Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 40 (2007) 5–16
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Methodological Review

Data integration and genomic medicine

Brenton Louie [a,*], Peter Mork [b], Fernando Martin-Sanchez [d], Alon Halevy [b], Peter Tarczy-Hornoch [a,b,c]

[a] Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, USA
[b] Department of Computer Science, University of Washington, Seattle, USA
[c] Department of Pediatrics, University of Washington, Seattle, USA
[d] Bioinformatics Unit, Institute of Health Carlos III, Madrid, Spain

Received 8 December 2005
Available online 9 March 2006

Abstract

Genomic medicine aims to revolutionize health care by applying our growing understanding of the molecular basis of disease. Research in this arena is data intensive, which means data sets are large and highly heterogeneous. To create knowledge from data, researchers must integrate these large and diverse data sets. This presents daunting informatic challenges such as representation of data that is suitable for computational inference (knowledge representation), and linking heterogeneous data sets (data integration). Fortunately, many of these challenges can be classified as data integration problems, and technologies exist in the area of data integration that may be applied to these challenges. In this paper, we discuss the opportunities of genomic medicine as well as identify the informatics challenges in this domain. We also review concepts and methodologies in the field of data integration. These data integration concepts and methodologies are then aligned with informatics challenges in genomic medicine and presented as potential solutions. We conclude this paper with challenges still not addressed in genomic medicine and gaps that remain in data integration research to facilitate genomic medicine.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Genomic medicine; Biomedical informatics; Data integration; Bioinformatics; Knowledge representation; Genomics

1. Opportunities and challenges of genomic medicine

1.1. Genomic medicine

There are many descriptions of genomic medicine in the literature [1,2]. At its core, genomic medicine attempts to elucidate the molecular basis of disease and then translate this knowledge into clinical practice for the benefit of human health. There are many potential implications of genomic medicine for health care [3–5], including: individualized healthcare based on genetics [4], predictive methods for disease susceptibility [6], new drug targets for currently untreatable diseases [7], gene therapy [8], and genetic/molecular epidemiology which will aid in the study of pathogen transmission and disease profiles of different populations [9].

The field of genomic medicine can be seen as a vast mosaic of related disciplines. Due to the rapidly changing nature of the field it would be impossible to completely cover the entire scope of genomic medicine, so for the purposes of this review we identify a subset of the disciplines where the informatics challenges are better understood: modern human genetics, which attempts to identify single-genes responsible for a genetic disease [10], pharmacogenetics and pharmacogenomics, which seek to understand how genes or systems of genes are involved in differential response by individuals to drug treatment [11], microarray researchers who look at the expression of thousands of genes at a time, possibly for the purposes of disease re-classification [12], rational drug design, which attempts to use all available biological, clinical, and chem-

* Corresponding author. Fax: +1 206 543 3461.
E-mail address: brlouie@u.washington.edu (B. Louie).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2006.02.007

ical knowledge to make informed development decisions [13,14], and clinicians who attempt to use "just-in-time" information for patient care [15].

1.2. Genomic medicine and data overload

Genomic medicine is, by definition, data intensive. The Human Genome Project [16] has spawned hundreds of publicly accessible databases [17] which grow larger and more numerous every year. There is also increasing diversity in the type of data: DNA sequence, mutation, expression arrays, haplotype, and proteomic, to name a few. Systems biologists, for example, deal with many heterogeneous data sources to model complex biological systems [18]. The challenge to genomic medicine is to integrate and analyze these diverse and voluminous data sources to elucidate normal and disease physiology.

1.3. Genotype-to-phenotype

Despite the disparate appearances of all the sub-disciplines of genomic medicine, there is a common thread: they are all, in some fashion, concerned with the connection between *genotype* and *phenotype*. A genotype is defined as an individual's genetic makeup, defined by his or her DNA sequence, and a phenotype can be defined as the "visible properties of an organism that are produced by the interaction of the genotype and the environment" [19].

In the context of genomic medicine, the genotype to phenotype connection can be loosely defined as which polymorphisms (changes in DNA sequence) or haplotypes (groups of polymorphisms) apply to which disease or differing responses of a genotype to treatment for a disease [20].

1.4. Genomic medicine and data integration

It is unlikely that any one satisfactory solution will arise that will solve all the informatics problems faced by researchers in genomic medicine. Nevertheless, as the common thread of the genotype-to-phenotype connection binds all sub-disciplines in genomic medicine, so may there be generalized data integration problems shared by each. It is important to identify these generalized problems as researchers in data integration attempt to solve just these sorts of challenges. In fact, research in data integration may have indeed provided some approaches and concepts that could prove to be valuable to genomic medicine. Some relief from data overload could be at hand by aligning the proper data integration technologies with appropriate, generalized, data integration problems in genomic medicine.

Data integration and genomic medicine are separate disciplines and have evolved in relative isolation. Our intent of this review is to look at the intersection between data integration and genomic medicine with intent to balance the computing and the biomedical and highlight potential bridges between the two disciplines.

2. Review of data integration approaches and concepts relevant to genomic medicine

There is much literature regarding data integration in the areas of biomedical informatics and computer science [21,22]. To complement this body of literature we highlight the data integration methodologies most relevant to data integration problems in genomic medicine. Note that we have tried to identify data integration concepts that are not simply "conceptual," but fairly stable technologies that can be readily applied to identifiable data integration problems related to the burgeoning field of genomic medicine. Many of these technologies were used in research projects that are now commercial systems such as DiscoveryLink [23], GeneticXchange [24], or TAMBIS [25].

Data integration is fundamentally about querying across different data sources. The different data sources could be, but not limited to, separate relational databases or semi-structured data sources located across a network.

Table 1
A summary of the advantages and disadvantages of data integration architectures

Architecture	Advantages	Disadvantages
Data warehouse	Fast queries Clean data	Stale data Complex schema Maintain extra copy of data
Database federation	Current data Flexible architecture No copying of data	Slower queries Complex schema Little or no data cleansing
Database federation with mediated schema	Current data Flexible architecture Schema tailored to users	Slower queries Little or no data cleansing Mappings from source schemas to mediated schema needed
Peer data management systems	Current data Flexible architecture Schema tailored to users Mappings between schemas distributed across peers	Experimental Slower queries Little or no data cleansing

Download English Version:

https://daneshyari.com/en/article/519106

Download Persian Version:

https://daneshyari.com/article/519106

Daneshyari.com