P.C. de Groen

Division of Gastroenterology and Hepatology, Department of Internal Medicine
Division of Medical Informatics Research, Department of Health Sciences Research
Mayo Clinic College of Medicine
Rochester, Minnesota, USA

Synopsis

Towards Clinical Bioinformatics

Rapid advances in biotechnology have led to deeper understanding of the complexity of processes that define life. Examples include large scale sequencing projects such as the Human Genome Project, increasingly detailed and complex methods to measure gene expression, high throughput protein identification, 3-D protein structure and function analysis, and last but not least novel analytical tools and availability of ever expanding reference databases allowing various forms of data annotation and explanation. It is the hope that this deeper understanding can be translated into improvements in overall human healthcare. However, in order for this to occur, major progress needs to occur in a number of areas, examples of which are outlined in the five articles chosen to be incorporated in this section of the year book.

Before discussing the message within these articles, I will summarize some of the major deficiencies that currently limit our ability to incorporate the advances in biotechnology into the medical practice. First, although we now know the sequence of the human genome, we still have little understanding about the regulation of transcription, the folding and 3D structure of proteins, and most importantly the function of peptide domains, whole proteins or protein complexes. The same is the case for animal, plant, bacterial and viral molecules. Second, our understanding of molecular interactions is limited and decreases with increasing complexity as seen in signal and metabolic pathways, organelles, cells, organs or normally functioning or diseased people. Third, our ability to produce biological data, especially where it concerns gene sequence, gene expression and protein identification, has surpassed our capability to analyze and comprehend the results. So we need new, automated tools that will process and annotate the data, and provide meaningful and easily comprehensible visualization of the results. Fourth, the genome- (DNA as well as gene expression) and protein-related data only have meaning if they are coupled to or integrated with traditional, clinical practice-derived data (e.g., signs and symptoms, drug use, allergies, previous diseases, family history, laboratory results, etc.), the so-called phenotype. Integration with the phenotype in turn requires advances in electronic comprehension and analysis tools of the phenotype. Areas where major advances are needed include data architecture, comprehension of textual information (e.g., clinical notes, most often electronically stored in minimally structured or free text), computer assisted image analysis and metadata descriptions of all variables. And last but not least, the entire process – all steps as outlined above, need to be comprehensible to the practicing physician. This will require advanced preprocessing of data in the background unbeknownst to the practicing physician, reduction of data complexity with preservation of the key elements containing the essential information, and display of the information in a format that allows immediate decision-making in daily medical practice.

The mission of clinical bioinformatics (or medical bioinformatics, or biomedical informatics) in my opinion should be to provide the technical and scientific infrastructure and knowledge to allow optimized, individualized healthcare using all relevant sources of information (evidence-based medicine). Ideally, such healthcare is proactive, that is, based on the patient's genotype and behavior, it predicts development of possible disease; based on the prediction of disease it recommends intermittent diagnostic evaluations, and based on the results of all information then recommends changes in lifestyle, a medical regimen or procedures to maintain health rather than cure disease (see Figure 1). The information sources include the "classical" information as currently maintained in the health record (the "phenotype") as well as new tissue

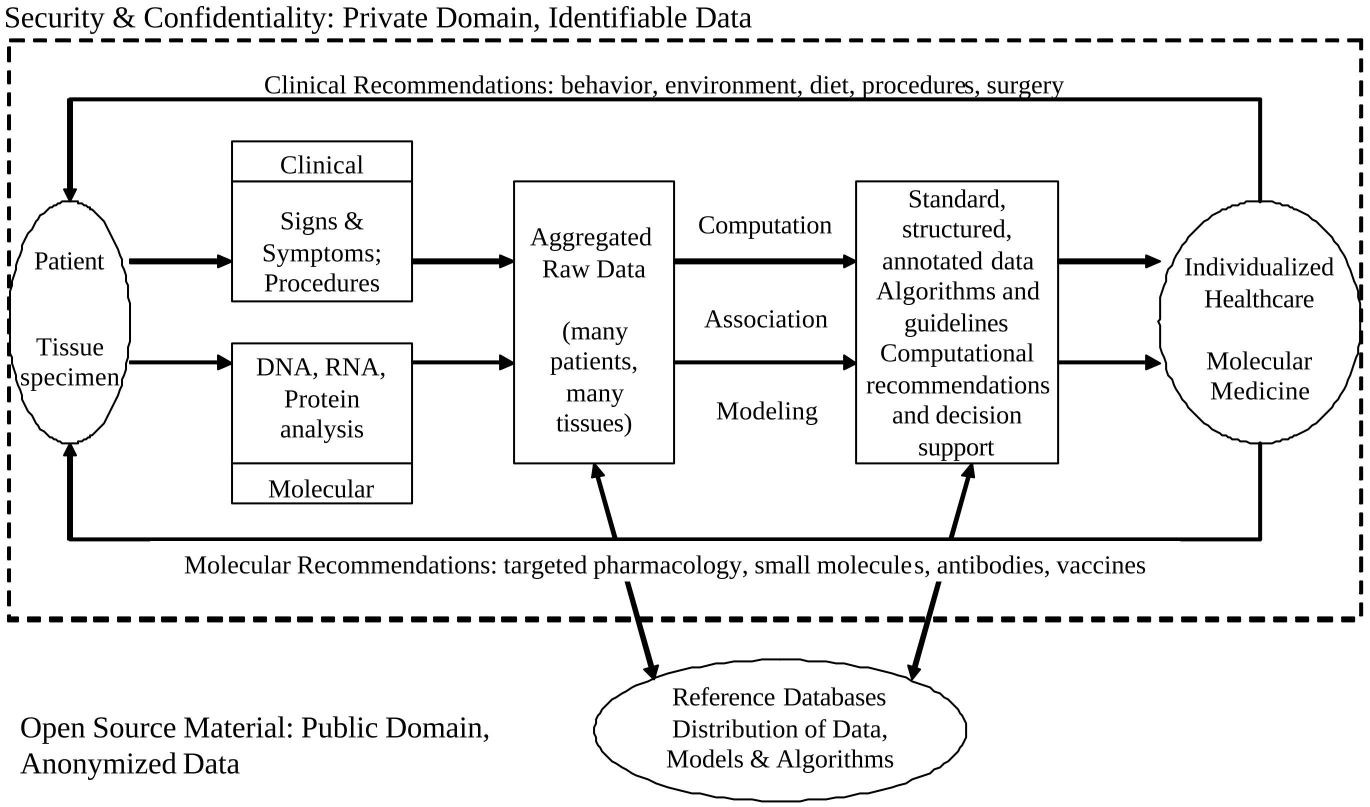
Security & Confidentiality: Private Domain, Identifiable Data
Clinical Recommendations: behavior, environment, diet, procedures, surgery
Patient
Tissue specimen
Clinical
Signs & Symptoms; Procedures
DNA, RNA, Protein analysis
Molecular
Aggregated Raw Data (many patients, many tissues)
Computation
Association
Modeling
Standard, structured, annotated data Algorithms and guidelines Computational recommendations and decision support
Individualized Healthcare
Molecular Medicine
Molecular Recommendations: targeted pharmacology, small molecules, antibodies, vaccines
Reference Databases Distribution of Data, Models & Algorithms
Open Source Material: Public Domain, Anonymized Data

Fig.1. Schematic representation of Clinical Bioinformatics. The upper half of the figure (from patient to individualized healthcare) represents the field of medical informatics; the lower half of the figure (from tissue to molecular medicine) the field of bioinformatics. In the merged view, these fields are combined and result in optimized, individualized healthcare based on clinical as well as molecular data. Only anonymized data will be distributed and included in reference database. Ideally, all models and algorithms become Open Source software.

and molecule-based information as soon as this is being collected. The change from late stage diagnosis (expensive cure) towards early detection or even prediction of disease (cheap cure or intervention) will not only improve the quality of the health of the individual, but likely at the same time reduce the overall costs of the healthcare systems.

A very important challenge for clinical bioinformatics is to extract from the continually expanding volume of information the knowledge, which will support, improve and change the medical decision-making process. The use of genetic and proteomic data in addition to clinical symptoms for medical decision-making will contribute to the expected, continued shift towards evidence-based medicine. This vision can only be realized with an enormous investment into (1) technology able to produce the genomic and proteomic data and the initial comparison of produced results with reference databases; (2) creation of standardized databases that combine clinical history, symptoms and signs, laboratory and procedural results, and genetic and proteomic data in raw as well as intelligently processed formats; (3) technology that assures confidential access to these data by those who need access (patient, healthcare providers, research staff), and full-proof security against unauthorized access; (4) extraction of knowledge out of these huge databases, their expert interpretation and matching against existing computational models; (5) development of novel explanatory and predictive models for the above, abstraction of the results to the clinical level, and incorporation of the extracted knowledge into algorithms and standardized clinical guidelines where feasible; and finally (6) implementation of the new guidelines into the clinical decision-making process.

The five articles within this section each describe components of the required infrastructure. The article by Antoniadis et al. deals with data complexity; it describes a data reduction method for microarray experiments that still permits efficient classification of the results. The value of understanding the relationship between phenotype and genotype is elegantly displayed in the article by Beerenwinkel et al. After careful analysis of sequence variations in the protease and reverse transcriptase genes in nearly 500 HIV type 1 virus isolates for 14 antiretroviral drugs, the authors were able to generate decision algorithms that are able to predict drug resistance of HIV type 1 virus for a subset of antiretroviral drugs based solely on genotype. Understanding the genotype of bacterial species and the interplay between the human and the bacterial genome is the topic of the article by Cariou et al. Nagl focuses on the need to understand the sequence-structure-function paradigm, to extract information out of integrated data resources, and to develop mathematical modeling for multivariable nonlinear dynamic systems. Both Nagl and Martin-Sanchez et al. stress the importance of data integration.

If the vision of clinical bioinformatics will come true then inclusion of tissue-based and molecular data will contribute to changes in practice standards and workflow in the healthcare system in general, in particular in the clinical decision-making process. On the one hand, such change supports the current trends towards use of cutting-edge information technology in healthcare and the creation of the networked healthcare systems. On the other hand, the design and architecture of today's electronic medical information systems will need a profound re-thinking, in order to be able to support the altered workflow and the new decision pathways of the era of "applied genomics".

How do we realize the mission? I propose research projects that demonstrate proof of concept. Such projects should have a limited number of clearly achievable deliverables, likely will be in the format of clinical trials, and should provide potential benefit to the patient. If results are as we expect - that is, the initial projects show proof of concept and indeed provide benefit to the patient - this information will need to be widely distributed to obtain the support of the average citizen, the patient, the healthcare providers, industry and the decision makers towards (1) development of a comprehensive clinical bioinformatics system based on common standards and (2) inclusion of all data of each citizen in order to create a system that is dynamically enhanced with addition of each bit of information collected within the population. Simply formulated, the main message should be "to apply the experience of the many to the benefit of the one", and to explain that this principle will contribute to the notion of disease prevention and maintenance of health. Such comprehensive system would continue to grow daily and its algorithms would continue to develop using the information obtained during healthcare provided to the population that is reflected within its databases. In addition, we need to convince the average citizen that the information we create and store is safe and secure, and will only be used to either optimize individualized healthcare (identifiable data – the unique patient being treated, any location) or used for creation of knowledge, disease algorithms and guidelines (anonymized data shared by researchers – algorithms and guidelines applicable to a specific disease or population). To accomplish these types of use, the information should be stored in systems that adhere to international or "de facto" standards. Lastly, rapid dissemination of the algorithms and guidelines, and subsequent implementation on a large scale with benefit to all citizens likely occurs quicker when the algorithms and guidelines are provided without copyrights as "Open Source" material.

I am indebted to Klaus Abraham-Fuchs, Peter Ghazal, Casimir Kulikowski, Georges de Moor, Stelios Orphanoudakis, and Gerd Spekowius for comments related to parts of the manuscript. The work was supported by the Mayo Clinic.

Address of the author:
Piet C. de Groen, M.D.
Consultant
Division of Gastroenterology and Hepatology
Department of Internal Medicine,
Division of Medical Informatics Research
Department of Health Sciences Research
Associate Professor of Medicine
Mayo Clinic College of Medicine
Program Director Mayo Clinic/IBM
Computational Biology Collaboration
Mayo Clinic
200 First Street S.W.
Rochester, MN 55905, USA
Tel: +1 507 284 3917 - secretary
Fax: +1 507 284 0538