) which are supplied by various collaborating databases.

"GO is not a way to unify biological databases." It promotes the sharing of vocabulary. Aspects such as domain structure, 3D structure, evolution, etc. are not included in GO.

Contacts

- Midori Harris, Ph.D.
  midori@ebi.ac.uk

- Micheal Ashburner, Ph.D.
  m.Ashburner@gen.cam.ac.uk

Genia Ontology

Project

The GENIA project seeks to automatically extract useful information from texts. Their initial methods are customized for micro-biology knowledge, but they forsee that their "basic methods should be generalizable to knowledge acquisition" in other fields.

Current work is on extracting event information about protein interactions. Such a task requires accessing many different sources of knowledge, thus necessitating the development tools such as a parser, ontology, thesaurus, domain dictionaries, and a learning model.

Ontology

The GENIA ontology is intended to be a formal model of cell signaling reactions in human. It is to be used as a basis of thesauri and semantic dictionaries for natural language processing applications such as:

- Information retrieval and filtering
- Information extraction
- Document and term classification & categorization

“Another use of the GENIA ontology is to provide the basis for integrated view of multiple databases including CSNDB developed at National Instiitute of Health Science. "

The current GENIA ontology is based on a taxonomy of entities involved in molecular reactions. It was developed as a semantic classification. Figures map out the GENIA ontology, included in the figures are links that point to notes for annotations.

Contacts

- Jun’ichi Tsujii, PhD. PI
  tsujii@is.s.u-toko.ac.jp

- Yuka Tateisi, Research Associate - GENIA (Ontology development, parsing)
  yucca@is.s.u-tokyo.ac.jp

GeneX Ontologies (NCGR)

Project

NCGR and the Computational Genomics Group at the University of California, Irvine, are participating in the GeneX™ Project to provide an Internet-available repository of gene expression data with an integrated toolset that will enable researchers to analyze their data and compare their results with other such data. This body of data will allow more confidence to be placed on the conclusions reached through analysis, as well as sharing the considerable cost of generating these datasets.

With large-scale sequencing comes the ability to query organisms for their partial or complete transcriptomes, the transcriptional response to a challenge. The myriad technologies for this are quite expensive but can provide huge amounts of data, only a small amount of which is generally of interest to the investigator.

The GeneX Project plans to make the greatest use of gene expression data by creating an Internet-available relational database of public data derived from these multiple technologies, as well as making the same database technology available for local installation.

Ontology

Refer to link for schema:
http://www.ncgr.org/genex/doc/GeneXSchema.pdf

Contact

GENEX PROJECT TEAM
- Honghui Wan, Project Manager

- Todd F. Peterson, Software Developer

- Jiaye Zhou, Adjunct Scientist and Software Developer

IMGT, the international ImMunoGeneTics database

Project

IMGT is an integrated information system specializing in Immunoglobulins, T cell receptors, and Major Histocompatibility Complex (MHC) molecules of all vertebrate species, created in 1989 by Marie-Paule Lefranc (Université Montpellier II, CNRS).

IMGT works in close collaboration with EBI. IMGT consists of sequence databases: IMGT/LIGM-DB, a comprehensive database of IG and TR from human and other vertebrates, with translation for fully annotated sequences, and IMGT/HLA-DB, a database of the human MHC, genome and structure databases (IMGT/3Dstructure-DB), Web resources, and interactive tools. The IMGT server provides a common access to all Immunogenetics data.

Contact

Marie-Paule Lefranc
Professor Université Montpellier II
Laboratoire d'ImmunoGénétique Moléculaire, LIGM
UPR CNRS 1142, IGH
141 rue de la Cardonille
34396 MONTPELLIER Cedex 5 (France)
e-mail: lefranc@ligm.igh.cnrs.fr
Tel: +33 (0)4 99 61 99 65
Fax: +33 (0)4 99 61 99 01

MGED – Microarray Gene Expression Data Group

Project

MGED is developing an ontology to adopt a standards for DNA-array experimentation. Pulling in work by groups sequencing and decoding the *Arabidopsis Thaliana*, bacteria, archea, *Gallus gallus*, Drosophila, fungi, and homo sapiens.

Contact

- Christian J. Stoeckert, Jr., Ph.D.
  Research Associate Professor
  Dept. of Genetics
  Penn Center for Bioinformatics
  stoeckrt@pcbi.upenn.edu

PharmGKB – The Pharmacogenetics andPharmacogenomics Knowledge Base

Project

PharmGKB is a collaborative effort created at Stanford University and funded by the NIH. The aim of the project is to aid researchers in understanding how genetic variations elicit differences in responses to drugs. PharmGKB is a research tool that is available to the public, its is basically a central repository for clinical information. It consists of a database of clinical information and genomic, molecular/cellular phenotype data from research participants at different medical centers.

“The NIH Pharmacogenetics Research Network currently funds clinical and basic pharmacokinetic and pharmacogenomic research in the cardiovascular, pulmonary, cancer, pathways, metabolic and transporter domains. In addition, network members are actively involved in understanding the ethical, legal and social issues that surround pharmacogenetics and pharmacogenomics research and are developing appropriate policies to guide the PharmGKB. “

Contacts

- Russ B. Altman, MD, PhD. Principal Investigator
  russ.altman@stanford.edu

- Teri E. Klein, PhD, Director
  teri.klein@stanford.edu

RiboWeb Project

Goal to represent biological data for molecular modeling. One of the fundamental goals of modern molecular medicine is to understand how the structure of biological macromolecules produces their function. Understanding for this function comes from multiple experimental, theoretical and statistical data sources that appear in the literature. RiboWeb links models and structural information with experimental data sources. Initially the project focused on the 30S ribosomal subunit and now has been expanded to structural data to the entire prokaryotic ribosome. "RiboWeb is composed of the following integrated information resources:

- A knowledge base containing an ontology-based representation of the primary data relevant to the structure of the ribosome as well as supplementary functional data. In particular, we have encoded the main experimental results of approximately 200 articles that are key for ribosomal structure modeling in this knowledge base.
- Links within the knowledge base to the Medline references reporting these data and the special-purpose databases containing ribosomal sequences (*e.g.*, the Ribosomal Database Project) as well as secondary and tertiary structures (*e.g.*, the Protein Data Bank).
- Software components that test for compatibility and consistency between the primary data and structural models; that compute, display and evaluate new models based on user-specified interpretations of the primary data; and that use the knowledge base for other purposes such as supporting intelligent literature searching.
- A structured query constructor, designed to be used by biologists, that guides the user in finding the information in the knowledge base for which s/he is searching.
- An XML-based syntax designed to exchange basic RNA molecular information"

Ontology

RIBOWEB is an online knowledge-based resource that supports the creation of three-dimensional models of the 30S ribosomal subunit. It has three components: (I) a knowledge base containing representations of the essential physical components and published structural data, (II) computational modules that use the knowledge base to build or analyze structural models, and (III) a web-based user interface that supports multiple users, sessions and computations.

Contact

- Richard O. Chen, Ph.D.
  rchen@smi.stanford.edu
- Ramon Felciano, Ph.D.
  Felciano@smi.standford.edu
- Russ B. Altman
  altman@smi.stanford.edu

Signal Ontology

Project

Catalogs cellular transduction/signaling pathways of all model systems.

Ontology

Utilizes a Protégé-2000 similar system/browser. Refer to: http://marine.ims.u-tokyo.ac.jp:8086/~spark/SO/

References

Refer to the following link:
http://www.hgc.ims.u-tokyo.ac.jp/organize/takagi/List_of_papers.html

Contacts

- Toshihisa Takagi, PhD. PI
  takagi@ims.u-tokyo.ac.jp

- Takako Takai, PhD
  takako@ims.u-tokyo.ac.jp

STAR – mmCIF

Project

'CIF' is an acronym for the Crystallographic Information File. CIF is a subset of STAR (Self-defining Text Archive and Retrieval format). It is used for archiving all types of text and numerical data. The aim of CIF is to expand upon its "compatibility, flexibility, and to incorporate these in electronic publications."

They have developed a dictionary of data items sufficient for archiving small molecule crystallographic experiments and its results. This dictionary was adopted by the IUCr at its 1990 Congress in Bordeaux. CIF is now the format in which structure papers are submitted to *Acta Crystallographica C*. Software has been developed to automatically typeset a paper from a CIF.

Ontology

"STAR defines a set of encoding rules similar to saying the English language is comprised of 26 letters. A Dictionary Definition Language (DDL) is defined which uses those rules and which provides a framework from which to define a dictionary of the terms needed by the discipline. Think of the DDL as a computer readable way of declaring that words are made up of arbitrary groups of letters and that words are organized into sentences and paragraphs. The DDL provides a convention for naming and defining data items within the dictionary, declaring specific attributes of those data items, for example, a range of values and the data type, and for declaring relationships between data items. "
[from: P. E. Bourne, H. M. Berman, B. McMahon, K. D. Watenpaugh, J. Westbrook, and P. M. D. Fitzgerald. The Macromolecular Crystallographic Information File (mmCIF). *Meth. Enzymol.* (1997) **277**, 571-590.]

Contacts

- John D. Westbrook
  jwest@rcsb.rutgers.edu
- Philip E. Bourne
  bourne@sdsc.edu

TAIR

Project

TAIR (The Arabidopsis Information Resource) is a "searchable relational database" cataloging scientific work done on the *Arabidopsis thaliana*. *Arabidopsis thaliana* is used a plant model with several advantages that allow it to be a model organism for studies of the cellular and molecular biology of flowering plants. TAIR is a collaboration between the Carnegie Institution of Washington Department Plant Biology, Standford University, and the National Center for Genome Resources. Grant No. DBI-9978564 by the National Science Foundation.

Contact

- Carneige Institution: Dept. of Plant Biology **650/325-1521**

- Sue Rhee, Staff Associate ext. 251
  rhee@acoma.stanford.edu
- Tanya Berardini, Curator ext. 325
  tberardi@acoma.stanford.edu
- Suparna Mundodi ext. 342
  smundodi@acoma.stanford.edu

TAMBIS Project

TAMBIS applet utilizes a full version of a conceptual model of molecular biology and bioinformatics. The model will enable you to form descriptions in the following areas:

- motifs in proteins;
- protein similarities;
- protein tertiary and secondary structure;
- enzymes, their substrates, products and cofactors;
- various functions and types of nucleic acids.

The goal of Tambis is to offer a stable system for queries and to access the Swiss-Prot, Enzyme, Cath, Prosite, and Blast bioinformatics resources.

Ontology

A program to answer queries by retrieving information. It does so in a conceptual fashion. Meaning it asks the browser to define/describe concepts so it may retrieve instances of said concept. TAMBIS only displays those concepts and relationships which can be held by the context stated. It does this by asking the user to model the "parents, children, and siblings of the concept, as well as what relationships to which other concepts are valid."

TAMBIS allows queries to be formed over several knowledge bases.

Contacts

- Norm W. Paton, PhD., PI
  norm@cs.man.ac.uk

- Carole Goble, PhD
  c.goble@cs.man.ac.uk

Virtual Head and Neck Anatomy WorkshopProject

The intent of the workshop was to address the feasibility of developing a digitized, 3-D, interactive head and neck anatomy program to be made available on CD-ROM and the Internet. This goal would be evaluated by using the Visible Human database as guide for possible recommendations.

Ontology

Rosse Foundational Model.

Contacts

Brent Jaquet, Director
Office of Communications and Health Education at NIDR.
brent.jaquet@nih.gov

Annex 4 -
Conclusions of the Proceedings from the Open Source Software Framework for Organ Modeling and Simulation Conference
July 23-24, 2001

(a) Information Technologies Have Become Essential

Living systems, and the human body in particular, are the most complex systems known. A deep understanding of how they function will give us unprecedented power to improve health. Mimicking biological systems will give us an extraordinary set of new tools that could increase economic productivity while reducing pollution and our need for natural resources.

But unlike other areas of science, our understanding of biological systems cannot be reduced to insights captured in a few equations. As we probe deeper, the systems appear ever more intricate and more diverse. Understanding these systems requires both an enormous number of detailed experiments and finding a way to tie this information together and make sense out if it. The explosion of information available from sequencing entire genomes and growing sophistication in many other fields means that simple models of behavior are being replaced with more complex, more realistic models involving the interaction of thousands of phenomena. Few important phenomena are likely to be explained by a "one-gene theory", for example. Most disease states can be understood only by following the interaction of many different genes working in complex networks.

The complexity of this kind of analysis has grown to the point where biological systems can best be understood by using modern computers. Computers were essential for sequencing the human genome and will be even more important in understanding how the genome works by developing computer simulations of their functions. These tools can also make it easier to visualize the operation of complex systems – how cells assemble the miniature machines they need or how defects in electrical networks degrade the performance of a heart – and see what may happens by intervening with new drugs, surgeries, or other therapies.

(b) Inventing a New Model for Collaboration

Building the software needed to describe the dynamic operation of cells and organs requires developing a new way of representing the research results. Just as the definitions of how to represent information on the Internet led to the growth of the World Wide Web, the Digital Human project will develop a language allowing research teams to combine their results and build simulations capable of addressing complex, practical problems.

Such simulations have already proven themselves capable of producing useful results. Computer modeling provides crucial help for biomedical research, the design of artificial hips and organs, anticipate the effects of crash tests, design robots, and create animated humans for games and movies. Much more can be achieved in the next few years.

Progress will, however, be much faster if the diverse community now developing simulations is able to work together efficiently and developers can save time by building on each other's work. But without an effective community, and a common vocabulary, most of these projects must start from scratch. Simulations can and must be built with out the Digital Human consortium. But an effective community will make the process faster, reduce duplication of effort and costs, and reduce bugs and errors. The community would ensure that diverse groups benefit from each other's work, and from the testing and bug reporting that would result from widespread use and testing. The core mission of the Digital Human Consortium is helping such a community to form and operate effectively.

The Digital Human Consortium will provide a forum and a framework to develop models and simulations that can interoperate for larger scale modeling of complex systems such as gene regulatory networks and multi-level organ systems. The consortium will ensure that the models and simulations are valid and accurate, and it will provide a framework allowing interoperation and reuse of models and simulations that developed by a diverse research community. The work must combine many disciplines including computer science, cell biology, molecular biology, physiology, pathology, pharmacology and anatomy.

There should be no illusions about the difficulty of this task. The consortium recognizes the need to advance incrementally but the virtue of doing so within the framework of a broad reference model. While useful products can be expected from the tools developed by the Digital Human consortium during the next few years, we may never have a complete understanding of human systems. Countless discoveries are needed in biology and medicine and new information tools must be developed. But now is the time to begin.

(2) The Utility of Biological Simulation

Biological simulations are being built for a wide range of purposes including medical research, education and training, medical practice, robotics and biomimetics, and human factors. All would benefit from a shared set of valid simulation tools that would prevent duplication of effort – letting each group spend more time on solving the problems that interest them most instead of working on software. Here are some examples.

(a) Biomedical Research

Modeling and simulation has been used for years in biomedical science as an adjunct to experimental work performed in "wet lab" experiments. Data gathered from *in vitro* and *in vivo* studies can be analyzed *in silico* and combined with insights from many other experiments to generate new hypotheses that can be tested in the laboratory. Most computer models and simulations have been developed in isolation and few attempts have been made to share computational strategies and data outside of conventional publication channels.

Recently, several communities of researchers and clinicians have realized the benefits of working in consortia working on models that span multiple levels of biological organization, integrating anatomy, physiology, biomechanics, cell biology and biochemistry. These include integrated models of the vertical organization of some of the major organs (heart, lung, muscle) as well as horizontally-integrated models of major physiological systems (circulatory, respiratory, immune). Visualization and simulation technology may soon allow users to move seamlessly between different spatial resolutions (molecular to organ level) and different temporal states (development through aging; varying physiologic state) within an integrated simulation.

Simulations of the Heart

Investigators working in cardiac modeling and simulation provide a particularly compelling example. Sophisticated models of cellular, tissue and organ systems have been built from a variety of data sources: diagnostic images, electrophysiological measures, biomechanics, bioelectric fields and ionic studies. The teams have used this model to build sophisticated simulations that provide insight into the physiology of the heart not possible from studies limited to a single level of analysis. The models have, for example, allowed a detailed understanding of the mechanisms of heart disease, such as arrthymias, ischemia and myopathy that allow them to explore a range of potential new strategies for therapies. Clancy and Rudy (1999), for example, showed that a mutation in the SCN5A gene produces a structurally defective sodium channel that causes cardiac arrhythmia when inserted into an integrated, quantitative computer model of a cardiac cell.

Modeling the Molecular Biology of the Cell

A significant application this strategy will be development of a context in which to understand the function of new gene products derived from the human genome project, genes can be screened for normal and abnormal function (so-called “phenotype screening”) using validated computer models and simulations of cells and organs. Thus, a candidate gene product whose function is unknown can be inserted into the requisite computational model, and the consequences of its expression can be studied within these higher order simulations.

Success in sequencing the human genome, as well as the sequencing of many other animal and plant species, has greatly accelerated research to understand the complex functions of individual genes, and the way the expression of one gene can affect the actions of others. Understanding these operations requires understanding complex sequences of operations that are in many ways analogous to complex electric circuits. Several genes may need to be expressed, and several others suppressed, for a biological function to occur.

Simulations allow researchers to assemble information that has been gathered about the functions of many different genes, and their reaction to their environments, and understand how networks of hundreds of genes operate together. These simulations allow experimental biologists to make conjectures about the responses of complex biological processes in a simulated environment, without having to conduct studies *in vitro*, on animals, or in human patients. These predictions, of course, eventually need to be validated *in vivo*. But the models provide a powerful tool to help point *in vivo* research in the most promising directions.

For example, predictive models of generic cell types such as red blood cells, eukaryocytes and prokaryocytes could be used to screen the effects of novel drugs in pharmacological research, identifying candidate drugs that show efficacy on simulated receptors in simulated cells. Similarly, patient-specific organ models could be developed from diagnostic images and physiologic data and used to predict the effect of novel pathogens on the individual tissues of a particular patient. Von Dassow et al (1999) showed the value of predictive modeling in biology, when their *simulated* Drosophila embryo was able to generate accurate patterns of developmental segmentation, based solely on the activity of 136 coupled equations with 50 parameters for the processing of gene products.

Clinical Practice

Accurate computer models can play a key role in developing new medical procedures, helping physicians plan radiation therapies, design prosthetics and artificial organs, and communicate with patients and other health providers.

Interventional Planning

The computer models built by members of the Digital Human consortium will provide reference standard for image analysis, anatomical landmarking, pathological classification, image-guidance for therapies and procedures, and patient comparison. The generic models can be extended to represent models of individual patients by using information from a variety of new imaging devices (MRI, CAT, PET). These simulations can, for example, combine new imaging modalities and the development of computer-based diagnostic systems for detection of tumors and other lesions. These models can allow surgical teams to plan procedures on accurate models of an actual patient's condition and aid therapists planning to target tumors with specific doses of radiation or chemicals. The models could greatly reduce risks and errors.

In the long run, Digital Human simulations can speed the development of new drugs and therapies. Accurate models would let physicians explore the impact of different therapies on the specific pathology and disease condition of an individual to be displayed and customized for very individualized therapies. These may include heart surgeries, customized drug interventions, and tumor and cancer resections, with full knowledge of the exact spread of the problem and the margins of safe and effective therapy.

Artificial Organs and Prosthetics

Computer models are already being used to design artificial hips, hearing aids, prosthetics and other devices fitted precisely to the requirements of individual patients. The Digital Human will provide a reference model that would increase the accuracy and validity of these designs, as well as speeding the development of a much wider variety of devices. By combining a vast amount of measured information into a single model, the Digital Human simulations would provide a powerful tool for learning how to mimic the operation of human organs – whether the heart, or kidneys, or the ear. They would also help ensure an accurate interface between artificial organs and the environment in which they will function (including their performance under extreme conditions that would be otherwise difficult to test).

A New Kind of Medical Record

'Body-double', patient-specific image models can be created to serve as a repository for diagnostic, pathologic and other medical information about a patient. These will serve as a three-dimensional (3-D) template for enhancing communication between patient and physician, and provide a reference framework to examine pathologic and age-related changes that occur over time.

a) Medical Training and Education

Computer simulations are becoming critical for extracting meaning from the complex information emerging from biological research. It is also becoming critical for students to learn this material for the first time, and to help experts keep pace with discovery. Much of the information about biological operations can be made much more vivid, and understandable, if it is shown visually. Text and two dimensional drawings in texts and journals can not convey information as forcefully as a simulation that allows a student to see the full dimensions of something like a heart, see how the components operate, and understand the impact of different diseases and clinical interventions. At the microscopic level fo the cell, the operations of organelles, cell walls, self-assembled motor structures can be simulated and visualized in compelling ways. Simulations allow students to explore and practice in ways that do no harm. And they make it possible for students to understand the diversity of biological systems helping prepare them to expect the unexpected.

Medical Schools

Medical schools are finding it increasingly difficult to attract new instructors willing to teach introductory courses – particularly human anatomy. Departments of Anatomy are being abolished or incorporated into other departments. The generation of basic science faculty adept at teaching gross anatomy is dying out. Graduate programs in anatomy no longer require training and teaching, but rather emphasize research in neurobiology, molecular biology and cell biology.

While the simulations made possible by the Digital Human consortium can obviously not provide a comprehensive solution, they could provide crucial new tools. Powerful simulations can let students learn more about the structure and function of anatomy than traditional techniques. The new tools would permit a new kind of pedagogy – based on exploration and apprenticeship – much more powerful than conventional work with texts and the occasional cadaver. The simulations could capture the expertise of existing teachers and give new teachers room to invent new tools and new approaches to instruction built around state-of-the-art models of human function captured in simulations built for research purposes.

Achieving this kind of instruction, of course, would require a unique collaboration beteen computer scientists, cognitive scientists, anatomists, physiologists to develop a new generation of models, simulations, educational programs that can support true user interaction with simulated human organs, including validated physical and physiological properties, such as real-time tissue deformability, realistic bleeding and accurate haptics ("touch and feel"). These simulators will support high bandwidth access will facilitate distributed visualization and simulation of models for medical education and research and development applications.

Continuing Education for Surgeons and Other Medical Specialists

One immediate benefit of an integrated Digital Human will be to provide simulators for practicing difficult procedures for medical professionals at all levels.

There is a growing public awareness that physicians and other healthcare workers make mistakes. Many of these mistakes are purely technical in nature; sometimes these errors are fatal. Recent studies suggest that up to 100,000 Americans die every year from medical errors. The future trend is toward even greater liability risks, regulatory oversight, and higher entry-level skills. Repeated certification and skill demonstration is now obligatory. Complex surgical procedures such as hip replacement, skull base surgery, complex liver surgery, can be rehearsed in the virtual environment using the patient's anatomy prior to the actual procedure, and health practitioners can be certified using accurate models and simulations based on the Digital Human. Medical schools are struggling to remain solvent. Academic medical centers are urgently seeking cost-effective solutions to expensive training and residency programs. Thousands of medical personnel throughout the world need to train and practice invasive procedures. The cost of using operating room time for training surgical residents has been estimated at $53 million in the United States alone. Opportunities to learn and practice these vital skills on animals and humans diminish as public expectations rise at the same time as hospitalizations and length of stay decrease.

Computer-based medical simulation can be used to train healthcare providers in a spectrum of medical skills from planning and diagnostics, through minimally invasive procedures, up to the most complex, high-risk procedures. The advent of high performance computing on the desktop, coupled with the enhanced realism of computer graphics models of the human body, makes this technology available now for *safe and effective* training. Simulation can be used to bridge the information gap between patient and textbook and between practitioners and patient for patient education.

a) Biomimetics and Robotics

Biological systems perform extraordinary feats that could open revolutionary new dimensions in computing, data storage, environmentally benign chemical manufacturing, and many other areas. Robot designers continue to struggle to imitate aspects of locomotion, cognition, and navigation mastered by the most simple animals. These efforts could be greatly assisted by Digital Human simulations that provided powerful explanations of the operation of real biological systems.

b) Human Factors

Many engineering designs are based on models of their impact on humans can operate safely and effectively. These can range from the design of vehicle seats and parachute harnesses to the design of safe cockpits and automobiles. Accurate simulations could predict the impact of a wide variety of extreme events on the human body. Combined with mechanical simulations of vehicles, the Digital Human simulations could predict the impact of a variety of extreme events on the human body (side collisions, rapid acceleration). They could even anticipate the impact of phenomena that can not be measured directly.. such as the impact of prolonged weightlessness in a long-duration NASA mission and the effectiveness of different interventions.

3) What Must Happen to Build the Digital Human Consortium

The Digital Human consortium will build simulations capable of achieving these ambitious goals by providing a forum where a diverse group of developers can share, test, and build on each other’s work. Researchers will be able to express new insights into the role of a specific gene in a language that would permit easy integration with other work. Drug designers, clinicians, teachers, human factors experts and others would be able to draw on validated, up-to-date simulations build by others and apply their creative energies to using the tools to achieve specific goals. Under current circumstances, each group builds redundant models.

But getting to a point where many groups can contribute to, and share in the Digital Human model, requires (1) building a community that could define the technical, legal, and other aspects of sharing, and (2) designing specific technical tools for ensuring interoperability of components (tools that define the interface between components, for example, and represent the geometry of objects in ways that permit a viewer to represent the combined operation of all components.)

The Digital Human is a software consortium that is building a collaborative approach for the design and development of biomedical simulations and models. In this scenario, developers of a heart model would be able to plug their software into another group’s lung model, and these models could interact in a meaningful and accurate simulation of actual cardiovascular-respiratory interaction. Similarly, software components modeled after molecules, cells and tissues could be integrated in a hierarchy to produce a valid representation of a functional organ such as a heart or liver. To achieve this goal, it is critical that developers engage in a collaborative software development process in which biomedical models and simulations are verified and validated by the larger biomedical research community.

a) Building the Community

.

There's no hiding from the daunting difficulty of improving communication among of the diverse, creative individuals and groups working in areas related to biological simulation. While funding agencies can encourage participation, in the long-run the Digital Human consortium will succeed only if it presents unambiguous benefits to the participants and if the transaction costs of participation – primarily the investment of precious time – are low. The minimum goals of a successful community are:

- A process for developing a technical architecture permitting the widest possible collaboration and sharing/reuse of components.
- Simple, clear rules for managing intellectual property
- Efficient procedures for peer-review and testing, bug reports, issue tracking software/biological validation, and procedures for releasing approved versions
- Easy procedures for version control, managing continuous build _ test _ revise cycles
- Clear identification of authors, sources of data and methods (both to trace and correct problems and to ensure adequate credit is given to creators)
- Ease in building business around extensions and services

The experience gained by the Open Software community provides a valuable model. The Mozilla process, for example, has resulted in successful projects even in projects involving millions of lines of code and a thousand developers.[1] It proves that given the right incentives, a diverse group of developers can maintain their independence and creativity while gaining enormous efficiencies by sharing each other's work. New information tools can greatly facilitate the process by making it easier to share work and conversations online and providing semi-automated checks of technical validty.

Few simulation projects in biomedical research benefit from sharing interoperable software components. In most cases individual researchers are managed as stand-alone, "stove-pipe", projects. But there is a growing sense that the complexity of the task has made this style of operation increasingly inefficient and frustrating for the participants.
In the Digital Human Consortium there will be stringent requirements not just for technical validity of the code but strict peer review and evaluation to ensure that the underlying biological models are valid. Careful procedures to verify the sources and accuracy of data used to build biomedical models and simulations are essential if the

[1] Frank Hecker, "Lessons from Open Source Software Development:The Mozilla Experience",
Proceedings of the Open Source Software Framework for Organ Modeling and Simulation Conference
July 23-24, 2001

tools are ever to be adopted as a legitimate platform for experimentation and clinical practice. But if the open consortium operates as hoped, the number of reviewers and valuators can be very large, and the process of review and improvement can be continuous.

4) Our Proposal

We propose to build a management process for the Digital Human Consortium that will roughly follow the successful model of large-scale open source projects. Ideally the funding agencies would be comprised of the following elements:

- Senior officials from public agencies (and companies) funding major portions of the code development would constitute a policy making board of directors.
- A Steering Committee would be appointed to manage the day-to-day operation of the project, including managing the required collaborative web-sites and data-bases. These people would work nearly full time on the project, would *facilitate* (and importantly not *direct*) the process, gaining consensus on policies and procedures, making “tie-breaking decisions” when disputes arise, and ensuring consistency among the projects.[2]
- Individual development efforts would be organized by “Project Leads” (also known as “component owners” or “module-owners”) having primary responsibility for a given component (e.g., “liver”, “user interface tools”, etc).

The Project Lead would typically work with a handful of other developers (say, 5-9 individuals); the Project Lead and his or her associated developers would together be the primary individuals responsible for creating the code and related material associated with their component. (Although other individuals may contribute code for use with the component, based on experience in open source projects the Project Lead and associated developers will likely produce 90% or more of the code and other material associated with the component.) Project Leads would have permission to enter and change code in the official version of the Digital Human; they may also approve such access for other individuals, including developers on their own teams.

In addition to being responsible for the technical development of their own modules, Project Leads would also be responsible for coordinating with the Project Leads for other modules, to ensure that the work performed by their team is coordinated with work performed by other teams. An overall Architecture Committee (or Technical Coordination Board), consisting of the Project Leads from all of the components of the project (or a representative subset thereof), would be responsible for overall technical decisions related to development activities for the Digital Human.

[2] We propose that the following individuals serve as initial members of this group: Adam Arkin, Brian Athey, Jim Bassingthwaighte, Parvati Dev, Tom Garvey, Frank Hecker, Gerry Higgins, Chris Johnson, Henry Kelly, Bill Lorensen, Andrew McCulloch, Ken Salisbury, Shankar Sastry

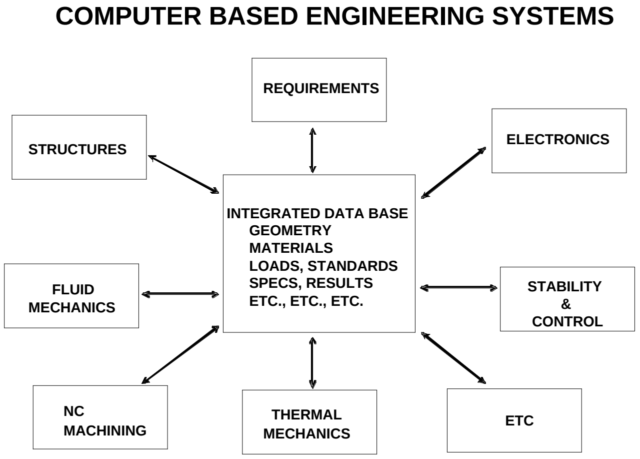
The small teams responsible for the various components would encourage collaboration and participation from a much larger group of people who would contribute components and review the work. This larger group would include several hundred individuals from academic, government, or industry research groups and would not necessarily be associated with any of the funding agencies. The members of the team would be authorized to work with pre-release versions of the Digital Human code, design documents, bug reports, and other project material, but would not have permission to change the official version of the Digital Human project code and data.

a) Technical Architecture

A key element of the Digital Human project will be to ensure that software components developed by different developers will work efficiently together. This means, for example, that a functioning heart model could be assembled by combining simulations of valves and other heart components built by different groups – and those individual components are easy to replace. A valve modeling the characteristics of a particular individual could, for example, be substituted for a generic valve.

An effective technical architecture would:

- Encourage creative, competing solution
- Adaptable to new concepts and discovery and accommodate existing models and simulations, while providing guidance for models yet to be developed.
- Not tied to a specific platform or programming language
- Highest possible compatibility with existing models.
- Rooted in biology (principles of biological organization, structured by natural representations of ontology and object interaction)-- no forced programming artifacts
- Minimize bureaucratic and computational overhead
- Accommodate both vertical components (e.g., modules at organ, tissue, cell levels) and horizontal components (e.g., user interface, security).



The engineering community has developed sophisticated approaches for developing technical architectures. The STEP standard[3], for example, provides a way to create drawings and simulations

ieering Systems and Data,
ıg and Simulation Conference

of complex aircraft and other systems that may involve thousands of components and hundreds of different designers. While the details of the way these systems manage geometry, pass information about fluid flows, and other aspects of visualization and simulation will differ, the experience these groups have had in developing functioning, interoperable components will be examined closely.

Over the long term, a large number of projects (under the "project leads" described above) will need to be developed. Topics will include

- Developing a unified ontology that would permit clear identification of components from gross anatomy to molecular components of cells
- Defining geometry so that components fit together properly and provide a precise basis for modeling physical connections and material flows.
- Defining models of physical motion and deformation
- Defining signal flows (chemical, electrical)
- Defining material flows
- Defining chemical transformations (including gene expression)
- User interface tools (including visualization, tools for building circuits of gene expression, etc.)
- Applications (teaching tools, research tools, human factors models)

Undoubtedly many more topics, and subtopics, will need to be introduced over time. Since it will not be feasible to undertake a complete set of these tasks at the beginning, we propose that the Digital Human project begin with four projects: (1) a unified ontology, (2) Defining a geometry model, (3) Building gene expression networks, and (4) Building post-secondary teaching tools.

(b) Anatomy Training and Surgical Simulation

The first application team to be formed will focus on Anatomy Training and Surgical Simulation, as this has been identified as a priority by the meeting's participants. The absence of qualified teachers in anatomy coupled with the obsolescence of the medical school basic science curriculum, suggests that this is one of the most important application priorities that could be targeted by the Digital Human Consortium.

(c) Unified Ontology

Ontology is an explicit specification of a conceptualization. For the Digital Human, it is necessary to define the objects and relationships that represent all of the molecular, cellular, tissue, organ and system objects. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based software program represents knowledge. Thus, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Most biological simulation models are founded on a sharply defined ontology, which allows a terse mapping of biology onto computer architecture. This is an important source of the power of such models.

A great deal of effort has been focused on the development of ontology in biology. For example, the Gene Ontology Consortium develops knowledge representation for eukaryotic cells (see http://www.geneontology.org/). Another example is the Bionome project, (http://www.ibc.wustl.edu/moirai/moirai.html) which models biochemical reactions and pathways that are representations as interactions of concentrations, without spatial distribution except as separated into compartments.

In contrast to these efforts, the Digital Human needs to develop an ontology that can unify both higher-level organ models and lower-level molecular and cellular models. As a first task, it is suggested that the Digital Anatomist Foundational Model of Rosse et al (1998; http://www1.biostr.washington.edu/~onard/AMIApapers/D005094.pdf), which specifies higher order structures and their relationships, with the emerging BioSPOICE ontology being developed by Garvey, Lincoln and Arkin.

While most work in ontology has focused on providing precise descriptions of objects such as organs, tissues, and cells, it will also be important to build systematic descriptions of the processes and actions of these components. The Biospice project, for example, will define chemical flows and transformations in cells. An analogous non-spatial ontology is typical for whole-body metabolic models such as QCP (http://www.biosim.com/: named for “quantitative circulatory physiology”), which specifies interactions between certain endocrine concentrations, blood pressure, etc., and simulates interventions like hemodialysis, change in diet, change in environment, various pumps, drips, stimulators and pharmacological agonists and antagonists. The system quantifies the homeostatic actions of many organ-systems, but it only names some chemicals in the chains: it contains no anatomical maps. Similarly, the Cardiome seeks to “Integrate biophysical models of the cardiac action potential, excitation-contraction coupling, and cross-bridge cycling into tissue and organ-level models and develop a unified, Web-based interface to these cellular models that can serve as a common entry point to a database of model parameters”, requiring what a model ‘is’ to be pre-defined. This aims at a tightly integrated structure for the collective model, where the internal structure of a part follows as standard pattern.

(d) Geometry

Biological objects, such as cells, tissues, organs and organisms have some geometric features that are difficult to model in a realistic manner using conventional engineering methods. Since modeling involves simplification, engineering approaches such as STEP may provide a useful framework for static and certain dynamic models of organs and their relationships. More complicated behaviors such as deformation may be modeled using well-understood, physics-based models.

A fundamental property of the Digital Human will be to coordinate spatial interactions between different models and simulations. A reasonable, highest-common-factor geometrical communication standard for surfaces (membranes or volume boundaries) is the triangulated mesh, specifying at least (*x*,*y*,*z*) positions for vertices and listing triples of vertex IDs to give triangles that will move with them. All other geometrical descriptors can be used to generate such a mesh, with variable levels of detail. While it is hard for a model whose internal description scheme is a NURBS (Non-uniform Rational B-Spline) patchwork to generate one automatically from mesh data, it should be able to handle collision with an object whose shape is specified this way, and accept and use its transfer messages. Other surface descriptors with significant usage in the Digital Human community should have standards by which a model may communicate them, but an agreed mesh format is basic, and should be defined early on in the process.

Similarly, every model involving a deformable volume should be able to export information about it in terms of a mesh of tetrahedra, the natural solid generalization of triangles. Many other primitives are possible, but all can be 'factorized' into tetrahedra, while few can be exactly re-expressed in terms of others. In general, inclusion of formats should come from consensus rather than an isolated committee's preference for some form with advantages for particular modeling purposes.

The language for curves (center lines of blood vessels, *etc*.) must clearly include 3D networks with straight segments between vertices. Some more curvilinear formats are in wide use, such as piecewise cubic polynomial curves fitted together as B-splines – which of these formats to include in a first version of the Digital Human standard is a matter for discussion.

(e) Cell Modeling and Simulation

The BioSpice project is designed to produce interoperable, open-source simulation and verification tools for intracellular circuits and intercellular communication: given a circuit (with proteins, regulatory genes, *etc*., specified), the program will simulate concentrations and synthesis rates. In each of these schemas, a molecule 'is' a concentration represented by a number, and interacts with other concentrations by kinetics with a defined set of rates. One of the goals of the Digital Human consortium is to integrate various physical levels of analysis, and this includes integration of cellular, molecular, organ and systems-level phenomena.

information society technologies

FET - Future and Emerging Technologies

DIMACS — Center for Discrete Mathematics & Theoretical Computer Science

European Research Consortium for Informatics and Mathematics
ERCIM
www.ercim.org

This workshop is part of a series of strategic workshops to identify key research challenges and opportunities in Information Technology. These workshops are organised by ERCIM, the European Research Consortium for Informatics and Mathematics, and DIMACS the Center for Discrete Mathematics & Theoretical Computer Science. This initiative is supported jointly by the European Commission's Information Society Technologies Programme, Future and Emerging Technologies Activity, and the US National Science Foundation, Directorate for Computer and Information Science and Engineering.

More information about this initiative, other workshops, as well as an electronic version of this report are available on the ERCIM website at http://www.ercim.org/EU-NSF/