Accessing and Integrating Data and Knowledge for Biomedical Research

A. Burgun[1], **O. Bodenreider**[2]
[1]EA 3888, IFR 140, Faculté de Médecine, Université de Rennes I, 35033 Rennes, France
[2]National Library of Medicine, NIH, Bethesda, Maryland, USA

Summary

Objectives: To review the issues that have arisen with the advent of translational research in terms of integration of data and knowledge, and survey current efforts to address these issues.

Methods: Using examples form the biomedical literature, we identified new trends in biomedical research and their impact on bioinformatics. We analyzed the requirements for effective knowledge repositories and studied issues in the integration of biomedical knowledge.

Results: New diagnostic and therapeutic approaches based on gene expression patterns have brought about new issues in the statistical analysis of data, and new workflows are needed are needed to support translational research. Interoperable data repositories based on standard annotations, infrastructures and services are needed to support the pooling and meta-analysis of data, as well as their comparison to earlier experiments. High-quality, integrated ontologies and knowledge bases serve as a source of prior knowledge used in combination with traditional data mining techniques and contribute to the development of more effective data analysis strategies.

Conclusion: As biomedical research evolves from traditional clinical and biological investigations towards omics sciences and translational research, specific needs have emerged, including integrating data collected in research studies with patient clinical data, linking omics knowledge with medical knowledge, modeling the molecular basis of diseases, and developing tools that support in-depth analysis of research data. As such, translational research illustrates the need to bridge the gap between bioinformatics and medical informatics, and opens new avenues for biomedical informatics research.

Keywords

Medical Informatics; bioinformatics; databases, distributed knowledge bases

Geissbuhler A, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2008. Methods Inf Med 2008; 47 Suppl 1:91-101

Access to information, analysis of data, and integration of knowledge are key components of biomedical research. Scientists and physicians must be able to integrate their data with other data, to combine information from multiple sources, and to compare their results to prior knowledge. This paper illustrates the role of knowledge in biomedical research, with focus on omics disciplines, and surveys current efforts to address the needs of biomedical researchers for better access to information and better integration of data and knowledge.

1 Trends in Biomedical Research and their Impact on Bioinformatics

The current era of biomedical research can be characterized by what NIH Director E.A. Zerhouni calls the "four Ps" of medicine: Predictive, Personalized, Preemptive and Participatory[1]. Risk factors of diseases must be identified early in order to adapt counter-measures, especially for long-term, chronic diseases. Treatments must be tailored in order to take into account the characteristics of individual patients. Shifting the focus of medicine from the current doctor-centric, curative paradigm to preventing diseases will require the active involvement of patients. With the advent of personalized medicine, biomarkers, including genetic markers, will be tested for each patient in order to diagnose specific forms of diseases, predict disease progression and patient outcome, and propose the best therapeutic options. This scenario puts genomics and pharmacogenomics at the centre of medicine [1]. This new vision of personalized medicine is supported by very active biomedical research. As the role of "omics" disciplines[2] in biomedical research becomes more important, classical clinical studies must be adapted to these new approaches. New models of diseases have emerged from these studies. The genes identified through omics studies provide clues to possible pathogenetic mechanisms and are likely to be useful in developing diagnostic tests and adapting therapeutic responses. Discoveries typically begin at "the bench" with basic research. Then they must be translated into practical applications and progress to the clinical level, the patient's "bedside." In parallel, clinical researchers make novel observations about the nature and progression of disease that often stimulate basic investigations. This exchange of information

[1] http://nihroadmap.nih.gov/

[2] omics is a generic term for new disciplines enabled by high-throuput technologies, such as genomics, transcriptomics, and proteomics

describes translational research or translational medicine: researchers and physicians applying newly gained knowledge to the clinic - and back again to the bench[3]. Such recent changes in biomedical research have brought about new challenges for bioinformatics and medical informatics. The analysis of genomic studies and the new workflows between research and health care generate greater demand for accessing and integrating information.

1.1 New Diagnostic and Therapeutic Approaches

Disease classification based on gene expression patterns. Over the past decade, biomedical research has evolved to mine gene expression profiles for clues to the pathogenesis, prognosis and treatment of human diseases. In oncology, for example, this research rests on the premise that extraordinary insights into the molecular basis of cancer can be obtained by analyzing gene expression in patient-derived tumor samples, in addition to experimental models. DNA microarrays (DNA chips) are used to monitor the gene expression (i.e., a proxy for gene activity) of thousands of genes simultaneously across the human genome. This technique involves the extraction of RNA from tumor samples and its subsequent fluorescent labeling and hybridization to an array of DNA probes. Microarrays covering nearly the entire human genome are now available. In a series of experiments, Golub demonstrated that the classification of cancer -- specifically two principal forms of acute leukemia -- could be achieved by using DNA microarrays to monitor gene expression, without a prior molecular understanding of this distinction [2]. This finding implies that such methodologies can be applied to the molecular dissection of cancers. This approach has been used for the molecular classification of many tumor types, including lymphoma (e.g., [3]), prostate cancer (e.g., [4]), brain tumors (e.g. [5]), and lung cancer (e.g., [6]). Similar approaches have demonstrated that patterns of gene expression (or gene expression "signatures") may be found across different tumor types. For example, Golub et al identified a signature of metastatic propensity across prostate, breast, and lung cancers, suggesting that a genetic test performed at the time of diagnosis might predict the future behavior of some tumors [7]. While most studies of gene expression have been carried out on tissue samples, some have used peripheral blood samples (e.g. [8]), thus extending the applicability of this technique.

Pharmacogenomics. Gene expression-based approaches are also widely used in pharmacology (e.g., [9]). The expectation here is that genomic approaches might lead to the discovery of molecules and compounds capable of modulating biological processes in cells. Drug discovery typically starts with prior knowledge of a target gene that is biologically relevant to a disease state (e.g., a gene mutation in cancer). The protein product of this gene is then biochemically purified, and a collection of compounds screened in vitro for their ability to bind to the protein. Novel approaches to drug discovery are based on genomics. Gene expression-based methods are used to identify candidate drugs that modulate previously intractable targets. These genes and gene products can serve as potential therapeutic targets or tools in addition to providing diagnostic and prognostic markers, as well as endpoints for clinical trials. In cancer research, this approach has been applied to the discovery of substances that may induce the maturation of abnormal cells (e.g., acute myeloid leukemia cells), inhibit androgen or estrogen action in cancer cells, inhibit angiogenesis associated with tumor cell proliferation or inhibit the activity of the causal protein in some tumors (e.g., Ewing sarcoma [10]).

The functional consequences of genetic polymorphisms have been examined for several drug-metabolizing enzymes [11]. Variants leading to reduced or increased enzymatic activity compared to the wild-type alleles have been identified. The possible application of genotyping has been discussed for several pathologic conditions. Among many other examples, the acetylator status has long been used for predicting isoniazid-induced hepatic toxicity in tuberculosis [12], and associations between genetic variability and response to beta-adrenergic medications have been explored [13]. The association between gene expression and response to treatment holds the promise of personalized medicine, as doctors will be able to individualize drug therapy and provide specific therapies to those most likely to respond, while avoiding therapies in those most likely to suffer adverse effects.

1.2 New Issues Related to the Analysis of Genomic Studies

Clinical trials provide an evaluation framework for interventions. Parameters are measured in patients under different types of interventions and the values of these parameters are compared across groups of subjects in order to identify associations between interventions and outcomes. Traditional clinical trials generally involve many subjects in which only few parameters are

[3] http://nihroadmap.nih.gov/clinicalresearch/overview-translational.asp

measured. Conversely, omics studies typically generate a large number of measurements on the limited number of test subjects (relatively to the number of parameters measured). This imbalance has created new issues involving statistics and bias [14]. Omics studies offer a potentially powerful approach to identifying new biomarkers, but many of them are plagued by a lack of consistency and reproducibility (e.g., [15]). In principle, the inconsistency may be due to false positive studies, false negative studies or true variability among heterogeneous groups. In order to avoid biases and get more reliable results, the data from individual experiments at different centers could be pooled and public data repositories used for comparative data analysis [16]. Moreover, the goal of omics approaches is also to acquire comprehensive, integrated understanding of biology by studying all biological processes in addition to analyzing parameters individually (e.g., [17]). Therefore, solutions exploiting prior knowledge about gene functions (e.g., in gene annotations databases) and multi-scale biological models have been proposed and are discussed in section 3.3.

1.3 New Workflows in Biomedical Research

In the context of translational research and translational medicine, information sharing between medical research, epidemiology and clinical medicine has been identified as a strong requirement. Translational research creates a bidirectional information transfer that accelerates trials and evaluates their clinical potential. In this framework, clinical data and biomarkers must be collected early in order to extract new knowledge and form new hypotheses from the mass of collected data. Therefore the relationship between research, population studies and health care rests on the integration of the data and knowledge from these three areas: research (scientific publications, public databases, experimental results), epidemiology (e.g., cohort studies), and healthcare (clinical data stored in patient records).

Two main challenges have to be overcome when automatically interrelating data from these different areas. First, these data are annotated to different terminologies and data referring to the same entity may be represented by different identifiers [18]. For instance, the disease "acute myeloid leukemia" is coded D015470 in bibliographic databases indexed with MeSH, 91861009 in clinical records coded with SNOMED Clinical Terms® (SNOMED CT®)[4], and C3171 in research records annotated to the NCI Thesaurus[5] [19]. The second issue is that the data to be integrated are complementary in nature but intrinsically different (omics - pathology - anatomy - physiology). Ontologies have been proven useful for data integration (e.g., [20, 21]). Several ontologies have been developed in bioinformatics and in the biomedical domain. However, they are still incomplete (neither all concepts nor relations are present) and fragmented (ontologies are orthogonal and few bridges are established between complementary ontologies) (e.g., [22]). Enrichment and integration of biomedical ontologies are therefore important stakes for translational medicine and bioinformatics, as well as for the future links between these two disciplines (e.g., [23, 24, 25])

[4] http://www.ihtsdo.org
[5] http://www.nci.nih.gov/cancerinfo/terminologyresources

2 Effective Data Repositories

Pooling experimental data requires the standard annotation of the experiments. It also requires interoperability among data repositories supported by standard services and workflows. Interoperable data repositories constitute an enabling resource for meta-analysis.

2.1 Repositories of Experimental Data

Public datasets have been created in response to the growing demand for publicly available repositories for high-throughput gene expression data. Such public repositories represent an important resource for the biological research community as they provide unrestricted access to microarray data published by other researchers. As such, they complement local in-house gene expression databases by providing reference data for comparative studies. Among them, the Gene Expression Omnibus (GEO) repository developed by the National Center for Biotechnology Information (NCBI) is publicly accessible on the NCBI website at http://www.ncbi.nlm.nih.gov/geo [26]. GEO archives and helps disseminate microarray and other forms of high-throughput data generated by the scientific community [27]. GEO data can be viewed from the perspective of the experiment or the gene. The experiment-centric view presents the entire study, while the gene-centric view displays quantitative gene expression measurements for one given gene across a dataset, with links to gene annotations. Other efforts to archive experiments and make them accessible to the whole community include the Stanford Microarray Database (SMD) [28] (http://smd.stanford.edu) and the ArrayExpress database of microarray [29] (http://

www.ebi.ac.uk/arrayexpress), developed by the European Bioinformatics Institute. All these repositories promote standard exchange formats such as MAGE-TAB [30]. Moreover, data submitted to these repositories are required to have a common set of core elements. As many other resources in this domain, including local experimental databases, data sets in public repositories are compliant with the standards that define a minimum information about a microarray experiment. Broad adhesion to these standards facilitates the publication and retrieval of data, as it ensures consistency across datasets.

In addition to such wide-scale projects, more focused initiatives seek to collect all published data on a given medical topic. Specific pipelines and services have been developed in conjunction with such focused databases. For example, the Oncomine initiative seeks to collect all published cancer microarray data (http://www.oncomine.org). To date, this effort has accumulated 18,000 cancer gene expression experiments. Automated analyses can be performed to identify the genes, pathways, regulatory networks, and functional networks activated and repressed in human cancer. As described in [31], all cancer microarray data deposited in GEO and SMD are automatically copied to Oncomine and then standardized.

Data repositories may be extended with clinical data. With focus on three types of tumors -- breast carcinoma, bladder carcinoma and uveal melanoma -- the Integrated Tumor Transcriptome Array and Clinical data Analysis (ITTACA) centralizes public datasets containing both gene expression and clinical data on these tumors [32]. This system enables users to carry out different class comparison analyses, including the comparison of expression distribution profiles, tests for differential expression and patient survival analyses and to compare personal results with the results in the existing literature (http://bioinfo.curie.fr/ittaca).

2.2 Standard Annotations

The generation of large amounts of data and the need to share and compare these data bring about challenges for both data management and data annotation and highlight the need for standards. The Microarray Gene Expression Data (MGED) society is an international organization created in 1999 for facilitating sharing of functional genomics and proteomics array data. MGED has defined the Minimum Information About a Microarray Experiment (MIAME) that corresponds to the minimum information that must be reported about a microarray experiment to enable its unambiguous interpretation and reproduction. This standard has been used for years worldwide. The Microarray Gene Expression Object Model (MAGE-OM) and resulting markup language (MAGE-ML) provide a mechanism for standardizing data representation for data exchange purposes [33]. Moreover, a common terminology, the MGED Ontology (MO) has been developed by the Ontology Working Group of the MGED society to complement these standards. The objective of MO is to provide common 'terms for annotating experiments in line with the MIAME guidelines, i.e., to provide the semantics to describe a microarray experiment according to the concepts specified in MIAME' [34]. (http://mged.sourceforge.net/ontologies/index.php.)

Similar efforts in the field of functional annotation have established standard vocabularies for the annotation of genes and gene products [35]. With the aim of contributing to the unification of biological information, the Gene Ontology (GO) has been developed since 2000 [36, 37] and has been adopted by most model organism databases, such as the Gene Ontology Annotation (GOA) database [38] (http://www.ebi.ac.uk/GOA).

Moreover, some research communities have decided to standardize their data models and data types to address interoperability issues. One of the requirements for a federated information system is interoperability, i.e., the ability of one computer system to access and use the resources of another system. In order to meet this need, the U.S. National Cancer Institute Center for Bioinformatics (NCICB) has created the cancer Common Ontologic Representation Environment (caCORE) to address interoperability issues in the field of cancer research [39]. The caCORE system includes controlled terminologies such as the NCI Thesaurus (NCIT) [40], as well as common data elements (CDEs), which are named identifiers for the entities and attributes found in databases.

However, despite these standardization efforts, not all the data created, stored, and made available in the biomedical domain are homogeneously represented. Because most biomedical systems have been developed independently of each other, these systems do not have a common structure, nor do they share common data elements. Because determining the correspondences between heterogeneous data sources is complex and time-consuming, automated support is needed [41]. Several approaches have been proposed, either based on the comparison of data-elements (schema-level approaches) or based on the comparison of value sets of data elements coming from distinct sources (instance-level approaches) [42, 43, 44].

2.3 Infrastructures and Services

Biomedical research requires to pool and to integrate information from diverse data sources, which is facilitated by the use of common data models and common ontologies. Additionally, coordinated research efforts typically span multiple institutions. Therefore, there is a need for an infrastructure that supports such collaborative efforts, with the objective of enabling more efficient access to the resources and sharing distributed computational resources. To address this need, the U.S. National Cancer Institute (NCI) has initiated a nationwide effort, called the cancer Biomedical Informatics Grid (caBIG), to develop a federation of interoperable research information systems [45]. At the heart of the caBIG approach to federated interoperability is a Grid middleware infrastructure, called caGrid. [46]. Moreover, this infrastructure is based on the caCORE system mentioned earlier, which supports the creation of interoperable biomedical information systems. Similar efforts in Europe have established grid infrastructures for sharing computational resources in bioinformatics (e.g., http://www.embracegrid.info) and enabling cooperative research in biomedical research [47], for example in infectious diseases [48] and immune diseases [49], as well as in cancer research [50].

More generally, grid technologies are expected to facilitate the launch and ongoing management of coordinated cancer research studies involving multiple institutions, to provide the ability to manage and securely share information and analytic resources. Additionally, grid computing supports high-throughput data analysis and predictive classification studies on large datasets [51]. Grid computing can also support the modeling of complex biological systems, which requires advanced computer simulations to bring together knowledge at all the different levels of biological understanding -- from the cell (e.g., gene function) to the organism (e.g., physiology) -- in order to provide a coherent theory of biology, which can then be applied to clinical medicine.

In conjunction with the development of distributed databases and grid computing, an increasing number of tools in biomedical informatics have been developed as Web Services, with potential applications in genomic medicine (e.g., [52]). Web Services offer two major benefits for the biomedical community: interoperability and reusability. Web Services use standard communication protocols over the Internet, which makes them virtually platform-independent. Instead of developing a specific service locally, developers can reuse Web Service components in their own applications. With the objective of implementing complex data analysis processes, Web Services must be associated with workflow management systems (e.g., [53]). Environments such as Taverna provide a language and software tools to create and execute workflows and to construct highly complex analyses over public and private data and computational resources [54, 55].

In the near future, these efforts will hopefully be strengthened by the creation of publicly available registries that describe all these services in a standard manner. For example, Stevens et al [56] recommend the use of ontologies to express the semantic information associated with the description of Web Services. The design of broad-coverage formal models of tasks and their representation as formal ontologies will facilitate the discovery of services, their selection and their composition into dynamic workflows [57].

2.4 Meta-analysis

One advantage of integrating large numbers of microarray studies and compiling them in a data-warehouse is that it makes it possible to compare the results of different studies and to determine which methods are robust and produce consistent results across a range of studies. There are, however, many problems associated with the comparison of gene expression profiles across disparate microarray data sets. In studies performed in 2004 and 2007 by several teams, the authors demonstrated that the consistency of replicates in each experiment exhibits a large degree of variation. Different technologies seemed to show good agreement within and across labs using the same RNA samples. The variability between two labs using the same technology was higher than that between two technologies within the same lab. Moreover, the source of RNA samples can make a difference in microarray data [58, 59, 60].

Several methods have been developed to address these variability issues in multiple, independent data sets generated on various platforms. Among others :

- Comparative meta-profiling is used in Oncomine to compare differential expression measured in each data set [61]. In this approach, users first select appropriate studies for comparison, and then use meta-analysis to identify the genes that are significantly overexpressed or underexpressed across multiple independent studies.
- SubMap is an unsupervised subclass mapping method, which reveals common subtypes between independent data sets. This method revealed the correspondence between several cancer-related data sets. Notably, it identified common subtypes of breast cancer associated with estrogen receptor status, and a subgroup of lymphoma patients who share similar survival patterns,

thus improving the accuracy of a clinical outcome predictor [62].

The approach associating data integration and meta-analysis helps address statistical methodological issues [63]. Data related to the same pathologic condition from different laboratories may be analyzed (e.g. [64]). For example, Bhanot et al have used classification models with non-Hodgkin's lymphoma-related microarray data from different laboratories [65], and Lyman et al have used meta-analysis techniques to detect predictors of recurrence-free survival in breast cancer [66]. Data integration may also be used with data corresponding to different diseases, for example different types of cancers [67]. Different kinds of experimental data can be integrated (e.g., microarray and proteomics). Moreover, data from different species can be integrated. For example, English and Butte evaluated 49 obesity-related genome-wide experiments including microarray, genetics, proteomics and gene knock-down from human, mouse, rat and worm. They created an integrative model and showed that intersecting the results of experiments significantly improved the sensitivity, specificity and precision of the prediction of obesity-associated genes [68].

3 Integrating Knowledge

Computable forms of knowledge include knowledge bases and ontologies. Existing resources are often incomplete and need to be enriched and integrated. Incorporating prior knowledge into the analysis of gene expression datasets has been shown to improve the results.

3.1 Knowledge sources and ontologies

Multiple knowledge bases. The number of data sources has grown tremendously over the last decade. Frey et al mention that around 900 biological public databases (e.g., genomic, proteomic, metabolomic, and others) were available in 2007, representing a vast amount of information about genes, proteins, diseases and their interrelations [1]. Besides repositories of experimental data, many knowledge resources are also publicly available. Such resources typically compile manually curated knowledge extracted from the biomedical literature and other sources. For example Entrez Gene provides information about genes, Online Mendelian Inheritance in Man (OMIM) provides information about genetic diseases and GOA provides the functional annotation of gene products.

Multiple ontologies. Ontologies have been developed to represent the entities of biomedical interest and their relations, in multiple subdomains and for multiple levels of granularity. Figure 1 shows ontologies from genomics (white), chemistry (blue), anatomy (yellow), and diseases (green). Some reference ontologies are domain-specific such as the Chemical Entities of Biological Interest (ChEBI) for chemical entities or the Foundational Model of Anatomy (FMA) for anatomical entities [69]. Some ontologies are level-specific such as GO at the cellular level, or SNOMED at the organism level. Ontologies can be overlapping in part. For example, subcellular anatomical entities are defined in both the FMA and the Cell Component axis of GO [70]. In contrast, some ontologies may reuse the entities defined in other ontologies. For example, reasoning over the anatomical location of diseases in a clinical ontology can be delegated to the anatomical ontology in which the anatomical entities are defined [71].

Ontology repositories. The use of ontologies is a key element to interoperability among resources. For this reason, high-quality ontologies must be available to the community, ideally at no cost and without any constraints

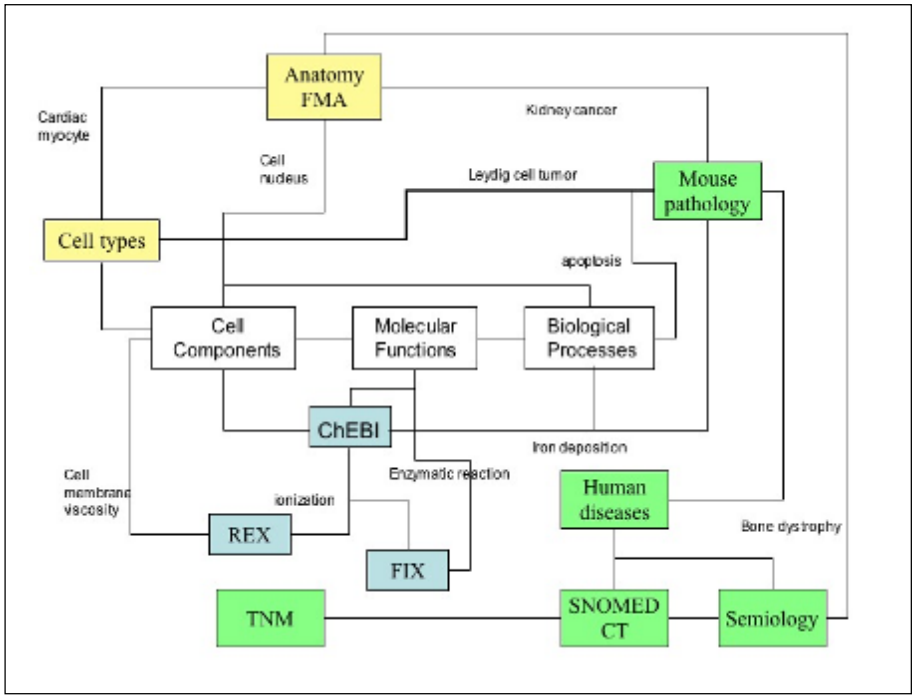

Fig. 1 Interrelations among biomedical ontologies

impeding their use or redistribution. The Open Biomedical Ontologies (OBO) are a collection of controlled vocabularies freely available to the biomedical community. Web-based ontology portals such as the BioPortal (http://www.bioontology.org/tools/portal/bioportal.html) allow users to browse, search, and visualize ontologies (and metadata) in the library, and to submit an ontology to the library. Ontology portals also tend to include features popularized by the "Web 2.0" movement, including the collaborative review of ontologies by users [72]. The need for innovative technology and methods that allow scientists to record, manage, and disseminate biomedical information and knowledge in machine-processable form gave rise, in part, to initiatives such as the National Center for Biomedical Ontology (NCBO) created in 2005[6] [73].

Ontology federation. The development of OBO ontologies is regulated within the OBO Foundry, which defines a set of shared principles governing ontology development [74]. Knowledge integration will also benefit from the development of top-domain ontologies, such as BioTop [75]. Such ontologies define the top-level classes of biomedical ontologies and can be used for linking finer-grained domain ontologies. Of note, some recently created ontologies were designed to be interoperable and to incorporate accurate representations of biological reality [74]. For example, the PRotein Ontology (PRO) includes connections to other ontologies, including GO. It is expected that the connection of protein forms to GO classes using appropriate relations will support accurate functional annotation. Analogously, relations defined between protein classes and the OBO Disease Ontology will facilitate disease understanding [76]. Until the development of federated biomedical ontologies is fully orchestrated by organizations such as the OBO Foundry - if it ever is, there will be a need for creating *ad hoc* bridges across existing ontologies, which is one of the objectives of the Unified Medical Language System (UMLS)[7] developed by the US National Library of Medicine. The UMLS Metathesaurus integrates 1.4 million concepts from over one hundred terminologies in use in life sciences, as well as some 12 million relations among these concepts. UMLS concepts are not only inter-related, but may also be linked to external resources such as GenBank, providing easy access to the knowledge contained in these resources [77]. More generally, various approaches to aligning existing ontologies are discussed in [78].

Semantic Web for Health Care and Life Sciences. Knowledge integration efforts have benefited from the development of Semantic Web technologies [21]. In the past few years, the World-Wide Web Consortium (W3C) has developed a set of standards and tools to support the vision of a flexible, integrated, automatic and self-adapting Web. Some of these technologies are now mature and have started making an impact in the life sciences. Semantic Web languages include the Resource Description Framework (RDF), a variety of data interchange formats (e.g., RDF/XML, N3, Turtle, N-Triples) and notations, such as RDF Schema (RDFS), and the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain. OWL provides formal computational definitions, as well as tools for reasoning, in order to facilitate ontology development and ontology maintenance. Therefore most health science ontologies, including those originally developed in OBO format [79], have been converted to OWL [80, 81].

3.2 Knowledge Enrichment

Standard terminologies, such as the Gene Ontology, are widely used in databases and knowledge bases as controlled vocabularies for functional annotations and largely facilitate comparative functional analysis. However, the functional annotation of gene products is not always consistent across databases and often remains incomplete. Although GO curators adhere to the same protocols and standards while assigning GO annotations, specific annotation procedures and the specialization of curators vary across groups. Methods have been developed to assess the consistency of GO annotation across model organism databases (e.g., [82]).

Enriching biological knowledge bases. Determining the function of uncharacterized proteins remains a major challenge and is an active field of research. Various knowledge sources have been explored, including large scale protein-protein interaction assays, global mRNA expression analyses and systematic protein localization studies in [83]). Various techniques have been explored as well to generate functional annotation predictions, among which information theory-based semantic similarity, based on existing GO annotations [84].

Methods based on natural language processing and statistical techniques have been widely used for years for mining free text and extracting GO annotations. While the content of most biological databases is acquired through careful manual curation of literature and data, the increasing volume of biomedical literature to be reviewed and

[6] http://www.bioontology.org/

[7] http://umlsks.nlm.nih.gov

the increasing number of gene products in need of annotation are likely to overload the manual curation process. Consequently, text mining techniques are often employed to retrieve and extract functional annotation from the literature. For example, GoPubMed uses GO to organize the results of a PubMed search [85]. The BioCreAtIvE initiative, with tasks such as gene name normalization and identification of functional annotation from free text, demonstrated that term recognition techniques are suitable for real applications in biology [86]. However, automatic annotation techniques generally require additional knowledge processing and had lesser performance than gene identification tasks. Daraselia et al also showed the usefulness of combining NLP techniques (protein annotation extracted from Medline) with additional knowledge (information from protein-protein interactions datasets) [87].

Enriching biomedical ontologies. Analogous to the methods devoted to quality assurance and enrichment of knowledge bases, methods have been developed for the evaluation of ontologies, including terminology enrichment and consistency checking.

Terminology enrichment techniques are used for identifying missing relations in terminologies. For example, GO lacks explicit associative relations across its three hierarchies, which may impede the consistent clustering of gene products according to functional characteristics. For instance, while the gene APOC3 is associated with both the molecular function 'lipid transporter activity' and the biological process 'lipid transport', APOH is only annotated with 'lipid transporter activity'. To address this issue, various approaches to suggesting new relations among biological terms have been proposed, based on lexical and statistical phenomena. Biological terms are often found as proper substrings of other terms. Compositionality of terms has been used to suggest semantic relations among GO terms directly [88, 89] or through ChEBI terms [90]. Moreover, Mungall proposed a formal language, Obol, for defining allowed compositional patterns among terms from OBO ontologies [91]. Statistical and data mining techniques have also been applied to biological knowledge bases annotated to the GO in order to automatically extract candidate relations among GO terms and help enrich ontologies with associative relations [92].

When ontologies are represented with formal languages and defined in reference to formal upper-level ontologies, it becomes possible to validate existing relations among classes and to identify new relations. OWL, the Web Ontology Language, is often used to represent the concepts and the relations in ontologies. OWL is more expressive than XML, RDF, and RDF-S, because it contains additional features for describing properties and classes formally. Such features include equivalence and disjointness among classes, cardinality of relations (e.g., "exactly one"), characteristics of properties (e.g., symmetry), and enumerated classes. Using the formal semantics of the OWL language makes it possible to reason about these classes and their instances and to ensure the consistency of these ontologies.

3.3 Strategies for Analysis and Applications

Key to the analysis of omics data is the integration of prior knowledge. Of special interest are methods that include functional characteristics from the beginning of the data analysis process, integrate medical knowledge with biological knowledge, and combine mining techniques with inference-based knowledge processing.

The analysis of transcriptomic data is classically carried out in two steps. First, data are clustered according to gene expression levels in order to create three clusters: over-expressed, under-expressed and invariant. Only subsequently is functional information introduced in order to characterize the clusters "functionally". One of the limitations of this approach is that functional similarity does not contribute to the clustering process. Methods including functional annotation from the beginning of the analysis have been proposed (e.g., [93]). These methods rely, for example, on semantic similarity measures among genes based on functional annotations [94].

Moreover, besides gene expression, proteomic patterns, functional characteristics of genes and the medical features associated with a sample (e.g., phenotype, clinical history, environmental factors, experimental conditions) could contribute to the clustering process. Such characteristics can be represented as UMLS concepts [95], NCIT or SNOMED CT concepts [96, 97]. Once annotated to these ontologies, the datasets can be clustered in such a way that the annotations themselves participate in the clustering, along with the expression profiles of the genes. More generally, knowledge integration has been shown to increase the power of analysis in several genomic studies. Butte has developed an approach based on the UMLS [95], while other authors have integrated Entrez Gene and GO [98]. Chabalier has proposed a method for integrating information from the KEGG pathway database and the GO annotation repository into a disease ontology [99].

Various data mining techniques have been applied to biomedical data analysis (e.g., [100], [101]). Among data

mining techniques, association rule mining, used widely in the area of market basket analysis, can be applied to the analysis of biological data as well. Based on the frequencies of co-occurrence between a gene G and a phenotype P, a typical rule would be: "if P is present, then G is present". Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression profiles. The mining techniques may include negative rule generation (e.g., [102]) in addition to positive rule generation. Ideally, data mining techniques should be combined with inference-based knowledge processing. For example, the classification capabilities associated with ontologies may be used to aggregate annotations in order to improve the support and confidence values of association rules. More generally, knowledge bases and inference may contribute to increase the power of data mining techniques.

4 Conclusion

As biomedical research evolves from traditional clinical and biological research towards omics sciences and translational research, specific needs have emerged, including integrating data collected in research studies with patient clinical data, linking omics knowledge with medical knowledge, modeling the molecular basis of diseases, and developing tools that support in-depth analysis of research data. As such, translational research illustrates the need to bridge the gap between bioinformatics and medical informatics [103], and opens new avenues for biomedical informatics research.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Frey LJ, Maojo V, Mitchell JA. Bioinformatics linkage of heterogeneous clinical and genomic information in support of personalized medicine. Methods Inf Med 2007;46 Suppl 1:98-105.
2. Golub TR. Genomics: global views of leukaemia. Nature 2007 Apr 12;446(7137):739-40
3. Staudt LM, Dave S. The biology of human lymphoid malignancies revealed by gene expression profiling. Adv Immunol 2005;87:163-208.
4. Mendiratta P, Febbo PG. Genomic Signatures Associated with the Development, Progression, and Outcome of Prostate Cancer. Mol Diagn Ther 2007;11(6):345-354.
5. Bredel M, Bredel C, Juric D, Harsh GR, Vogel H, Recht LD, et al. High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. Cancer Res 2005 May 15;65(10):4088-96.
6. Weir BA et al. Characterizing the cancer genome in lung adenocarcinoma. Nature 2007 Dec 6; 450(7171):893-8.
7. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet 2003 Jan;33(1):49-54.
8. Osman I, Bajorin DF, Sun TT, Zhong H, Douglas D, Scattergood J, et al. Novel blood biomarkers of human urinary bladder cancer. Clin Cancer Res 2006 Jun 1;12(11 Pt 1):3374-80
9. Wei G, Twomey D, Lamb J, Schlis K, Agarwal J, Stam RW, et al. Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. Cancer Cell 2006 Oct;10(4):331-42.
10. Stegmaier K, Wong JS, Ross KN, Chow KT, Peck D, Wright RD, et al. Signature-based small molecule screening identifies cytosine arabinoside as an EWS/FLI modulator in Ewing sarcoma. PLoS Med 2007 Apr;4(4):e122.
11. Tomalik-Scharte D, Lazar A, Fuhr U, Kirchheiner J. The clinical role of genetic polymorphisms in drug-metabolizing enzymes. Pharmacogenomics J 2008 Feb;8(1):4-15
12. Donald PR, Parkin DP, Seifart HI, Schaaf HS, van Helden PD, Werely CJ, et al. The influence of dose and N-acetyltransferase-2 (NAT2) genotype and phenotype on the pharmacokinetics and pharmacodynamics of isoniazid. Eur J Clin Pharmacol 2007 Jul;63(7):633-9. Epub 2007 May 16.
13. Wechsler ME. Managing asthma in the 21st century: role of pharmacogenetics. Pediatr Ann 2006 Sep;35(9):660-2, 664-9.
14. Lay J, Liyanage R, Borgmann S, Wilkins CL, Problems with the "omics" TrAC Trends in Analytical Chemistry, 2006 Dec,25(11):1046-56.
15. Dupuis J, O'Donnell C. Interpreting results of large-scale genetic association studies: separating gold from fool's gold. JAMA 2007 Feb 7;297:529- 31.
16. Larsson O, Wennmalm K, Sandberg R. Comparative microarray analysis. OMICS 2006 Fall;10(3): 381-97. Review.
17. Stransky B, Barrera J, Ohno-Machado L, De Souza SJ. Modeling cancer: integration of "omics" information in dynamic systems. J Bioinform Comput Biol 2007 Aug;5(4):977-86. Review.
18. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998 Nov;37(4-5):394-403. Review.
19. de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW NCI Thesaurus: using science-based terminology to integrate cancer research results. Medinfo 2004;11(Pt 1):33-7.
20. Brazhnik O, Jones JF. Anatomy of data integration.J Biomed Inform 2007 Jun;40(3):252-69.
21. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics 2007 May 9;8 Suppl 3:S2.
22. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Brief Bioinform 2008 Jan;9(1):75-90. Epub 2007 Dec 12.
23. Lambrix P, Tan H. SAMBO - A System for Aligning and Merging Biomedical Ontologies. Journal of Web Semantics 2006, 4, 3
24. Kumar A, Yip YL, Smith B, Grenon P. Bridging the gap between medical and bioinformatics: an ontological case study in colon carcinoma. Comput Biol Med 2006 Jul-Aug;36(7-8):694-711
25. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform 2006 Sep;7(3):256-74. Epub 2006 Aug 9. Review.
26. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002 Jan 1;30(1):207-10.
27. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 2007 Jan;35(Database issue):D760-5.
28. Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, et al. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics 2006 Nov 6;7:489
29. Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, et al. The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Res 2005 Jan 1;33(Database issue):D580-2.
30. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 2007 Jan;35(Database issue):D747-50.
31. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 2007 Feb;9(2):166-80.
32. Elfilali A, Lair S, Verbeke C, La Rosa P, Radvanyi F, Barillot E. ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis. Nucleic Acids Res 2006 Jan 1;34(Database issue):D613-6.
33. Ball CA, Brazma A. MGED standards: work in progress. OMICS 2006 Summer;10(2):138-44.
34. Whetzel PL, Parkinson H, Causton HC, Fan L, Fostel J, Fragoso G, et al. The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics 2006 Apr 1;22(7):866-73.
35. Blake JA, Bult CJ. Beyond the data deluge: data integration and bio-ontologies. J Biomed Inform

2006 Jun;39(3):314-20
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000 May;25(1):25-9
37. The Gene Ontology Consortium. The Gene Ontology project in 2008. Nucleic Acids Res 2007 Nov 4 [Epub ahead of print].
38. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, et al. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D262-6.
39. Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, et al. caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform 2007 Apr 2 [Epub ahead of print]
40. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007 Feb;40(1):30-43.
41. Shvaiko P, Euzenat J. A survey of schema-based matching approaches. Journal on data semantics 2005,4:146-71.
42. Alonso-Calvo R, Maojo V, Billhardt H, Martin-Sanchez F, García-Remesal M, Pérez-Rey D. An agent- and ontology-based system for integrating public gene, protein, and disease databases. J Biomed Inform 2007 Feb;40(1):17-29.
43. Zhao H, Ram S. Combining schema and instance information for integrating heterogeneous data sources. Data & Knowledge Engineering 2007, 61 (2): 281-303.
44. Mougin F, Burgun A, Bodenreider O. Mapping data elements to terminological resources for integrating biomedical data sources. BMC Bioinformatics 2006 Nov 24;7 Suppl 3:S6.
45. caBIG Strategic Planning Workspace. The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. Medinfo 2007;12(Pt 1):330-4.
46. Saltz J, Oster S, Hastings S, Langella S, Kurc T, Sanchez W, Kher M, et al. caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. Bioinformatics 2006 Aug 1;22(15):1910-6.
47. Schroeder M, Burger A, Kostkova P, Stevens R, Habermann B, Dieng-Kuntz R. Sealife: a semantic grid browser for the life sciences applied to the study of infectious diseases. Stud Health Technol Inform 2006;120:167-78.
48. Emerson A, Rossi E. ImmunoGrid - the virtual human immune system project. Stud Health Technol Inform 2007;126:87-92.
49. Müller H, Pitkanen M, Zhou X, Depeursinge A, Iavindrasana J, Geissbuhler A. KnowARC: enabling Grid networks for the biomedical research community. Stud Health Technol Inform 2007;126:261-8.
50. Tsiknakis M, Kafetzopoulos D, Potamias G, Analyti A, Marias K, Manganas A. Building a European biomedical grid on cancer: the ACGT Integrated Project. Stud Health Technol Inform 2006;120:247-58.
51. Konagaya A. Trends in life science grid: from computing grid to knowledge grid. BMC Bioinformatics 2006 Dec 18;7 Suppl 5:S10. Review.
52. Maojo V, Crespo J, de la Calle G, Barreiro J, Garcia-Remesal M. Using web services for linking genomic data to medical information systems. Methods Inf Med 2007;46(4):484-92.
53. Romano P, Bartocci E, Bertolini G, De Paoli F, Marra D, Mauri G, et al. Biowep: a workflow enactment portal for bioinformatics applications. BMC Bioinformatics 2007 Mar 8;8 Suppl 1:S19. 8.
54. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, et al. Taverna: a tool for building and running workflows of services. Nucleic Acids Res 2006 Jul 1;34(Web Server issue):W729-32.
55. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 2004 Nov 22;20(17):3045-54.
56. Stevens RD, Robinson AJ, Goble CA. myGrid: personalised bioinformatics on the information grid. Bioinformatics 2003;19 Suppl 1:i302-4.
57. Wolstencroft K, Alper P, Hull D, Wroe C, Lord PW, Stevens RD, et al. The (my)Grid ontology: bioinformatics service discovery. Int J Bioinform Res Appl 2007;3(3):303-25.
58. Wang H, He X, Band M, Wilson C, Liu L. A study of inter-lab and inter-platform agreement of DNA microarray data. BMC Genomics 2005 May 11;6(1):71.
59. Jarvinen A, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi O, et al. Are data from different gene expression microarray platforms comparable? Genomics 2004;83:1164-8.
60. Thompson KL, Afshari CA, Amin RP, Bertram TA, Car B, Cunningham M, et al. Identification of platform-independent gene expression markers of cisplatin nephrotoxicity. Environmental Health Perspectives 2004;112:488-94.
61. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. Large-scale metaanalysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA 2004;101:9309-14.
62. Hoshida Y, Brunet JP, Tamayo P, Golub TR, Mesirov JP. Subclass mapping: identifying common subtypes in independent disease data sets. PLoS ONE 2007 Nov 21;2(11):e1195.
63. Cahan P, Rovegno F, Mooney D, Newman JC, St Laurent G 3rd, McCaffrey TA. Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. Gene 2007 Oct 15;401(1-2):12-8. Review.
64. Fishel I, Kaufman A, Ruppin E.Meta-analysis of gene expression data: a predictor-based approach. Bioinformatics 2007 Jul 1;23(13):1599-606.
65. Bhanot G, Alexe G, Levine AJ, Stolovitzky G. Robust diagnosis of non-Hodgkin lymphoma phenotypes validated on gene expression data from different laboratories. Genome Inform 2005; 16(1):233-44.
66. Lyman GH, Kuderer NM. Gene expression profile assays as predictors of recurrence-free survival in early-stage breast cancer: a metaanalysis. Clin Breast Cancer 2006 Dec;7(5):372-9.
67. Yang X, Sun X. Meta-analysis of several gene lists for distinct types of cancer: a simple way to reveal common prognostic markers. BMC Bioinformatics 2007 Apr 6;8:118.
68. English SB, Butte AJ. Evaluation and integration of 49 genome-wide experiments and the prediction of previously unknown obesity-related genes. Bioinformatics 2007 Nov 1;23(21):2910-7. Epub 2007 Oct 5.
69. Rosse C, Mejino Jr, JLV. The Foundational Model of Anatomy ontology. In: Burger A, Davidson D, Baldock R, editors. Anatomy ontologies for bioinformatics: principles and practice, New York: Springer; 2008. p. 59-117 .
70. Agoncillo AV, Mejino JL Jr, Rickard KL, Detwiler LT, Rosse C; Structural Informatics Group. Proposed classification of cells in the Foundational Model of Anatomy. AMIA Annu Symp Proc 2003;:775
71. Rubin DL, Bashir Y, Grossman D, Dev P, Musen MA. Using an ontology of human anatomy to inform reasoning with geometric model. Stud Health Technol Inform 2005; 111:429-35
72. Supekar K, Rubin D, Noy N, Musen M. Knowledge Zone: a public repository of peer-reviewed biomedical ontologies. Medinfo 2007;12(Pt 1):812-6.
73. Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS 2006 Summer;10(2):185-98.
74. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007 Nov;25(11):1251-5.
75. Stenzhorn H, Beisswanger E, Schulz S. Towards a top-domain ontology for linking biomedical ontologies. Medinfo 2007;12(Pt 2):1225-9.
76. Natale DA, Arighi CN, Barker WC, Blake J, Chang TC, et al. Framework for a protein ontology. BMC Bioinformatics 2007 Nov 27;8 Suppl 9:S1
77. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-70.
78. Euzenat J, Schvaiko P Ontology matching. Springer-Verlag, Berlin Heidelberg (DE); 2007. pp. 333.
79. Day-Richter J, Harris MA, Haendel M; Gene Ontology OBO-Edit Working Group, Lewis S. OBO-Edit--an ontology editor for biologists. Bioinformatics 2007 Aug 15;23(16):2198-200.
80. Moreira DA, Musen MA. OBO to OWL: a protege OWL tab to read/save OBO ontologies. Bioinformatics 2007 Jul 15;23(14):1868-70.
81. Aranguren ME, Bechhofer S, Lord P, Sattler U, Stevens R. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. BMC Bioinformatics 2007 Feb 20;8:57.
82. Dolan ME, Ni L, Camon E, Blake JA. A procedure for assessing GO annotation consistency. Bioinformatics 2005 Jun;21 Suppl 1:i136-43.
83. Jiang T, Keating AE AVID: an integrative framework for discovering functional relationships among proteins.BMC Bioinformatics 2005 Jun 1;6:136
84. Tao Y, Sam L, Li J, Friedman C, Lussier YA. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics 2007 Jul 1;23(13):i529-38.
85. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res 2005 Jul 1;33(Web Server issue):W783-6.
86. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of infor-

mation extraction for biology. BMC Bioinformatics 2005;6 Suppl 1:S1.
87. Daraselia N, Yuryev A, Egorov S, Mazo I, Ispolatov I. Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. BMC Bioinformatics 2007 Jul 10;8:243.
88. Ogren PV, Cohen KB, Hunter L. Implications of compositionality in the gene ontology for its curation and usage. Pac Symp Biocomput 2005;:174-85.
89. Bada M, Hunter L. Enrichment of OBO ontologies. J Biomed Inform 2007 Jun;40(3):300-15. Epub 2006 Jul 26.
90. Burgun A. Desiderata for domain reference ontologies in biomedicine. J Biomed Inform 2006 Jun;39(3):307-13. Epub 2005 Oct 17.
91. Mungall CJ. Obol: integrating language and meaning in bio-ontologies, Comp Funct Genomics 2004, 5,509-20.
92. Bodenreider O, Aubry M, Burgun A. Non-lexical approaches to identifying associative relations in the gene ontology. Pac Symp Biocomput 2005;:91-102.
93. Chabalier J, Mosser J, Burgun A. A transversal approach to predict gene product networks from ontology-based similarity. BMC Bioinformatics 2007 Jul 2;8:235.
94. Wang H, Azuaje F, Bodenreider O. An Ontology-driven clustering method for supporting gene expression analysis. CBMS 2005::389-94.
95. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. Nat Biotechnol 2006 Jan;24(1):55-62.
96. Shah NH, Rubin DL, Supekar KS, Musen MA. Ontology-based annotation and query of tissue microarray data. AMIA Annu Symp Proc 2006;:709-13.
97. Shah NH, Rubin DL, Espinosa I, Montgomery K, Musen MA. Annotation and query of tissue microarray data using the NCI Thesaurus. BMC Bioinformatics 2007 Aug 8;8:296.
98. Sahoo SS, Zeng K, Bodenreider O, Sheth A. From "glycosyltransferase" to "congenital muscular dystrophy": integrating knowledge from NCBI Entrez Gene and the Gene Ontology. Medinfo 2007;12(Pt 2):1260-4.
99. Chabalier J, Mosser J, Burgun A. Integrating biological pathways in disease ontologies. Medinfo 2007;12(Pt 1):791-5.
100. Bresson C, Keime C, Faure C, Letrillard Y, Barbado M, Sanfilippo S, et al. Large-scale analysis by SAGE reveals new mechanisms of v-erbA oncogene action BMC Genomics 2007, 8:390.
101. Wang H, Zheng H, Simpson D, Azuaje F Machine learning approaches to supporting the identification of photoreceptor-enriched genes based on expression data BMC Bioinformatics 2006, 7:116.
102. Artamonova II, Frishman G, Frishman D. Applying negative rule mining to improve genome annotation. BMC Bioinformatics 2007 Jul 21;8:261.
103. Maojo V, Kulikowski C. Medical informatics and bioinformatics: integration or evolution through scientific crises? Methods Inf Med 2006;45(5):474-82.

Correspondence to:
Anita Burgun
Département d'Information Médicale
CHU Pontchaillou
rue Henri Le Guilloux
F-35033 Rennes Cedex
France
E-mail: anita.burgun-parenthoine@univ-rennes1.fr