Biodash: A Semantic Web Dashboard for Drug Development

Eric K. Neumann and Dennis Quan

Pacific Symposium on Biocomputing 11:176-187(2006)

BIODASH: A SEMANTIC WEB DASHBOARD FOR DRUG DEVELOPMENT

ERIC K. NEUMANN[†]
W3C, MIT
Cambridge, MA 02139 USA

DENNIS QUAN
IBM T. J. Watson Research Center, 1 Rogers Street
Cambridge, MA 02142 USA

A researcher's current scientific understanding is assembled from multiple sources of facts and knowledge, along with beliefs and hypotheses of their interpretations. A comprehensive and structured aggregation of all the relevant components is to-date not possible using standard database technologies, nor is it obvious how to include beliefs, such as models and hypotheses into such a bundle. When such information is required as the basis for important decision-making (e.g., in drug discovery), scientists often resort to using commercial presentation applications. This is sub-optimal for the effective use of knowledge, and alternatives that support the inclusion of meaning are urgently needed. This paper describes a prototype Semantic Web application, BioDash[1], which attempts to aggregate heterogeneous yet related facts and statements (using an RDF model) into an intuitive, visually descriptive and interactive display.

1. Introduction

1.1. Today's Research Informatics Problems

Scientific research relies on researchers sharing heterogeneous knowledge, experimental data, and interpretations in meaningful ways that go beyond transmitting data fragments. Although computational methods and data-exchange protocols are common to modern scientific practice, they are a small part of the overall process. Critical interpretation of experimentally derived information and the consolidation of knowledge that include alternative views and hypotheses are essential to the scientific process of debate and rebuttal.

The need for improved information systems is being recognized throughout the pharmaceutical and biotech industry. In a recent report on drug development [1], the U.S. Food and Drug Administration (FDA) stated the need for "a knowledge base built not just on ideas from biomedical research, but on reliable insights into the pathway to patients." Much still needs to be done to meet these

[†] contact: eneumann@alum.mit.edu
[1] http://www.w3.org/2005/04/swls/BioDash/Demo

goals, and current web and enterprise architectures cannot satisfy the functionality specified.

Currently, knowledge is captured in either rigid, hard-to-define databases, or in applications (PowerPoint and Excel) designed for human viewing. The former misses the inclusion of explicit scientific meaning, while the latter are difficult to query, making it hard to find or reuse knowledge (*knowledge cul-de-sacs)*. In addition, most analysis and visualization applications are not interoperable via open standards, and cannot "see" connections across data sets that could be presented to users. Finally, items such as context and hypotheses are not well encoded, severely limiting data interpretation.

The essential problem in data management is not how to store large amounts of data but how best to distill insights from them (through analysis), and associate these interpretations with the data. In so doing, knowledge would be organized for higher-level reasoning and decision-making. The Semantic Web (SW) (www.w3.org/2001/sw/), as proposed by Tim Berners-Lee, is supposed to allow meaning (i.e., semantics) to be associated with information on the Web through a universal mechanism that machines can process as well [2, 3]. It is based on two key standards: Resource Description Framework (RDF) for describing objects and the relations between them; and the Web Ontology Language (OWL, based on RDF) for specifying the supported ontologies (semantic systems of concepts and relations). OWL ontologies (one or more per RDF document) are used to define the logical types of objects and how they can relate to one another within an RDF document.

We define six areas where SW technologies could offer critical support to the life sciences: (1) database conversions and wrappers; (2) unique identifiers that are supported by the SW URI model; (3) coordination and management of terminologies and ontologies; (4) tools and viewers conversant in RDF-OWL; (5) knowledge encoding: theories, hypotheses, models; (6) semantics accounts and channels: store and share annotations based on SW. This research addresses the first four and suggests directions for the latter two.

1.2. *BioDash: A Life Science Scenario of How Things Should Be*

A real-world illustration of the need to organize and utilize complex, distributed forms of information is seen in drug discovery. The likely success of a new drug is indicated through the combined analysis of target classes, high-throughput (HT) screening, ADME, toxicity, efficacy, animal testing, and efficient clinical trials. These can determine whether a new drug is successfully launched or terminated.

This paper describes our work on BioDash, a prototype of an "information dashboard" for drug discovery, which involves a scenario based on the drug

target Glycogen Synthase Kinase 3 beta (GSK3b) [4], in which multiple forms of knowledge (genomic, biopathway, disease, chemical, and SNP data) residing in disparate repositories are brought together through SW technologies to support the discovery process.

We will begin by summarizing the basic notions of the Semantic Web – universal identifiers, the "quantitization" of knowledge, and Semantic Lenses – from the perspective of life science data integration. Afterwards, we proceed to describe how the BioDash user experience takes advantage of Semantic Web integration technologies. Ultimately, this integration is only worthwhile if it can enable users to gain insights they would otherwise be unable to if the data were kept separate. We give several examples of how BioDash permits the user to "experience" data integration across topics. Finally, we end with a discussion of additional applications that could benefit from the technology.

2. The Semantic Web Data Model within BioHaystack

2.1. *Building on the Web Model*

The World Wide Web succeeded in large part because it allows users to retrieve information from an ever-broadening range of sources through a single tool: the Web browser. In the days before the Web, users had to jump tediously from one system to another to perform complex retrieval tasks. At present, there is a lot less system-hopping thanks to hyperlinks However, there are still barriers to making effective use of that information. For example, applications do not successfully negotiate data from other applications due to differences in data formats.

Ironically, one application that is often caught "hoarding" data is our old friend, the Web browser. The data needed for various tasks are found on public domain Web pages buried in tables, bullet listings, or even prose. The characteristics that make Web pages easily consumable for humans, i.e., context-specific page layouts and inspired uses of formatting, are the very things that inhibit machine processing, which depends on data being laid out in a consistent, predetermined, "boring" fashion. Differences in data formats have made collating data from multiple web pages hard for humans and impossible for machines. While the Web has standardized the way humans retrieve information, until now it has done little to standardize data representations.

As data become easier to consume by applications, new visualization capabilities will be possible, and browsers will evolve to take advantage of them. BioHaystack, on which BioDash is built, is a prototype of a life science "Semantic Web Browser" that specifically supports information formatted for the Semantic Web. It is able to handle RDF and OWL documents, by aggregating, filtering, and rendering RDF data files into viewable and interactive

displays. BioHaystack also allows one to create new RDF information and store the new contents, and will be discussed in further detail in Sections 2.5 and 3.1.

2.2. *A more universal data exchange format*

The Web was premised on the idea that human-readable content must be written in a common format (HTML) and made available to Web browsers from content servers through the HTTP protocol. The Semantic Web requires that data published to the Web should utilize the RDF format, making it easier for applications other than the data's origin to read and incorporate them.

The key idea behind RDF is that by introducing some syntactical simplifications on XML, a number of important capabilities are enabled: (1) personal or domain-specific annotations, classifications, and other forms of knowledge can be added to any application's data without interfering with its normal function; (2) information retrieval is made easier, because RDF-enabled Web browsers and search engines can index and extract classification metadata from any RDF file; (3) arbitrary RDF data files, containing pieces of knowledge from multiple applications, can be easily merged to form a larger whole (information integration); (4) automated, rules-based processing is possible using off-the-shelf RDF inference engines.

2.3. *LSID: A More Universal Naming Scheme*

SW also requires objects that are described by RDF data files, such as gene sequences, research papers, or 3D structures, to be referred to by universal names in accordance with the Universal Resource Identifier (URI) standard, an extension of the original URL system (e.g., http://www.w3.org/). A universal naming scheme simplifies the processing of data from a variety of sources, because the application does not need to have specific, hard-coded support for each naming scheme. This allows cross-referencing between data sources to be done implicitly using URI's.

One such effort currently underway is the Life Sciences Identifier (LSID) project [5]. In our demonstration we use LSID's as external references for OMIM records as well as for Uniprot proteins in both the target data as well as the WNT pathway data:

urn:lsid:uniprot.org:uniprot:P49841

This LSID names the protein record in Uniprot that is referred to as P49841. It consists of parts separated by colons: A prefix "urn:lsid:", The authority name; The authority-specific data namespace; and the namespace-specific object identifier ("P49841").

2.4. *Statements: the Quantum Unit of RDF*

The second constraint is that RDF data files are decomposable into fundamental units of information called statements (or triples). A statement has three parts: a subject, a predicate, and an object[2]. Here are some examples of statements:

- GSK3b is-type Protein
- GSK3b has-name "Glycogen Synthase Kinase 3 beta"
- GSK3b interacts-with betaCatenin

Each of these elements are specified by their URIs (the object could also be a string values), eliminating ambiguity for machine processing. The statement from the above set could be recorded as follows:

<urn:lsid:uniprot.org:uniprot:GSK3b>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<urn:lsid:uniprot.org:uniprot:Protein>

Using the LSID resolver at http://lsid.biopathways.org, Uniprot and NCBI records are returned to BioDash as RDF statements. By breaking data files into quantum units, applications that "see" RDF statements they do not "understand" can safely ignore these (a consequence of logical monotonicity). Formats that are not based on a quantum unit, such as standard XML formats, have blurred boundaries between units of information, so the complete structure must be understood in advance. Consequently, RDF data can be combined by simple concatenation, with metadata included and additional properties appended, without any need to re-program applications that already read these files.

2.5. *Semantic Lenses*

SW requires browsers that not only collect and render the semantic documents visually, but also aggregate select information referenced by documents and data objects. Automated rules can then be applied to filter relations or create new ones so that only the relevant parts of information aggregations are shown and users are not overwhelmed by the data. The browser component that makes this possible is an intelligent information filter and viewer called a "Semantic Lens" that is created to isolate specific meaning within an arbitrary chunk of information. In BioHaystack, lenses are defined using the *Adenine* language, which supports a full set of UI components including lenses and parts[3], as well as query objects for finding and extracting

---

[2] Here "object" is used in the grammatical sense, as in object of a verb or a preposition, instead of in the sense of "entity". To avoid confusion, the SW community refers to entities as "resources".

[3] See http://haystack.csail.mit.edu/documentation/ui.pdf

additional RDF information. XML documents can be converted into RDF using XSLT scripts that can be called through Adenine. Since Adenine is declarative, there is often no need for any additional Java, C, or perl coding. BioDash is defined by the set of lenses it utilizes for Topic Views, Pathway Views, and SNP View.

```
hs:member ${
        rdf:type     data:RDFQueryAspect ;
        data:sourceExistential ?s ;
        data:targetExistential  ?t ;
        rdfs:label  "" ;
        data:existentials @( ?s ?t ?type ) ;
        data:statement ${
                data:subject            ?type ;
                data:predicate          biopax:LEFT ;
                data:object             ?s
        } ;
        data:statement ${
                data:subject            ?type ;
                data:predicate          biopax:RIGHT ;
                data:object             ?t
        }
}
```

Figure 1. Semantic Lens written in Adenine for rendering BioPAX pathways

Lenses are the SW equivalent of cascading style-sheets used by HTML browsers to enhance the HTML being viewed. However, since they handle logic and are active, semantic lenses can be applied on the back-end as well. For example, consider the problem of displaying a pathway encoded with the BioPAX standard (www.biopax.org [6]). As with any machine-readable format, BioPAX allows myriad details of the pathway to be encoded. BioPAX also defines the notion of a reaction to have a left hand side and a right hand side. Most tools today—Web browsers included—will present all of this detail on a single screen, making it difficult to decipher basic properties of the pathway such as which proteins are interacting with each other. For the purposes of an overview, it is often much more useful to filter out everything but this basic level of detail. Figure 1 shows the definition of such a filter. It directs BioHaystack to draw an arrow between the LEFT and RIGHT properties of a reaction. By defining families of lenses, completely different views of the same data can be constructed for multiple concerned parties.

3. Results: The User Experience

3.1. *The BioHaystack Semantic Web Browser*

BioDash is built on the BioHaystack Semantic Web Browser [7], a Java application that enables users to navigate, visualize, annotate, and organize data

in highly-customizable ways (www.w3.org/2005/04/swls/BioDash/Demo). Similar to a traditional Web Browser, a Semantic Web Browser provides a graphical interface to data available both locally and on the network. A user can begin his or her browsing session by entering a URI into the "Go to" box. The "pages" that are shown are graphical displays containing hyperlinks to URI's (but not in HTML); by clicking on a hyperlink, a user is taken to a "new object page". The toolbar also provides the familiar "Back", "Forward", "Refresh", and "Home" buttons.

Compared to standard Web Browsers, BioHaystack provides users with improved flexibility in how they view information. Rather than viewing data through layouts predetermined by a Web site designer, BioHaystack allows users to choose the view that is most appropriate to the task at hand. Different views for the same data can be provisioned for bioinformaticists, chemists, pathologists, or other roles. Additionally, the browser itself is capable of data integration, allowing users to incorporate data from custom data sources, such as local files or secondary data stores.

3.2. *Building a Drug Target Model*

Central to most drug development strategies is the mapping of gene/protein target information to bioactive compounds. Targets are typically organized into classes, where a handful of target classes map to 80% of approved drugs. The objective is to identify new classes of targets, which are usually identified and validated based on the significance they play in a disease- a key piece of knowledge for all drug R&D. Additionally, information of "anti-targets" and secondary targets are of interest as well, since these can be used to improve compound selectivity. Such information comes from either empirical HT screening assays (compounds to isolated proteins), or from *in silico* modeling of molecular interactions between ligands and proteins.

BioDash organizes information about targets, investigated compounds, therapeutic areas, and other relevant data into *therapeutic topics*, defined by the LS-Ont bridge ontology (http://www.w3.org/2005/04/swl/BioDash/ls-ont.rdf). The BioDash topic view, seen in Figure 2, incorporates a series of views or lenses (see semantic lenses below) that give an overview of the status of the topic. The *Target Overview* lens shows the chemical entities being considered that target (arrows) GSK3b. The *Primary* (project team disease focus) and *Alternative* (potential future applications) *Disease* lenses render information about the diseases in which the target has been implicated; here we use descriptions from OMIM. The *Group Members* lens gives a listing of the people involved in the effort, as well as their roles and emails. Finally, we have

included published antagonists of GSK3b in the RDF demonstration set, including their chemical structures and properties.

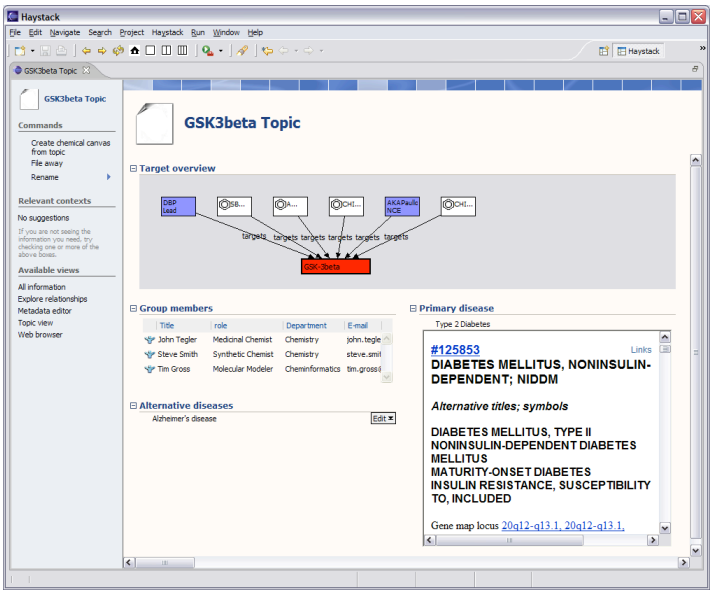

Figure 2. The Topic View containing the Target Overview, Primary Disease, and Group members lenses

3.3. *Finding Multiple Intervention Points*

A powerful way to understand the interaction of drugs with biological systems is to see the relation between compounds and the molecular pathways that they are presumed to affect. Such a perspective can be especially insightful when searching for optimal intervention points that modulate a key process with reduced chances for adverse effects. Examining the data this way often highlights differences in tissue specificity, downstream effects, and regulation type. In addition, by considering molecular processes, multi-target therapies can be developed whereby drug combinations can more effectively modulate a process.

To support this mode of investigation, we have incorporated a *Pathway View* into BioDash that can render pathways encoded using the BioPAX representation. Figure 3 shows a screenshot of BioDash rendering the WNT pathway, in which GSK3b plays a role. The Pathway View shows information that is quite distinct from the Topic View. First, information for the Pathway View comes primarily from public pathway databases (e.g., BioCyC), whereas

the Topic View is populated with internal topic-tracking status data. Second, the information rendered by the two views is represented with completely different ontologies. Additionally, the Pathway View is rendered as a full-screen graph, while the Topic View is a segmented display with both graphical and tabular diagrams. Finally, while the two displays both depict GSK3b, the various data sets use different names for GSK3b.

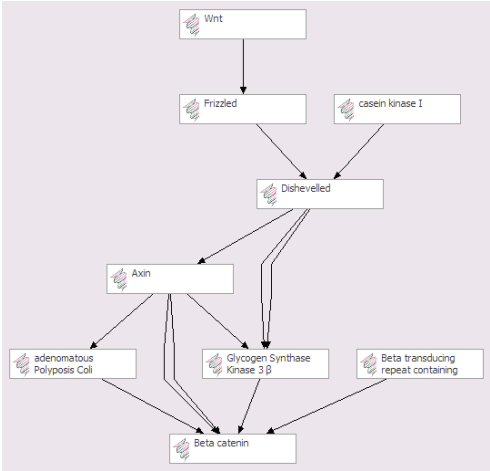

Figure 3. A BioPAX-encoded WNT pathway, rendered using a semantic lens.

Despite these differences, it is still possible to aggregate information from the two views. The Pathway View is designed to accommodate other forms of information, using proteins marked with Uniprot IDs as "pivot points" (see LSID above). In our scenario, the pivot point between the GSK3b Pathway View and the WNT Pathway View is GSK3b itself. If the user drags the red GSK3b icon from the Topic View onto the Pathway View, BioDash merges the two diagrams together (see Figure 4). The significance of this merge is twofold: first, contrary to the commonly held belief that ontologies require significant development effort to interoperate, hardly any coordination was required between the drug topic ontology and the BioPAX ontology; only a common Uniprot identifier was needed. Second, the merge exposes information (using a rule) that was not present in either of the two diagrams alone (but present in the data set): the fact that one of the chemical entities under consideration also targets casein kinase I, another player in the WNT pathway.

3.4. *Sensitivity to Polymorphisms*

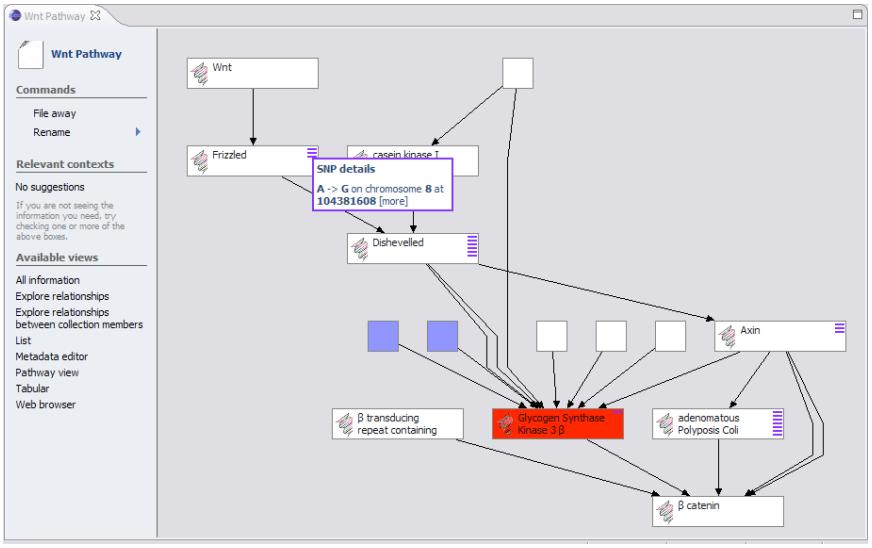

Figure 4. Non-synonymous SNP information (shown as purple dashes) aggregated onto proteins in the WNT pathway along with the compounds that target GSK3 beta all represented together as RDF.

With the advent of *personalized medicine*, pharmacogenomics will play an increasing role in assessing the safety and utility of drugs as determined by the variations of an individual's genetic background. The genotype each person inherits from their parents, tends to follow the distributions of their ancestral sub-population. Genomic variation information today can be obtained for most gene loci from single nucleotide polymorphic (SNP) databases such as dbSNP (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp). Consequently, families of polymorphisms can be aggregated onto sets of genes (proteins) that all belong to a common pathway. When overlaid onto pathways, the SNP variations along with their known population distributions can be used to predict what branches of a biological process might be more susceptible to genotypic variations, possibly affecting drug responses or causing side effects for different individuals. This information is queried from dbSNP and the returned XML results can be converted into RDF using XSLT and displayed graphically as purple bars shown on the right hand side of the protein tiles in the Pathway View. By clicking on a bar, a SNP summary pop-up appears that allows more information to be retrieved. Since the polymorphic-pathway is semantically defined as RDF, additional computational reasoning can be performed.

We specifically queried for non-synonymous polymorphisms (i.e., protein sequence changes) for each gene, since these have the highest probability of

affecting the function of a pathway component directly, enabling one to understand the functional range of each component in a pathway context. At such a time when clinical data on individuals becomes available, extending this model to handle individual genotypic (via genetic diagnostics) plus clinical evidence to assess which polymorphisms influence therapeutic responses would be straightforward. Thus polymorphic mapping onto pathways could serve as a scaffold for aggregating clinical data into a semantic structure to analyze complex interactions between pathway components and their polymorphisms.

4. Conclusion

In the SW paradigm, we begin to consider biological, chemical, and clinical information as part of a viewable and computable web of related facts and hypotheses, not simply as disassociated data fragments. Many traditional data models were defined at a time when data were submitted in chunks. However, databases such as Entrez [8] and Reactome [9] have much more intrinsic connectivity to related information of diverse forms, though they represent the semantics implicitly. If the semantics were explicitly defined using RDF/OWL, emerging SW applications could make full use of their information. Some data sources including UniProt (www.isb-sib.ch/~ejain/rdf/) have already been converted to RDF. Even so, SW tools such as BioDash can already take advantage of structured life science resources by converting XML files into RDF or mapping databases to RDF using wrappers. In addition, most life science data objects and documents can be uniquely tracked with URI's, either through LSID's or URL's appended with identifiers.

In this project we demonstrate that relevant facts can be collected from multiple sources, combined semantically, and viewed using a SW browser. Semantic Web Browsers will be necessary since the full complement of RDF-based information is too complex for humans to take in all at once. In Drug Discovery, processes are segmented from each other and information from one set needs to be provided to subsequent steps (e.g., selected targets for defining HT screening), using the knowledge perspectives local to each step. It is also possible to postulate hypotheses as RDF statements, and share these points of view as part of the topic. Furthermore, such additions could be distributed using RSS (based on RDF) newsfeed technology. Some open issues still require consideration: standard do not exist yet for semantic lenses; models for aggregation or knowledge sharing are lacking; and memory limitations on the client-side may suggest that large aggregations be performed on back-end servers. Nonetheless, BioDash offers a practical test-bed for asking these questions in different contexts over a broad range of research areas.

The use of aggregators and Semantic Web Browsers can be applied to other areas requiring embedded semantics: medical language systems [10], health care

management [11], chemistry [12], cancer research [13], clinical trial management [14], and analytical workflows (myGRID) [15]. Shifting emphasis to knowledge representations allows aggregation and reasoning between all information sets, and can support the managing of information across different communities. Semantic Lenses offer an intelligent and powerful means to organize interlinked information specific to a user's needs, supporting the construction and use of collective knowledge.

Acknowledgement

We would like to thank Melissa Cline, Joanne Luciano, Eric Prud'hommeaux, Susie Stephens, and John Wilbanks for their contributions to this project.

References

1. FDA Report, U.S. FDA, March 2004.
2. T. Berners-Lee, J. Hendler, O. Lassila. Sci. Am. 284:34-43 (May 2001).
3. E. Neumann, Science STKE 2005, pe22 (2005).
4. Cohen P, Goedert M. Nat Rev Drug Discov. 2004 Jun;3(6):479-87. Review. PMID: 15173837
5. T. Clark, S. Martin, T. Liefeld, Brief Bioinform. March. 5, 59–70 (2004).
6. J. Luciano, Drug Discovery Today, Vol 10, No. 13, 938-942 (2005).
7. D. Quan, D. Karger, Proceedings of the 13th International Conference on World Wide Web, pg 255-265 Association Computing Machinery Press.
8. D. L. Wheeler, *et al.,* Nucleic Acids Res. 33, D39–D45 (2005)
9. G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L Matthews, S. Lewis, E. Birney, L. Stein, Nucleic Acids Res. 33, D428–D32 (2005).
10. V. Kashyap, American Medical Informatics Association Annu. Symp. Proc. 2003, 351–355 (2001).
11. G. Goebel, K. L. Leitner, K. Pfeiffer, Medinfo 2004, 1618 (2004).
12. P. Murray-Rust, H. S. Rzepa, S. M. Tyrrell, Y. Zhang, Org. Biomol. Chem. 2 (22), 3192–3203 (2004).
13. S. De Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, L. W. Wright, Medinfo 2004, 33–37 (2004).

M. N. Kamel Boulos, A. V. Roudsari, E. R. Carson, Med. Inform. Internet Med. Sep. 27, 127–137 (2002).

14. R. D. Stevens, H. J. Tipney, C. J. Wroe, T. M. Oinn, M. Senger, P. W. Lord, C. A. Goble, A. Brass, M. Tassabehji, Bioinformatics 4 (suppl 1.), I303–I310 (2004).