Practical Approaches to the Development of Biomedical Informatics: the INFOBIOMED Network of Excellence

Hans-Peter Eich[a], Guillermo de la Calle[b], Carlos Diaz[c], Scott Boyer[d], A. S. Peña[e], Bruno G. Loos[e], Peter Ghazal[f], Inge Bernstein[g] and the INFOBIOMED Network

[a] *Coordination Centre for Clinical Trials, Heinrich-Heine-University, Düsseldorf, Germany*
[b] *Artificial Intelligence Lab./Facultad de Informática, Universidad Politécnica de Madrid, Madrid, Spain*
[c] Biomedical Informatics Research Group, Municipal Institute of Medical Research - IMIM, *Barcelona, Spain*
[d] *AstraZeneca R&D Mölndal, Mölndal, Sweden*
[e] *Academic Center for Dentistry Amsterdam (ACTA), Universiteit van Amsterdam, Vrije Universiteit and VUmc Amsterdam, Amsterdam, The Netherlands*
[f] *Scottish Centre for Genomic Technology & Informatics,University of Edinburgh Medical School, Edinburgh, United Kingdom*
[g] *The Danish HNPCC-register, Surgical Department, Hvidovre Hospital, Hvidovre, Denmark*

Abstract

Biomedical Informatics (BMI) is the emerging discipline that aims to facilitate integration of Bioinformatics and Medical Informatics for the purpose of accelerating discovery and the generation of novel diagnostic and therapeutic modalities. Building on the success of the European Commission-funded BIOINFOMED Study, an INFOBIOMED Network of Excellence has been constituted with the main objective of setting a structure for a collaborative approach at a European level. Initially formed by fifteen European organizations, the main objective of the INFOBIOMED network is therefore to enable the reinforcement of European BMI at the forefront of these emergent interdisciplinary fields. The paper describes the structure of the network, the integration approaches regarding databases and four pilot applications.

Keywords:
Bioinformatics; Medical Informatics; Biomedical Informatics; Genomic Medicine; Research Networks; Data Integration; Ontologies

1. Background

Since the 1960s Medical Informatics (MI) has been established as an independent discipline, and some universities have consolidated specific departments and training programs. It was clear at the time that a new discipline was emerging, with professionals trained in fields outside classical medicine, e.g. computer science, decision analysis, engineering, or economics [1]. Computer scientists, mathematicians, and engineers joined MI to begin a professional career in this field. Such professionals, medical informaticians, began to merge models and techniques from computer science fields, like artificial

intelligence, with knowledge about patient care. Later, the development of applications such as electronic health records, expert systems, hospital information systems, multimedia programs and many others have contributed to establish MI as a scientific discipline.

In the biological area, the consolidation of genetics as a scientific discipline, based on principles such as Mendel's Laws and the discovery of the physical structure of DNA led to an increasing amount of data and information that needed to be stored and analyzed. In the 1960s, this growth of data and the availability of computers led to the beginning of the discipline of computational biology [2]. A convergence appeared later, gathering topics from biology, biochemistry, engineering, and computer science, leading to Bioinformatics (BI). Some pioneers began to apply informatics methods and computer tools to molecular biology problems, even a decade before DNA sequencing was feasible. It was also shown that computers could dramatically speed up sequencing and determination of protein structures. Rapidly, BI began to develop. For instance, GENBANK, a DNA sequence database, was created in 1980 and SwissProt, for proteins, a few years later [3]. These and other computer systems led to the acceptance of BI as an independent discipline.

MI and BI have had problems in obtaining scientific recognition. Computer scientists considered that applied informatics is just a branch of computer science, and some biomedical professionals viewed informaticians as mere developers of computer programs without a real scientific merit. In this sense, advances in genomics might dramatically change this traditional perception. The techniques needed to advance genomic medicine might come from the intersection of these four areas: MI, BI, medicine and biology. That is the reason why a new area, Biomedical Informatics (BMI), is being brought at the intersection of both MI and BI to create a synergy between them. Only combined studies of gene interactions in humans and large epidemiological studies from many different populations can discover the complex pathways of genetic diseases. In such a postgenomic era, it is presumed that it will be easier to determine the risks of some specific populations in relation with certain diseases. Thus, personalized prevention and therapeutics could be established, with patient-customized pharmaceuticals or diet [4]. Therefore, a close collaboration between researchers in MI and BI can contribute to new insights in genomic medicine and towards the more efficient and effective use of genomic data to advance clinical care [5]. The need for integrated BMI efforts has been realized by different institutions. In this regard, the European Commission (EC) organized in 2001, in Brussels, a workshop to analyze the synergies between MI, BI and neuroinformatics. Later, efforts carried out at the EC and US institutions have led to various achievements. In Europe, the BIOINFOMED [6] study led to the publication of a White Paper about the challenges in BMI regarding genomic medicine. This study has led to launch the INFOBIOMED Network of Excellence (NoE) funded by the EC (6th Framework Programme) [7].

2. Research approaches in the network

In its initial phase, the NoE has carried out a comprehensive analysis of the state of the art in the areas of data models and ontologies, database integration, security, data mining and knowledge discovery, image analysis and processing, and decision support. These state of the art studies aim to provide basis for identifying the existing gaps and finding the best solutions that can be applied horizontally to the pilot applications. Although details cannot be fully reported in this paper, two areas - ontologies and data integration - have been considered as being crucial for the development of BMI solutions for common problems.

Ontologies were defined by Gruber as "explicit specifications of conceptualizations". They are much more than controlled taxonomies. They are conceived to represent the underlying meaning of a scientific domain or field. The OWL (Web Ontology Language), a proposed standard for developing and representing ontologies, linked to the concept of the

"Semantic Web", can be fundamental to link knowledge from many different and heterogeneous sources, including databases and documents accessible over the Web. The huge amount of data that biomedical genomic researchers must analyze generates important challenges that biomedical informaticians should address.

Bridges to overcome syntax and semantics gaps across different data sources are required for database integration. For instance, when the same concept is labeled with different names in different databases, it might be needed to map these names to the same concept within a specific ontology. In recent research approaches, specialized ontologies such as GeneOntology are used to create shared vocabularies, whose terms can be used to map terms from different database fields. For instance, the INFOGENMED project [8], funded by the EC, aimed to integrate public and private genomic and health databases to assist biomedical researchers in locating and accessing information over the Web.

Within the NoE, the use of ontologies in various areas related to database integration, data mining and information retrieval, is planned.

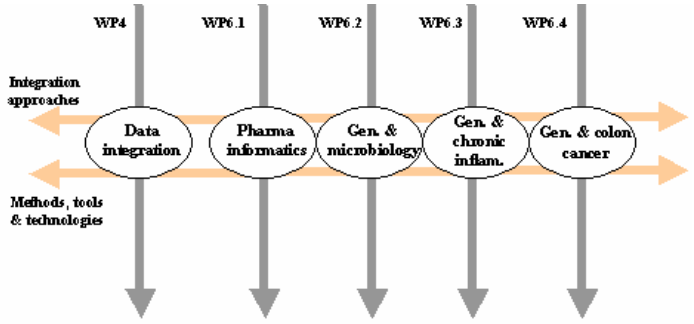

Figure 1-A diagram showing the horizontal and vertical integration approaches regarding database integration (WP 4), integration of tools, methods and technologies and the four pilot applications, WP 6.1 to 6.4

With respect to data integration, two different methods for heterogeneous database integration have been identified: data translation and query translation. In data translation approaches, data from heterogeneous databases at different sites are collected and placed in a local repository. In query translation approaches, queries are translated instead of data. Most recent approaches to database integration are based on a query translation approach.

Most approaches based on query translation can be classified into four categories: a) Pure mediation, using "mediators" and "wrappers". These are brokers between users and data resources; b) Single virtual conceptual schemas, using a unique semantic model to integrate the information available at the data sources. A broker module is responsible for retrieving, adapting, gathering and presenting the final results; c) Multiple virtual conceptual schemas, where each linked database is described with a different virtual schema; and d) Hybrid approaches, where a single schema is used to describe each data source. This schema is created using a shared vocabulary or ontology.

3. Pilot applications

The goal of the NoE is to be able to reuse BMI methods and tools within four pilot applications, showing the ability to carry out integrated ideas in different research domains.

Pharmainformatics, aimed at investigating the impact of BMI on the different stages of the drug discovery process, from target identification to clinical trials. The intensive use of new information technologies has been postulated as a way to accelerate and optimize the drug discovery process. To a large extent, information drives the development of new drugs, and the power of computers can be harnessed to sieve vast numbers of molecules with potential medicinal value. Computational procedures include the "in silico" creation,

characterization and filtering of molecular libraries. Computer-based "virtual screening" experiments can automatically assess the fulfillment of drug-likeness criteria or pharmacophoric patterns, as well as perform the simulated docking of large series of compounds to 3D models of their potential targets. A recent extension of the virtual screening strategy is the chemogenomics approach, which aims to link both chemical and biological spaces by a joint analysis of libraries of selected ligands and related targets. Early virtual screening of ADMET properties is an computational task that is becoming crucial for optimizing the flow along the drug discovery pipeline. The aim of this pilot is therefore to extend the use of existing bioinformatic /chemogenomic approaches and to link them to clinical data relating either to a specific disease or adverse event. The activities in this pilot focus on carrying out two specific research examples of how software, database / format and work processes available within the NoE relate and contribute to the drug discovery process. The two examples are: Complex Regional Pain Syndrome (CRPS) and Nuclear Hormone Receptors (NHRs).

They will illustrate the information continuum from pathology to pathway to target to ligand/approved drug, but they will be approached from different directions. In the case of CRPS, a top-down approach will be used. The starting point will be the pathology, i.e. CRPS, and the end point will be possible ligands/approved drugs. In the case of NHRs, the starting point will be the ligands/approved drugs and the end point is pathology/adverse events. In both cases, the goal of this pilot application is to identify gaps in technologies and information that can be focus of further research to improve the drug discovery process.

Genomics and infectious disease, focused on the study of host and pathogen genetic polymorphisms, protein interactions and transcriptional/translational control and how these impact on pathogen virulence and host immune responses to infection. To date in excess of 2000 viral and microbial genomes have been sequenced and genetic variation at the single nucleotide level of our genome is fast approaching eight figures. Comparative and functional genomic approaches combined with proteomic strategies are further helping to describe gene/protein interaction pathways. These recent advances are dependent on the use and development of novel BI tools. In medicine, the quantitative modeling of viral dynamics in patients treated with multi-drug regimes are gaining increasing effectiveness in treatment management. The determination of viral sequence variation for assessing escape mutants from therapeutic agents in individuals is fast becoming standard practice, e.g. in HIV infected patients. These advances require tools and new algorithm development in MI. Fundamental to the biology and virulence of an infection is a clear understanding of the host-pathogen interactions at the systemic and cellular levels and which opens new challenges and opportunities for advancing anti-infective therapies. These challenges and opportunities will require BMI approaches.

The general activities in this pilot are aimed at using pathway biology (of the interferon system) as a central hub for integrating BI and MI. Two distinct clinical relevant pathogens, HIV and Cytomegalovirus, will be used as exemplars in the pilot. Here, it will be necessary to further characterize a) the viral genome, load and dynamics at a given stage of disease and b) assess the host's genotype, e.g. polymorphisms in key genes defining the hosts innate resistance to viral infection and proliferation and those determining the efficacy of therapeutic drugs (and their combinations) in clinical use. Taking in account the higher complexity of the human genome, in this application the NoE concentrates on the interferon pathway, combining host and virus genotype data with clinical data in order to find new markers of host immunity and viral therapy resistance.

Genomics and chronic inflammation, aimed at investigating the complex susceptibility to adult periodontitis. About 10% of the adult population will develop severe forms of destructive, chronic periodontal disease (chronic periodontitis). This complex inflammatory disease is precipitated in susceptible subjects by infection of the periodontium (tooth

supporting tissue) by Gram-negative, anaerobic, mostly commensal oral microorganisms. Moreover, the environmental factor smoking contributes importantly to disease severity. Modifying disease genes determine the susceptibility of periodontitis. However, still very little is known about the interplay and relative importance of genetic factors, bacterial pathogens and environmental determinants, like smoking and stress. There is a great need to gain more insight in the complexity of periodontitis, to design new treatment strategies and devise preventive measures. Periodontitis is an excellent model to study complex chronic inflammatory diseases because of its multifactorial etiology (genetics, bacteria, and environment), relative high prevalence and broad and easy access to diseased patients' and normal tissues, genomic DNA, and access to the history of infections and other relevant data through patient records. The aim of this pilot is to build a periodontitis data warehouse based on patient information coming from different sources: genetics, infection, environment, intermediate phenotype and disease phenotype. This data warehouse will be explored by data analysis and data mining tools from the various partners.

Genomics and colon cancer targets at accumulating knowledge useful for the planning and organization of screening in families with a high-risk of developing colon cancer and supporting research on the subject. HNPCC (Hereditary Non Polypose Colon Cancer) is a dominantly inherited colorectal cancer syndrome, with a lifetime risk up to 90% of developing colorectal cancer for carriers of the genes. Furthermore there is an increased risk of developing endometrial cancer or cancers in the urinary tract. The aim of this pilot is to build-up a general IT-infrastructure based on open XML standards for communication, to link different kinds of medical departments working together in an HNPCC register. These standards should meet HNPCC needs (e.g. transmission of pedigrees including geno-phenotype), the needs of related fields (e.g. other onco-genetic diseases) and should be usable in different countries. Existing international standards will build a basis (e.g. HL7) for this purpose. In a proof of concept, the IT-infrastructure of the Danish HNPCC register will be transformed from mainly isolated databases with paper-based communication amongst them to linked and interoperable databases with XML communication.

4. Network structure

This program of activities requires the participation of organizations with distinct profiles. The INFOBIOMED NoE gathers European research groups with a strong background and experience. Also technological groups that will ensure the application of state-of-the-art information tools and technologies (ontologies, web services, etc.) and their interoperability are included. Finally, biomedical research labs will offer real data and validate the network tools and methods in a number of biomedical research areas, as described above.

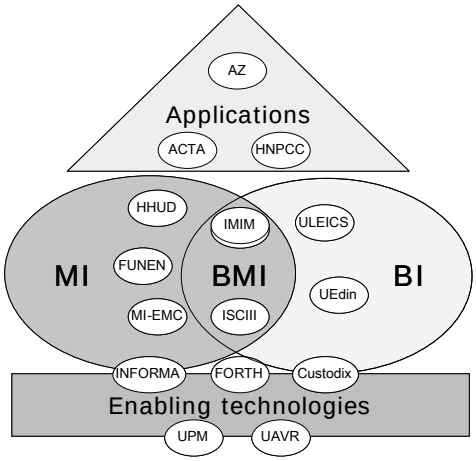

Figure 2-Profiles of the groups included in the network

The core of the INFOBIOMED NoE, formed by 15 renowned institutions that currently develop critical research in MI, BI and related fields, has been designed to offer the necessary critical mass to create a chain effect in the field at the European level that can foster the pursued integrative and structuring effort (figure 2): AstraZeneca R&D Mölndal (AZ), Academisch Centrum Tandheelkunde Amsterdam (ACTA), Danish HNPCC-Register (HNPCC), Heinrich-Heine University of Düsseldorf (HHUD), Municipal Institute of Medical Research (IMIM), University of Leicester (ULEICS), Danish Center for Health Telematics (FUNEN), University of Edinburgh (UEdin), Erasmus University Medical Center (MI-EMC), Institute of Health "Carlos III" (ISCIII), Informa S.r.l. (INFORMA), Foundation for Research and Technology – Hellas (FORTH), Custodix nv (CUSTODIX), Polytechnical University of Madrid (UPM) and the University of Aveiro (UAVR)

5. Conclusion

Following an initiative of the EC, INFOBIOMED offers an innovative networking approach that intends to exploit the potential synergies of already established scientific disciplines for the empowerment of an emerging one. This structuring effort seeks to deploy the promised benefits of the genomic revolution to society by combining the expertise and experience acquired through the, up to now, independent development of both BI and MI. Past and present integration initiatives in that respect suffer from an excessive isolation or only address the problem partially; INFOBIOMED represents the kind of global, integrative vision and joint effort required to overcome the obstacles that are delaying the development of true genomic medicine.

6. Acknowledgments

The present work has been funded by the European Commission (FP6, IST thematic area) through the INFOBIOMED NoE (IST-2002-507585).

7. References

[1] Schwartz, WB. 1970. Medicine and the Computer: The Promise and Problems of Change. *New England Journal of Medicine*, 283:1257-1264.

[2] Levitt M. 2001. The Birth of Computational Structural Biology. *Nat. Structural Biology*, vol 8, 5: 392-393.

[3] Bairoch A. 2000. Serendipity in Bioinformatics, the Tribulations of a Swiss Bioinformatician through Exciting Times. *Bioinformatics* 16: 48-64.

[4] Housman D. 1998. Why pharmacogenomics? Why now? *Nat Biotechnology* 16:492.

[5] Knaup P., Ammenwerth E., Brander R., Brigl B., Fischer G., Garde S., Lang E., Pilgram R., Ruderich F., Singer R., Wolff A.C., Haux R., Kulikowski C. Towards Clinical Bioinformatics: Advancing Genomic Medicine with Informatics Methods and Tools. *Methods Inf Med*: 2004 43: 302-307.

[6] Martin-Sanchez F., Iakovidis I., Norager S., Maojo V., de Groen P., Van der Lei J., Jones T., Abraham-Fuchs K., Apweiler R., Babic A., Baud R., Breton V., Cinquin P., Doupi P., Dugas M., Eils R., Engelbrecht R., Ghazal P., Jehenson P., Kulikowski C., Lampe K., De Moor G., Orphanoudakis S., Rossing N., Sarachan B., Sousa A., Spekowius G., Thireos G., Zahlmann G., Zvarova J., Hermosilla I., Vicente F.J. Synergy between Medical Informatics and Bioinformatics: Facilitating Genomic Medicine for Future Health Care. *J Biomed Inform*: Feb 2004 37(1): 30-42.

[7] www.infobiomed.org / www.infobiomed.net

[8] Babic A, Maojo V, Martin-Sanchez F, Santos M, and Sousa A. Ercim News. *Special Biomedical Informatics*. Jan 2005. nº 60. (www.ercim.org/publication/Ercim_News/enw60/)

Address for correspondence

Hans-Peter Eich, Coordination Centre for Clinical Trials, Heinrich-Heine-University Düsseldorf, Moorenstr. 5, 40225 Düsseldorf, Germany, Email: eich@uni-duesseldorf.de