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.

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

Within data integration are two orthogonal dimensions which refer to where data or knowledge about meta-data *resides*, and the *representation* of data and data models. For the purposes of illustration we divide these dimensions into two axes: (1) integration architecture, and (2) data and knowledge representation (Table 1).

2.1. Axis 1: integration architecture (where data resides)

2.1.1. Data warehouses

Data warehousing is the consolidation of all specified data into a single database with a generalized, global schema. Data warehouses are considered to be reliable and provide a researcher with fast access and excellent response time to user queries. This is a non-trivial aspect since performance is often cited as a key feature by biologists [23]. Since importing of data is under local control, this facilitates easier cleansing and filtering of the data.

Consolidating all data into a single location does present problems. The volume of data may simply be too large for the warehouse to handle. Updating the warehouse presents serious maintenance issues and can create problems in that queries are only as relevant as the latest update [26]. Also, it is difficult to create a global schema that captures all the nuances of diverse data types. Often, as a consequence, the richness of the individual data sources is lost if one captures only the common elements in the global schema or alternatively the complexity of the global schema needed to represent all the source schemas becomes unwieldy.

Given these constraints, warehouses may be best suited for the creation of highly curated datasets focused on a specific and narrow area of research. The UCSC Genome Browser [27], the EnsEMBL Database Project [28], DataFoundry [29], and BioMolQuest [30] are examples of data warehouse approaches in biology. Chaudhuri and Dayal [31] also have published a survey of data warehousing technology.

2.1.2. Database federations

Unlike a data warehouse, a database federation leaves data at the source. In a database federation, underlying databases remain autonomous and may be distributed across a network. The federation maintains a common data model and relies on schema mapping for integration of the disparate source, usually facilitated by software programs that interface with the source [32,33], often called "wrappers" [34]. The federation appears to the user as a single database [35]. Federations relieve the "temporal" problems of warehouses since the data resides at the source and therefore is always updated. BioKleisli was a pioneer in applying these data integration approaches to biological data [36]. Using these approaches they were able to answer one of the "impossible" Department of Energy queries, namely: "Find for each gene located on a particular cytogenetic band of a particular human chromosome, as many of its non-human homologs as possible."

Data cleansing is difficult in a federation. No data is housed locally so data cleansing must be done on-the fly [37]. Performance can also suffer because it is dependant on the query load capacities of the other members of the federation. Federations use a common data model and therefore face the same difficulties as warehouses in representing diverse data types. Database federations are best suited for situations where a researcher requires the most up-to-date information, or where the researcher must integrate a large number of related proprietary and/or public data sources. OPM [38], ACEDB [39], and the Entrez cross-database search [40] are examples of database federations.

2.1.3. Database federations with mediated schemas

A problem with database federations can arise in dealing with the various source schemas of different databases in a federation. A mediated schema addresses this problem [41]. Databases in the federation can not only be relational but can be semi-structured data sources [42]. In general, a mediated schema is a graph representing all entities and relationships in a domain of discourse with entities as nodes and relationships as edges [26]. Mediated schemas act as middleware in a database federation where data sources are mapped to the mediated schema by defining the entities they contain [43,44]. Queries are then posed to the mediated schema rather than the union of all of the source database schemas. This allows the user to pose much more general questions that cannot be answered using a traditional relational database [25]. They can also offer the advantage in that mediated schemas can be more focused and tailored relatively easily for a particular user or set of queries. Given a collection of data sources to integrate, a user can develop a mediated schema focused on the data of interest, allowing a rich model of a particular subset of the data they are interested in without having to develop a global schema that must take into account all possible queries or data of interest to all potential users. Mediated schemas can also be "modular," in that a number of then can be created and then swapped in and out as needed.

Mediated schemas may best be suited to situations where researchers need to ask highly complex questions that span disparate knowledge domains. The BioMediator data integration systems is a federated database that uses a mediated schema to ask diverse questions of data sources [45].

2.1.4. Peer data management systems

One of the main drawbacks of database federations with mediated schemas is the challenges in developing a single mediated schema that encompasses an entire domain of discourse. This limitation is somewhat analogous with the difficulties in creating a global database schema in data warehouses and federated databases. A mediated schema is easily developed for a small set of data sources but runs into scaling and maintenance issues as the number of sources increases. A possible way around this limitation is to develop multiple tailored and focused mediated schemas and integrate the mediated schemas in what is known as a peer data management system or PDMS.

A PDMS is an evolutionary step in data integration systems [46]. In a PDMS each data source provides a semantic mapping to either one or a small set of other data sources, or peers. This creates a semantic network of peers which the PDMS can traverse to answer a query.

Peer data management systems can be seen as an exception to our rule of identifying only stable technologies. Currently, there is no implemented peer data management system that we can cite as being a successful data integration project in genomic medicine. We mention peer data management systems here in that they are the next step beyond mediated databases and deserve mention here as an evolutionary step.

A PDMS addresses the problem of creating a global mediated schema by allowing discrete groups to create their own local, specialized mediated schemas and then providing appropriate mappings to glue together semantically related peers on the network. There are many competing projects regarding peer data management systems, a couple of which are the Hyperion Project [47], and the Piazza Project [48].

2.2. Axis 2: data and knowledge representation

The following are data and knowledge representation formalisms that can be employed in any data integration architecture. There are various advantages and disadvantages associated with each. Fig. 1 provides a summary of modeling languages in data and knowledge representation.

2.2.1. Relational schemas

The traditional relational data model [49,50] is centered around the concept of a table (or relation) which consists of rows (tuples) and columns (attributes). The relational model is a well-understood and robust method of representing data but one of its main criticisms is that modeling complex, hierarchically structured objects such as biological entities is not immediately intuitive to anyone except experienced database designers [51]. Also, the relational data model forces precise, unambiguous elucidation of relationships. The data must be regular and complete, or "structured." Unfortunately, our understanding of relationships in biological systems are rarely precise. Despite these drawbacks, the relational schema is by far the most common, familiar, and ubiquitous data representation model [52].

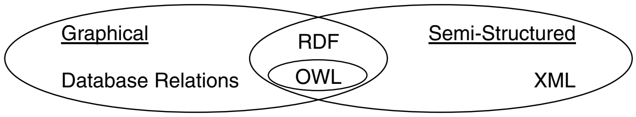

Fig. 1. A summary of modeling languages in data and knowledge representation.

2.2.2. Semi-structured data

Semi-structured data frees you from the rigid structure of the traditional relational data model. Semi-structured data is essentially data with a series of labels and associated values. It can be represented graphically as nodes (objects) connected by edges to values. XML [53] is a format recommended by the WWW Consortium [54] for data exchange on the web and is perfectly suited for describing semi-structured data [55]. The semi-structured data model is not as well understood as the structured model such as in the areas of data validation and search, but it permits more natural modeling of biological entities because it allows features like nesting [56]. A key limitation of XML is that it is difficult to model complex relationships; for example, there is no obvious way to represent many-to-many relationships, which are needed to model complex pathways. The PharmGKB project has used XML in its efforts to build a pharmacogenomics knowledge base [57].

Currently, most information on the World Wide Web is published in HTML, which is suitable for humans but less than ideal for computers. The Semantic Web is a vision of the World Wide Web as a globally linked, semi-structured database [58] expressed in RDF [59], a semi-structured data model in which arbitrarily complex relationships can easily be modeled. The rationale is that data published on the web are useful in some contexts and not others. These data are "hidden" from computers in that it is in HTML, a form not optimal for automated processing [60]. The Semantic Web utilizes XML for its syntactic foundation [61], and *ontologies* to give explicit meaning to information [62].

2.2.3. Ontologies

An ontology is defined as a "specification of a conceptualization" [63]. It is a description of concepts and relationships that exist for a particular domain of discourse, such as anatomy [64]. In regards to data integration, a mediated schema is essentially an ontology serving as middleware for a database federation. Within the Semantic Web, these ontologies are expressed using OWL [62], an ontology language built on top of RDF. Fig. 2 provides an example of gene concepts expressed in XML, and OWL.

Ontologies specify object classes, relationships, and functions [22]. In other words, ontologies are an embodiment of knowledge suitable for a computer. This has enormous implications not only in that a computerized reference source has greatly increased expressive power in regards to queries [25,36,45,65], but also in that a data or reference source can also be a source of *inference* (reasoning). Inferring across data sources is accomplished through computations that traverse a network of entities and relationships within an ontology. For example, Karp and Collado studied the *Escherichia coli* genetic network represented within the pathway database EcoCyc [66], and were able to elucidate a number of interesting

```
<?xml version="1.0"?>
<GeneList>
  <Gene symbol="CREB-17A" organism="D. melanogaster">
    <Sequence>ACTGCGACCGCCCTGCGCT</Sequence>
    <Sequence>ACTGCGACGGCCCAGCGCT</Sequence>
    <Product>cAMP-regulated enhancer-binding protein</Product>
    <Function id="0007616" status="confirmed"><Term>long-term memory</Term></Function>
  </Gene>
  <Gene symbol="CER" organism="D. melanogaster">
    <Sequence>TTCCGTCGCTAGCT</Sequence>
    <Function id="0007616" status="inferred"><Term>long-term memory</Term></Function>
  </Gene>
</GeneList>
```

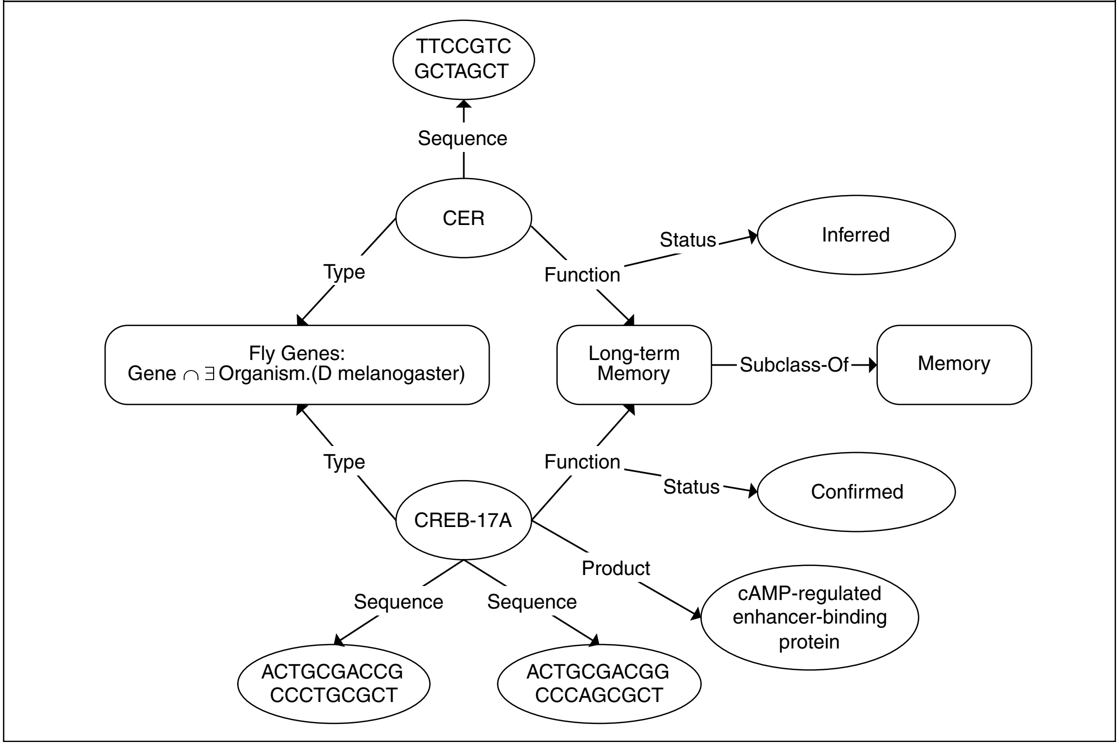

Fig. 2. Top: sample XML describing genes involved in long-term memory. Nested within the gene elements, are sub-elements related to the parent. The first gene includes two nucleic acid sequences, a protein product, and a functional annotation. Additional information is provided by attributes, such as the organism. This example illustrates the difficulty of modeling many-to-many relationships, such as the relationship between genes and functions. Information about functions must be repeated under each gene with that function. If we invert the nesting (i.e., nesting geners inside function elements), then we must repeat information about genes with more than a single function. Bottom: information about genes using RDF and OWL. Both genes are instances of the class *Fly Gene*, which has been defined as the set of all *Genes* for the organism *D. melanogaster*. The functional information is represented using a hierarchical taxonomy, in which *Long-Term Memory* is a subclass of *Memory*.

properties about the genetic network of *E. coli*, such as negative autoregulation being the dominant form of feedback for transcription [65]. This sort of reasoning enables researchers to discover global properties of the ontology that would be extremely difficult with unaided human cognition.

Ontologies also have the added benefit of facilitating interaction between researchers in different knowledge domains and enabling interoperability between databases and programs, both of which are vital to future collaborative work in genomic medicine [67]. The Gene Ontology Consortium [68] is attempting to produce a controlled gene product vocabulary applicable to all organisms and has many, though not all of the attributes of a formal ontology [69] (see Fig. 3).

3. Review of genomic medicine with relevance to data integration

The challenges researchers in genomic medicine face in integrating voluminous amount of heterogeneous data are readily apparent in the literature [70]. Indeed, it has been said that researchers are “trying to swim in a sea of data” [71]. As previously discussed, genomic medicine is complex. We illustrate some of the more “well-known” categories of genomic medicine and identify informatics problems within each.

3.1. Modern human genetics

Genetics, in regards to medicine, studies diseases caused by a single gene. Elucidation of these single-gene diseases

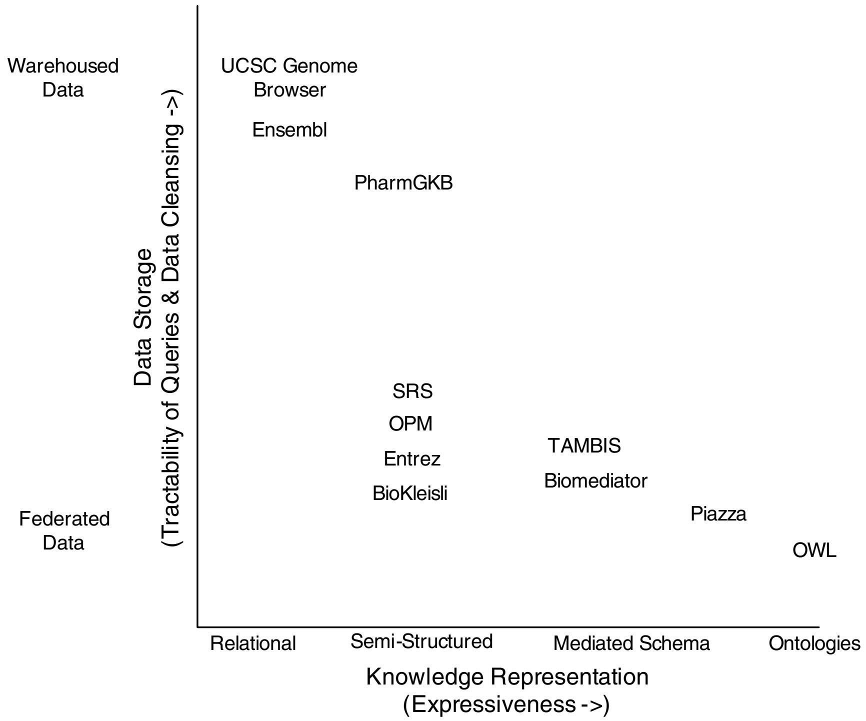

Fig. 3. The axes of data integration architecture and knowledge representation methodologies and where current data integration systems lie along this continuum. A system designed to have control of data and fast queries can have difficulty expressing complex biological concepts and integrating them. Systems that employ highly expressive knowledge representation methodologies are more able to represent and integrate complex biological concepts but have much less tractable queries.

can involve studying their inheritance through large families [10].

Researchers in genetics often have to traverse several data sources to answer questions. This can present problems such as having to know the query syntax and source contents of a great number of data sources. It can even be difficult to navigate using a single gene name due to the dynamic nature of the data sources and lack of standards [72]. The queries posed by researchers can be vague yet highly complex, essentially requiring a “join” of multiple databases to answer it [52]. Geneticists also require records that combine clinical and genetic information to assess the relevance of molecular markers [73]. Their work can also be made easier if tissue samples are linked to accurate medical records and genealogic information [74].

3.2. Genomics in clinical medicine and epidemiology

Clinical medicine aims to evolve into the era of molecular medicine. Diagnosis and treatment of disease will be based on knowledge of the underlying molecular defects rather than evaluation of the overt symptoms [2]. Epidemiology aims to use molecular information to understand transmission and virulence patterns of microbes or to utilize “predictive” genes to assess susceptibility of a population to a disease [9].

Clinicians require access to complete patient information as well as medical knowledge for “just-in-time” information at the point of care [15]. This includes the use of relevant patient genetic data that need to be put in context, usually through connection with public databases. The clinical patient record of the future will require integration of clinical and genomic data along with sound methodologies for acquiring, storing, and analyzing them [9,75].

3.3. Microarray studies

Microarrays are a relatively new technology and are used to measure the transcription levels of thousands of genes from a tissue sample in a single experiment [76]. The conventional wisdom is that by comparing expression levels of various genes from normal and diseased tissue it will be possible to produce a definition of a disease state [77].

Genes are represented on a microarray as “spots,” or uniquely identifying subsequences of the genes. In order to make sense of the experiment, external annotation is needed for each gene [13]. The external annotation often comes from public domain databases which are located in various places and is constantly changing and being updated [51]. There is also the notion of needing to construct valid paths between the datasets (nucleotide → gene → protein) in order to capture proper semantics [78]. Microarray data, being essentially expression information regarding genes, must be analyzed in the context of clinical and epidemiological data to make sense of the experiments. To facilitate this, researchers could benefit from having access to this integrated information.

3.4. Pharmacogenetics and pharmacogenomics

Patients with different genetic makeups can have different responses to drug treatment [79]. Pharmacogenetics and pharmacogenomics attempt to understand how individual genetics plays a role in variation to drug treatment or how systems of genes are involved in modulating drug response [11,80]. The discovery path can be based on a genotype-to-phenotype approach, summarized as: (1) identify suspected gene or system of genes, (2) identify variations within gene(s), (3) search for phenotypes associated with variation, and (4) confirm clinical relevance. A second approach, phenotype-to-genotype, includes the following steps: (1) Identify a phenotype that shows variation, (2) Search for genes that may explain this variation, (3) characterize genetic variation and check for association with the phenotype, and (4) confirm genetic basis for variation and its clinical relevance [81].

Research in pharmacogenetics and pharmacogenomics involves integration of highly diverse types of data including genetic, genomic, phenotypic, and clinical. The type of data schema to model these diverse data types is inherently complex and must change often to incorporate ever-increasing knowledge in the field [57]. An essential requirement for research in pharmacogenetics and pharmacogenomics is a curated knowledge base derived from high-quality, diverse data sets.

3.5. Rational drug design

Drug discovery requires an enormous investment of resources [82]. There is a great need to make informed decisions about proceeding with the costly development of a drug [13]. To increase productivity, the pharmaceutical industry would like to streamline the process by doing much of the work computationally [14]. Also, much recent work in drug discovery revolves around the segregation of populations according to disease subtypes in order to prevent failure of a drug due to supposed inefficacy [12].

Rational drug design involves integration of diverse, heterogeneous data types [13] which can include highly proprietary data. Researchers often ask vague questions that span multiple data sources [23]. The ultimate goal of rational drug design is molecular modeling of disease, prediction of specific compounds which interact with identified proteins, identification of proper patient population through disease sub-classification, and predicting absorption, distribution, metabolism, and excretion (ADME) of a drug "in silico" [14].

Microarrays are highly applicable to research in pharmacogenetics, pharmacogenomics and rational drug design [12], therefore the importance of relevant clinical and epidemiological data applies here as well.

3.6. Biobanks

Biobanks or populational repositories [83], like the centralized anonymous healthcare database in Iceland, integrate coded medical data resources that can be analyzed together with coded genealogy and genotypic information. Human genetic research databases store collections of information on large number of tissues and samples and manage large amounts of molecular epidemiological data of different populations (both of patient and control individuals). The integration of these genetic, clinical, environmental and lifestyle data will facilitate the unravelling of polygenetic disease causality and complex gene-environment interactions existent in disease pathogenesis and causation. The P3G Consortium (Public Population Project in Genomics) [84] includes among its objectives, the development of a common, open and accessible dataset and the building of a unique knowledge base for international collaboration and sharing of data.

4. Application of data integration concepts and approaches to genomic medicine

Theoretically speaking, a single omnipotent database containing the sum total of all biomedical knowledge would solve the data integration problems outlined above. As it turns out, there are strong arguments against this solution [52]. Research in genomic medicine is inherently complex and data intensive. The data is highly heterogeneous and, given political barriers it is unlikely that a single informatics solution will arise that solves all problems [22]. Taking this into account, there exist categories of general informatics problems in genomic medicine, some of which were introduced above. Approaches and concepts in the field of data integration can potentially be applied to provide partial relief from these problems. It is important to note that an improper data integration solution can create more problems that it solves, at least in wasted time and effort. You would not want to buy a semi-truck to commute a long ways to work and then have to deal with a constrained parking situation and likewise you do not want to buy a hybrid vehicle to haul loads of freight. Each data integration solution has optimal cost to benefit ratio so it is important to align these to their proper informatics problems. Table 2 summarizes the examples described in the text.

4.1. Data warehouse approaches

Data warehouses may be best suited for the creation of databases where performance, local control, and privacy are key issues such as in a clinical genetics database [85], in a biobank, or in highly curated reference data sources such as PharmGKB [86], or a clinical trials database for drug discovery [87]. In these cases, there is usually a high amount of human "interaction" with the data, often in

Table 2
Examples of application areas of informatics in genomic medicine and proposed most appropriate data management techniques

		Human genetics	Genomics in Clinical practice	Microarrays	PharmacoGx	Rational drug design	Biobanks
Data storage	Data warehouses	☑			☑		☑
	Database federation			☑			
	Mediated schema		☑		☑	☑	
	Peer data management	☑	☑				
Data and knowledge representation	Ontologies				☑	☑	
	Semi-structured data				☑		

A two-dimensional representation is shown in Fig. 4.

the form of quality control and curation. This provides a rate-limiting step so rapid change in data content or schema is alleviated, thus reducing one of the major problems most often identified with data warehouses.

4.2. Database federation approaches

Researchers who need to dynamically integrate vast amounts of related data located in disparate locations where the data sources may possibly be undergoing rapid change in data model or data content would be best served by a database federation rather than data warehouses. Microarray researchers may require annotation for each of the possibly thousands of spots on the array and the data they require may reside in large databanks located across a network [51]. The volume of data may simply be too great to house locally, due to limitations in data warehouse technology or resources of the researchers. Also, annotation data is notorious for rapid change. Researchers who have identified these problems such as in the case of gene expression studies [78], or who have formed ad hoc research collaborations [88], may be best served by database federations since no data, other than proprietary, need be housed locally and since data is kept at the source, it is always up-to-date.

4.3. Mediated schema approaches

Researchers may only have a general notion of the questions they want to ask [45]. The genotype-to-phenotype correlation in genomic medicine requires answering vague questions that span multiple disciplines. Mediated schemas, serving as middleware to database federations, may help in this regard as they facilitate "general" queries and integration of diverse types of data. Mediated schemas are good productivity tools in that researchers only have to understand one schema that applies to the entire federation, rather than having to understand the individual schemas of the sources. They also have the advantage of being "modular," in regards to the fact that it is unlikely that one schema will serve the needs of many groups of researchers. Any number of specific mediated schemas can be created to meet the needs of a particular group and swapped in an out of a federation, quite unlike the fixed schema of a data warehouse. Researchers in rational drug design and pharmacogenetics and pharmacogenomics often have to ask general queries that span knowledge domains [23] and may benefit from the use of mediated schemas.

4.4. Peer data management system approaches

Peer data management systems are a future possibility for data integration in genomic medicine. In this model, data sources remain autonomous but can connect to an existing, interwoven, information fabric by interacting through a relevant domain ontology and providing appropriate concept mappings. The world of disparate data sources essentially becomes a self-organizing, semantic network. This model could be suitable for those systems oriented toward facilitating the navigation between genotype and phenotype, as in human genetics or clinical genomics. The Piazza peer data management system represents important research in this regard [89]. This model is not without precedent as a similar "knuckles-and-nodes" approach to the problem of data integration has been proposed in the literature [52].

4.5. Ontologies

Ontologies aid in the integration of diverse data that spans disciplines. In other words, they can help to resolve the semantic inconsistencies that arise when attempting to integrate data from different sources. For example, the definition of "vector" is different in molecular biology and in mathematics [51]. Researchers who attempt to integrate data from disparate knowledge domains, such as in rational drug design, pharmacogenetics and pharmacogenomics, or within a knowledge domain may benefit by developing ontologies. The genomics working group of the American Medical Informatics Association [90] recently had a meeting at the MedInfo 2004 conference [91] that focused on ontologies in genomic medicine including representation of molecular knowledge in computable form as in Flybase [92], representation of phenotype in the Foundational Model of Anatomy [93], and cross-domain integration of biomedical information [94].

4.6. Semi-structured data

XML is a format that is better for representing complex, hierarchical biological objects for human visualization. Also, much research is ongoing in regards to being able to query XML documents [95], essentially giving an XML document the query capabilities usually associated with databases. An ontology is needed to control nomenclature and to provide semantic links between XML documents. Integration of data with complex data models could be best served using XML, such as in PharmGKB [57], and in conjunction with mediated schemas and database federations as in BioMediator [45].

There are a large number of ‘‘ad hoc’’ data integration solutions that exist today in genomic medicine [27,96–98]. These solutions are highly specialized for a specific purpose, such as genetics [99]. They are very useful in their own right but likely would not be ‘‘portable,’’ or applicable to researchers in different domains, such as medicinal chemistry and biology. The field of data integration attempts to provide general solutions and we have attempted to identify those general data integration solutions that apply to informatics problems in genomic medicine (Fig. 4).

5. Gaps remaining in data integration research to facilitate genomic medicine

There should be no doubt that numerous gaps remain regarding data integration to facilitate genomic medicine and it is certain that the number of gaps will increase as more data becomes available. The gaps reviewed here reflect some of the more urgent needs and challenges identified both in the literature and in our overview of genomic medicine above.

5.1. Data availability

The number of public databases in molecular biology is large and growing [17]. Access to this data has facilitated rapid advances in the field of bioinformatics [100]. Similarly, researchers in genomic medicine require access to large clinical data sources for integration with molecular biology databases to make inferences on the genotype-to-phenotype connection. Some of the more well-known clinical databases include OMIM [101], GeneClinics [102], and Medline [103], but the number of publicly available clinical databases is small in comparison to molecular biology. It appears that the reason for this revolves around issues regarding privacy and data modeling and collection [104].

5.1.1. Privacy issues

The issue of privacy is not as simple as it may seem. It is not that easy to ‘‘de-identify’’ and individual in that an individuals genome is essentially a fingerprint and thus uniquely identifying. In addition to individual privacy there is also the open issue of the privacy of the family of an individual with a ‘‘defective’’ genome. For example, if a parent is diagnosed with Huntington’s disease, a fatal autosomal dominant disorder, then there is a minimum 25% chance that any of their children will also have the disease. Even worse, if a child tests positive for Huntington’s

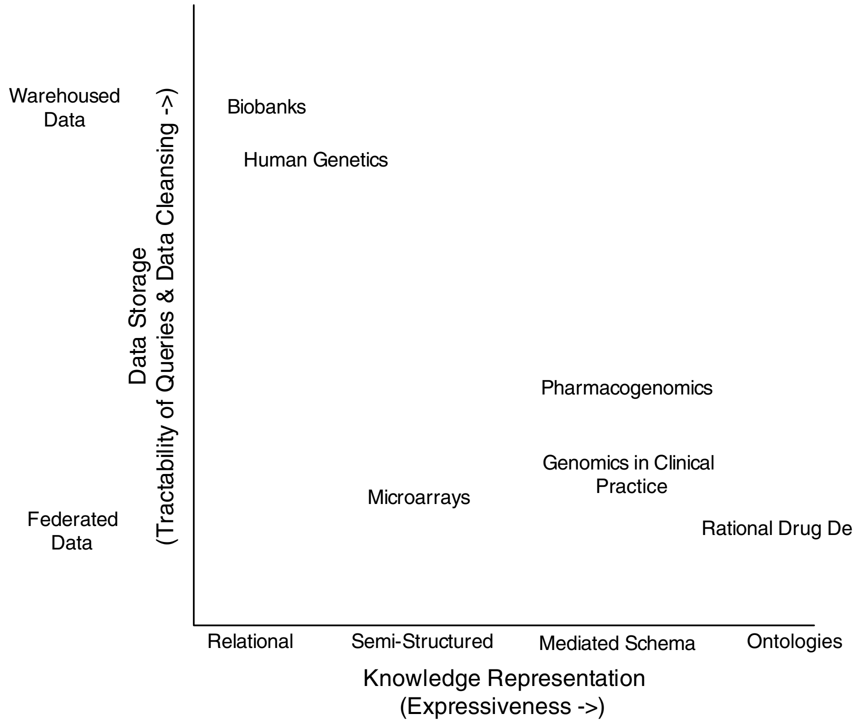

Fig. 4. The axes of data integration architecture and knowledge representation methodologies and where approaches in genomic medicine lie along this continuum according to their identified challenges regarding data integration and knowledge representation.

disease then there is a 100% chance that at least one parent will have the disease.

The issue of privacy versus access to data is a battleground in genomic medicine as well as healthcare. Data sharing is challenging in areas such as pharmacogenomics [105] in that they are inherently about linking genotype to clinical phenotype, such as individual response to drug treatment. In the healthcare arena, individual medical records are increasingly moving from paper to electronic format. The current solutions such as informed consent, de-identification, and mediation, are apparently inadequate [81,104]. For example, medical data released by medical institutions may appear to be anonymous but can often be re-identified [106]. The Health Insurance Portability and Accountability Act (HIPAA) was recently enacted to address medical records in the electronic era. Debate over HIPAA illuminates the competing interests of privacy advocates who argue for individual autonomy over medical records and healthcare industry advocates who argue for freedom to use information to aid treatment decisions and to further research [107]. The interdisciplinary nature of genomic medicine requires access and sharing of data to reap its potential benefits [20].

5.1.2. Data issues

The complexity of the clinical record has been identified as a major barrier to the collection of clinical data [104], and significant gaps have been identified between molecular biological data and its relevance to the clinic [73]. Update issues as a barrier to the usefulness of public databases [72] have been identified as well. Also, there is much useful information in scientific papers but is in the form of natural language text. Research in natural language processing and text mining is needed to populate databases which could be an enormously valuable resource [13,81].

5.2. Lack of standards

It has been identified that the lack of standardized vocabularies has hampered development of databases in the clinical arena [104]. Clinical data, or phenotype, is difficult to precisely define and represent [20,81], although some relevant initiatives have been recently proposed, such as the Human Phenome Project [108], Phenofocus initiative [109], and the IEEE: Bioinformatics Standards Committee [110].

Genomic medicine will require integration of diverse complex data types including genomic, proteomic, clinical, and even pharmacological and chemical [23]. Standards will be required for representation of clinical and genetic information to ensure proper semantic integration of heterogeneous data, and also for communication standards to ensure interoperability between disparate data sources [20]. HL7 [111] and SNOMED [112] are examples of existing foundational standards that can be used as tools for the development of future standards.

5.3. Bridging disciplines: collaborations or convergence?

For molecular biological information to be useful to genomic medicine it must be analyzed in context with clinical information [73]. Genetic and genomic information will have to be managed over many levels of health information from the molecular to the population level [80]. Researchers from all backgrounds will have to be able to understand the interrelations of all disciplines involved in genomic medicine [20]. Neuroinformatics could be considered a "bridging" discipline in that it involves managing information at many different levels of neuroscience from the micro to the macro-anatomic, and essentially "connects" genomic and clinical information in the area of neuroscience research [79]. Similar approaches are being pursued through what we could name "integrated approach to the study of diseases," including examples such as cancer informatics or cardiovascular informatics. All of this suggests close collaboration is needed between medical informatics and bioinformatics, as was illustrated in the European project BIOINFOMED, and it is even suggested that they should converge into a single discipline [20]. Biomedical Informatics is the emerging discipline that aims to put these two worlds together so that the discovery and creation of novel diagnostic and therapeutic methods is fostered. The INFOBIOMED [113] network of Excellence, recently funded by the European Commission, aims to set a durable structure for the described collaborative approach at a European level, supporting the consolidation of BMI as a crucial scientific discipline for future healthcare. There are already many synergies between medical informatics and bioinformatics and a convergence of the two fields could remove a major impediment to research in this arena, namely the differing motivations of collaborators [79], as well as sharing approaches to the facilitation of genomic medicine [75].

6. Conclusion

Research in genomic medicine and data integration has proceeded in relative isolation, although there have been some attempts at cross-pollination [23,36,114]. Our review of the literature in genomic medicine and data integration has illuminated the possibility of greater synergies between the two in the identification of generalized informatics problems in genomic medicine and applicable concepts and approaches in data integration.

Certainly there are many gaps and challenges that remain for data integration research to facilitate genomic medicine. The creation of a National Health Information Infrastructure [115].will address many of these gaps and challenges and will be needed in order for genomic medicine to be effectively applied in healthcare. Genomic medicine is ambitious and will require enormous scientific and political will to implement, but the implications and benefits are to great to ignore.

Acknowledgments

Dr. Martin-Sanchez's research in related areas has been supported by funding from the EC INFOBIOMED Network of Excellence, the INBIOMED project, Ministry of Health, Spain, and the MEDGENBASE project, Ministry of Science and Technology, Spain. Dr. Halevy's and Peter Mork's research in related areas has been supported by funding from: NSF IIS-0205635—ITR, NSF IIS-9985114—PECASE, SRI Subcontract 03-000225 (in support of DOI NBCHD030010)—CALO (DARPA via SRI), and NIH 1P20LM007714-01—BISTI Planning Grant. Dr. Tarczy-Hornoch, Brent Louie, and Peter Mork's research in related areas has been supported by funding from: NHGRI (R01HG02288, PI: Tarczy-Hornoch), NLM training grant (T15LM07442, trainees: Mork & Louie) and BISTI planning grant (P20LM007714, PI: Brinkley, co-PI: Tarczy-Hornoch).

References

[1] Collins FS, Guttmacher AE. Genomic medicine—a primer. N Engl J Med 2002;347:1512–20.

[2] Gerling IC, Solomon SS, Bryer-Ash M. Genomes, transcriptomes, proteomes. Arch Intern Med 2003;163:190–8.

[3] Collins FS, McKusick VA. Implications of the human genome project for medical science. JAMA 2001;285(5):540–4.

[4] Sander C. Genomic medicine and the future of health care. Science 2000;287(5460):1977–8.

[5] Collins FS. Medical and societal consequences of the human genome project. N Engl J Med 1999;341:28–37.

[6] Ansell S et al. Primer on medical genomics Part VI: genomics and molecular genetics in clinical practice. Mayo Clin Proc 2003;78:307–17.

[7] Hopkins A, Groom C. The druggable genome. Nat Rev Drug Discov 2002;1(9):727–30.

[8] Russell S, Peng K-W. Primer on medical genomics Part X: gene therapy. Mayo Clinc Proc 2003;78:1370–83.

[9] Maojo V, Martin-Sanchez F. Bioinformatics: towards new directions for public health. Methods Inf Med 2004;43(3):208–14.

[10] Saar K et al. Homozygosity mapping in families with Joubert syndrome identifies a locus on chromosome 9q34.3 and evidence for genetic heterogeneity. Am J Hum Genet 1999;65:1666–71.

[11] Klein TE et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenomics J 2001;1:167–70.

[12] Schadt EE, Monks SA, Friend SH. A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. Biochem Soc Trans 2003:437–43.

[13] Claus BL, Underwood DJ. Discovery informatics: its evolving role in drug discovery. Drug Discov Today 2002;7(18):957–66.

[14] Augen J. The evolving role of information technology in the drug discovery process. Drug Discov Today 2002;7(5):315–23.

[15] Tarczy-Hornoch P, et al. Meeting clinician information needs by integrating access to the medical record and knowledge resources via the web. In: Proceedings of the AMIA Annual Fall Symposium. 1997.

[16] Consortium IHGS. The human genome. Nature 2001;409:860–921.

[17] Galperin MY. The molecular biology database collection: 2004 update. Nucleic Acids Res 2004;32:D3–D22.

[18] Ideker T et al. Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. Science 2001;292: 929–34.

[19] NCI, http://plan.cancer.gov/glossary.html.

[20] Martin-Sanchez F et al. Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. J Biomed Inform 2004;37(1):30–42.

[21] Karasawas K, Baldock R, Burger A. Bioinformatics integration and agent technology. J Biomed Inform 2004;37(3):205–19.

[22] Sujansky W. Heterogeneous database integration in biomedicine. J Biomed Inform 2001;34(4):285–98.

[23] Haas LM et al. DiscoveryLink: a systems for integrated access to life sciences data sources. IBM Syst J 2001;40(2).

[24] GeneticXchange, http://www.geneticxchange.com/v3/index.php.

[25] Stevens R et al. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics 2000;16(2):184–5.

[26] Mork P, Halevy A, Tarczy-Hornoch P. A model for data integration systems of biomedical data applied to online genetic databases. AMIA Symp 2001:473–7.

[27] Karolchik D et al. The UCSC genome browser database. Nucleic Acids Res 2003;31(1):51–4.

[28] Hubbard T et al. The Ensembl genome database project. Nucleic Acids Res 2002;30(1):38–41.

[29] Critchlow T, Fidelis K. DataFoundry: information management for scientific data. IEEE Trans Inf Technol Biomed 2000;4(1):52–7.

[30] Bukhman Y, Skolnick J. BioMolQuest: integrated database-based retrieval of protein structural and functional information. Bioinformatics 2001;17(5):468–78.

[31] Chaudhuri S, Dayal U. An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record 1997.

[32] Levy A, Rajaraman A, Ordille J. Querying heterogeneous information sources using source descriptions. In: Proceedings of the 22th international conference on very large data bases. 1996. p. 251–262.

[33] Lambrecht E, Kambhampati S, Gnanaprakasam S. Optimizing recursing information-gathering plans. In: Proceedings of the sixteenth international joint conference on artifical intelligence. 1999. p. 1204–1211.

[34] Wiederhold G. Mediators in the architecture of future information systems. IEEE Comput Mag 1992;25(3):38–49.

[35] Haas LM. Optimizing queries across diverse data sources. In: Proceedings of the 23rd international conference on very large data bases. 1997. p. 276–285.

[36] Chung S, Wong L. Kleisli: a new tool for data integration in biology. Trends Biotechnol 1999;17(9):351–5.

[37] Muller H, Freytag J. Problems, methods and challenges in comprehensive data cleansing. Berlin: Humboldt-Universitt zu Berlin, Institut fr Informatik; 2003.

[38] Chen I, Markowitz V. An overview of the object-protocol model (OPM) and OPM data management tools. Informat Syst 1995;20(5).

[39] Stein L, Thierry-Mieg J. Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. Genome Res. 1998;8(12):1308–15.

[40] Entrez, http://www.ncbi.nlm.nih.gov/Entrez/.

[41] Chawathe S. The TSIMMIS Project: integration of heterogeneous information sources. In: Proceedings of IPSJ conference. 1994. p. 7–18.

[42] Ambite J et al. Ariadne: a system for constructing mediators for Internet sources. ACM SIGMOD Record 1998;27(2):561–3.

[43] Shaker R, et al. A rule driven bi-directional translation system for remapping queries and result sets between a mediated schema and heterogeneous data sources. In: AMIA Symposium. 2002.

[44] Adali S et al. Query caching and optimization in distributed mediator systems. ACM SIGMOD Record 1996;25(2):137–46.

[45] Donelson L, Tarczy-Hornoch P, Mork P. The BioMediator system as a data integration tool to answer diverse biologic queries. In: Proceedings of MedInfo, IMIA, San Francisco, CA; 2004.

[46] Gribble S, et al. What can databases do for peer-to-peer? WebDB Workshop on Databases and the Web, June 2001.

[47] Arenas M et al. The hyperion project: from data integration to data coordination. SIGMOD Record 2003;32(3):53–8.

[48] Tatarinov I et al. The Piazza peer data management project. SIGMOD Record 2003;32(3).
[49] Codd EF. A relational model for large shared data banks. Comm ACM 1970;13:377–87.
[50] Chen PPS. The entity-relationship model: toward a unified view of data. ACM Trans Database Syst 1976;1:9–36.
[51] Lacroix Z, Critchlow T. Bioinformatics: managing scientific data. San Francisco: Morgan Kaufmann Publishers; 2003.
[52] Stein LD. Integrating biological databases. Nat Rev Genet 2003;4(5):337–45.
[53] Eckstein R. XML pocket reference. Sebastopol: O'Reilly; 1999.
[54] W3C, W.W.W.C., Extensible markup language (XML) 1.0 (second edition). 2000, W3C.
[55] Abiteboul S, Buneman P, Suciu D. Data on the Web. San Francisco: Morgan Kaufmann Publishers; 2000.
[56] Achard F, Vaysseix G. XML, bioinformatics and data integration. Bioinformatics 2001;17:115–25.
[57] Rubin DL, et al. Automating data acquisition into ontologies from pharmacogenetics relational data sources using declarative object definitions and XML. In Pac. Symp Biocomput. 2002.
[58] Berners-Lee T, Hendler J, Lassila O. The semantic web. Scientific American; 2001.
[59] Miller E. An introduction to the resource description framework, In: Dlib Magazine. 1998.
[60] Palmer S. The semantic web: an introduction. 2001.
[61] Dumbill E. The semantic web: a primer. 2000.
[62] W3C, W.W.W.C., OWL web ontology language use cases and requirements. 2004.
[63] Gruber T. A translation approach to portable ontologies. Knowl Acquis 1993;5(2):199–220.
[64] Rosse C, Mejino J. A reference ontology for bioinformatics: the foundational model of anatomy. J Biomed Informat 2003;36:478–500.
[65] Karp P. Pathway databases: a case study in computational symbolic theories. Science 2001;293(14):2040–4.
[66] EcoCyc, www.ecocyc.org.
[67] Stead W et al. Integration and beyond: linking information from disparate sources and into workflow. J Am Med Inform Assoc 2000;7:135–45.
[68] Consortium TGO. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[69] Smith B, Williams J, Schulze-Kremer S. The Ontology of the Gene Ontology. In: Proceedings of AMIA Symposium 2003, 2003.
[70] Murray A. Whither genomics? Genome Biol 2000;1(1). comment003.1–comment003.6.
[71] Roos D. Bioinformatics–Trying to swim in a sea of data. National BioInformatics Institute; 2001.
[72] Mitchell J, McCray A, Bodenreider O. From phenotype to genotype: issues in navigating the available information resources. Methods Inf Med 2003;42(5):557–63.
[73] Dugas M et al. Impact of integrating clinical and genetic information. In Silico Biol 2002;2(0034).
[74] Annas G. Rules for research on human genetic variation–lessons learned from Iceland. N Engl J Med 2000;342(24):1830–3.
[75] Maojo V, Kulikowski CA. Bioinformatics and medical informatics: collaborations on the road to genomic medicine. J Am Med Inform Assoc 2003(10):515–22.
[76] Brown P, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet 1999;21(1 Suppl.):33–7.
[77] Diehn M, Alizadeh A, Brown P. Examining the living genome in health and disease with DNA microarrays. JAMA 2000;283(17):2298–9.
[78] Mei H et al. Expression Array annotation using the biomediator biological data integration systems and the bioconductor analytic platform. In: Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium. Washington, DC: American Medical Informatics Association; 2003.
[79] Miller PL. Opportunities at the intersection of bioinformatics and health informatics. J Am Med Inform Assoc 2000(7):431–8.
[80] Martin-Sanchez F, Maojo V, Lopez-Campos G. Integrating genomics into health information systems. Methods Inf Med 2002;41(1):25–30.
[81] Altman RB, Klein TE. Challenges for biomedical informatics and pharmacogenomics. Annu Rev Pharmacol Toxicol 2002;42:113–33.
[82] Tufts, http://csdd.tufts.edu/NewsEvents/RecentNews.asp?newsid=6. 2001.
[83] Kaiser J. Biobanks. Population databases boom, from Iceland to the U.S. Science. 2002;298:1158-61.
[84] P3G, www.p3gconsortium.org.
[85] Birch P, Friedman JM. Utility and limitations of genetic disease databases in clinical genetics research: A neurofibromatosis 1 database example. Am J Med Genet 2004;125C(1):42–9.
[86] PharmGKB, http://www.pharmgkb.org.
[87] TrialDb, http://ycmi.med.yale.edu/trialdb/.
[88] Kolker E et al. Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae. Nucleic Acids Res 2004;32(8):2353–61.
[89] Tatarinov I et al. The Piazza peer data management project. SIGMOD 2003;32(3).
[90] GEN-WG, A., http://www.amia.org/working/genomics/main.html.
[91] Medinfo2004, http://www.medinfo2004.org/.
[92] Consortium TF. The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic Acids Res 2002;30:106–8.
[93] Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. Proceedings, Medinfo 2004, 2004.
[94] Covitz P et al. caCORE: a common infrastructure for cancer informatics. Bioinformatics 2003;19(18):2404–12.
[95] Mork P, et al. PQL: a declarative query language over dynamic biological schemata. AMIA Symp 2002.
[96] GeneCards, http://bioinfo.weizmann.ac.il/cards/index.shtml.
[97] Ensembl, http://www.ensembl.org/.
[98] UCSC, http://genome.ucsc.edu/.
[99] Cyrillic, http://www.cyrillicsoftware.com/.
[100] Chicurel M. Bioinformatics: bringing it all together. Nature 2002;419:751–7.
[101] OMIM, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.
[102] GeneClinics, http://www.geneclinics.org/.
[103] MedLine, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.
[104] Kohane IS. Bioinformatics and clinical informatics: the imperative to collaborate. J Am Med Inform Assoc 2000(7):512–6.
[105] Rothstein M, Epps P. Ethical and legal implications of pharmacogenomics. Nat Rev Genet 2001;2:228–31.
[106] Sweeney L. Guaranteeing anonymity when sharing medical data,k the datafly system. In: Proc J Am Med Inform Assoc. Washington, DC: Hanley & Belfus Inc; 1997.
[107] Gostin LO. National health information privacy: regulations under the Health Insurance Portability and Accountability Act. JAMA 2001;285(23):3015–21.
[108] Freimer N, Sabatti C. The human phenome project. Nat Genet 2003;34(1):15–21.
[109] Phenofocus, www.phenofocus.net.
[110] Committee, I.B.S., http://grouper.ieee.org/groups/1953/index.html.
[111] HL7, http://www.hl7.org.
[112] SNOMED, http://www.snomed.org.
[113] INFOBIOMED, www.infobiomed.org.
[114] Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. Methods Enzymol 1996;266:114–28.
[115] Yasnoff WA et al. A consensus action agenda for achieving the National Health Information Infrastructure. J Am Med Inform Assoc 2004(11):332–8.