Font size: Times New Roman 10 all through, Bold Titles and Authors, Underline presenting author

THE INFOGENMED PROJECT

A. Sousa Pereira[(1)], Victor Maojo[(2)], Fernando Martin-Sanchez[(3)], Ankica Babic[(4)], Sofia Goes[(5)]

(1) University of Aveiro/IEETA, (2) Polytechnical University of Madrid (UPM),
(3) Institute of Health Carlos III (ISCIII), (4) Linköping University, (5) STAB Vida - GENÓMICA

Introduction

The Human Genome Project has launched numerous efforts to expand research in the biomedical sciences and explore new application areas. Thus, new technologies and tools are needed to extract data, information, and knowledge that can be applied to create novel diagnostic and therapeutic procedures in medicine.

The development of biochips and other technologies as well as the growth of Internet, are also contributing to expand medical applications of genetic information. Researchers at different world sites are working towards developing new informatics tools to ensure an optimal integration of genetic and clinical information. These tools are not currently widely available. Thus, various companies from the medical informatics industry are developing new models of computerized medical records to integrate genetic and clinical data. Those data can be available at different remote sources with different formats and logical structures.

A new breed of systems and software tools is necessary to convert the enormous amount of data that geneticians and molecular biologists can obtain at their labs in information that physicians and health workers can use. In the latter environments, professionals will have to manage a kind of information that they are not familiar to. Not only they will need methods to search, access, and retrieve genetic information but also methods to gather, classify and interpret such information.

The main problems to integrate all the biomedical knowledge are:

- Many different sources of relevant information are spread over the Web; information needs to be located, accessed, and retrieved;
- Exchange of data is difficult since databases can present a wide range of formats;
- Coding and terminology are not unified, being sometimes difficult to discern quality;
- Medical coding systems are not ready for managing genetic information;
- Existing bioinformatics tools are designed for researchers;
- Physicians lack of guidance in accessing relevant data.

Taking into account the above-presented background a consortium of 5 organizations submitted a proposal to the Information Society Technologies Programme (IST) of the EU, being a project approved on April 2002.

Objectives

The main objective of the project is to build a virtual laboratory for accessing and integrating genetic and medical information for health applications, which we have denominated INFOGENMED. Two main issues will be involved in this project: the integration of heterogeneous medical and genetic databases over the Internet, using an innovative approach, and integrating medical and genetic terminologies in a vocabulary server.

Users of the INFOGENMED will be able to directly connect with remote medical and genetic databases, located at separate places. The user interface will be accessible using any standard Web browser, connected with a server of medical and genetic terms, based on Artificial Intelligence techniques, such as, for instance, conceptual graphs. This server will provide user navigation and searching capabilities in multiple remote databases, in different languages. Once the information is located at heterogeneous remote databases, a set of tools creates a unification of the contents of those databases, creating a virtual database. Users will have the perception that they are working with a single, local database through a Web browser, that is, in fact, the virtual database created by the system. In addition, the interface tool can provide collaborative capabilities to allow the cooperation between people located at remote locations.

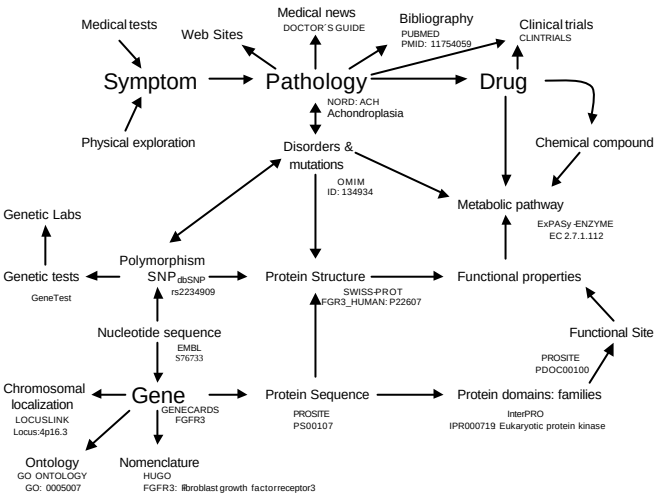

Figure (1) Clinical pathway for Achondroplasia

The information retrieved from remote databases will be stored for its integration with local health or clinical information, stored by the specific users. Results will be validated at different sites, involving health practitioners and patients. We have selected, as application domain, the integration and clinical use of information about rare genetic diseases. These diseases are subject to many investigations during the last years because of the achievements of the Genome project and their importance in public health. As mentioned later, these diseases have not been the focus of a great part of past research and there is an increasing interest in them.

Figure (1) shows the clinical pathway for one rare genetic disease, giving an example of the integration of clinical and genetic information.

We can thus summarize the project objectives:

- Determination of the needs of genetic and medical information in various pathologies, considered as "rare genetic diseases", in health environments;
- Design of the methods and development of tools for the integration of heterogeneous databases over Internet;
- Design and implementation of an interface to aid users to search, find and retrieve the contents of remote databases, based on a vocabulary server for the integration of medical and genetic terms and concepts;
- Development of an assistant to help health practitioners to use the designed methods and tools;
- Integration of the complete system and validation in the area of rare genetic diseases.

Methodology

As previously stated the main objective of INFOGENMED is to facilitate the access to information from remote heterogeneous databases to users of the system, through a standard Web browser-based interface. This interface will be linked with a server of medical and genetic vocabularies that will provide users with navigation and searching. Once the data and information are located a set of tools creates a unification of the contents of those databases, building a virtual database.

We can identify different types of databases which are relevant for the project objectives:

- Clinical (patient information) PHRs
- Medical (disease information) Rare Diseases
- Genetics (dbSNP)
- Hybrids (OMIM)

A preliminary study was performed identifying relevant web-sites for the INFOGENMED. Some examples are presented:

Databases with clinical and genetic information

-GeneCards
http://bioinfo.weizmann.ac.il/cards/index.html
GeneCards is a database of human genes, their products and their involvement in diseases. It offers concise information about the functions of all human genes that have an approved symbol, as well as selected others.

-OMIM - http://www.ncbi.nlm.nih.gov/omim/
This database is a catalogue of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins and elsewhere, and developed for the World Wide Web by NCBI, the National Center for Biotechnology Information. The database contains textual information, pictures, and reference information.

-GeneReviews - http://www.geneclinics.org/
(formerly GeneClinics), online publication of expert-authored genetic disease reviews with International genetics Lab Directory and International genetics and prenatal diagnosis Clinic Directory

-GDB–The Genome Database – http://gdbwww.gdb.org

-GenAtlas - http://bisance.citi2.fr/GENATLAS/ is a repertory of three types of objects: genes, diseases, and markers

-Medline - PUBMED – http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubM
ed MEDLINE is a bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences.

-HGMD – Human Gene Mutation Database - http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html at the Institute of Medical Genetics in Cardiff. The Human Gene Mutation Database (HGMD) represents an attempt to collate known (published) gene lesions responsible for human inherited disease.

HUGO Mutation Database
http://www.genomic.unimelb.edu.au/mdi/centraldb.html
The goal of the HUGO Mutation Database Initiative is to collect and disperse accurately, efficiently, and completely a record of predominantly disease causing variation together with other variations where relevant (e.g., inorganic variation).

Nomenclature and coding systems

-Human Gene Nomenclature Database - http://www.gene.ucl.ac.uk/nomenclature/
Currently providing approved symbols and literature Aliases for over one third of the genes in the human genome

-Gene Ontology - http://www.geneontology.org/
The goal of the Gene OntologyTM Consortium is to produce 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.

-ICD-9-MC - http://www.cdc.gov/nchs/about/otheract/icd9/maint/maint.htm The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) is based on the World Health Organization's Ninth Revision, International Classification of Diseases (ICD-9). ICD-9-CM is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States.

-MeSH – http://www.nlm.nih.gov/mesh/meshhome.html
MeSH is the National Library of Medicine's controlled vocabulary thesaurus. MeSH consists of a set of terms or subject headings that are arranged in both an alphabetic and a hierarchical structure. There are more than 19,000 main headings in MeSH.

Databases on rare diseases

-Rare Diseases at the Spanish Institute of Health "Carlos III" – CISATER http://cisat.isciii.es/er/

-ORPHANET - http://orphanet.infobiogen.fr/

-NORD – National Organization for Rare Diseases
http://www.rarediseases.org/cgi-bin/nord

Conclusions

The project started on the 1st September, being the first results available the end of February 2003. These results include the Analysis of the state of the art, User requirements and the Functional analysis of the system. Further information and results can be obtained in project website: http: //infogenmed.ieeta.pt