ELSEVIER

A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources

Dietrich Rebholz-Schuhmann[1,2], Christoph Grabmüller[1], Silvestras Kavaliauskas[1], Samuel Croset[1], Peter Woollard[3], Rolf Backofen[4], Wendy Filsell[5] and Dominic Clark[1]

[1] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[2] Computerlinguistik, Universität Zürich, Binzmühlestrasse 14, 8050 Zürich, Switzerland
[3] GlaxoSmithKline, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, UK
[4] Albert-Ludwigs-University Freiburg, Fahnenbergplatz, D-79085 Freiburg, Germany
[5] Unilever R&D, Colworth Science Park, Sharnbrook MK44 1LQ, UK

In the Semantic Enrichment of the Scientific Literature (SESL) project, researchers from academia and from life science and publishing companies collaborated in a pre-competitive way to integrate and share information for type 2 diabetes mellitus (T2DM) in adults. This case study exposes benefits from semantic interoperability after integrating the scientific literature with biomedical data resources, such as UniProt Knowledgebase (UniProtKB) and the Gene Expression Atlas (GXA). We annotated scientific documents in a standardized way, by applying public terminological resources for diseases and proteins, and other text-mining approaches. Eventually, we compared the genetic causes of T2DM across the data resources to demonstrate the benefits from the SESL triple store. Our solution enables publishers to distribute their content with little overhead into remote data infrastructures, such as into any Virtual Knowledge Broker.

Type 2 diabetes mellitus (T2DM) is a disease with unresolved questions

The genetic causes of diabetes are still not fully understood, although different types of diabetes can be distinguished, including neonatal diabetes (transient and permanent), noninsulin-dependent, maturity-onset diabetes of the young (MODY) and T2DM in adults [1]. Several genes are under investigation for their involvement in the development of this disease [2–5].

In the case of neonatal diabetes, the causes can be found in modifications of the insulin gene [6,7] as well as in other molecular defects (e.g. involving transcriptional and translational factors). For MODY, only six genes account for 80% of the disease development; using selected clinical traits, it is possible to distinguish eight genetic subgroups of MODY [1,8]. By contrast, all loci associated with the risk of diabetes explain no more than 1% of the risk variance and, for most loci, there is a lack of clues about their function in diabetes pathogenesis [9,10]. In addition, only specific phenotypes in diabetes, such as insulin resistance and β cell dysfunctions, indicate heritability [11]. Finally, the genetic parameters only marginally improve the prediction of the disease risk and only if they have been added to the phenotypic factors in the analysis [9].

Furthermore, diabetes is linked to other diseases, such as obesity, which has its own genetic preconditions [12,13]. In addition, the risk for T2DM increases under the genetic predisposition for obesity. In recent years, even patients with type 1 diabetes mellitus (T1DM) have developed obesity, leading to an obscured border between T1DM and T2DM [14]. Therefore, clinical symptoms and genetic criteria have to be reassessed for improved diagnostics and treatments possibly leading to novel drugs [3,14–16].

Trying to determine the causes of T2DM (e.g. insulin resistance in comparison to β cell dysfunction) leads to novel hypotheses for improved disease treatment. For this goal, pharmaceutical companies have to bring together their experts from different disciplines, such as molecular biologists, medicinal chemists and

Corresponding author:. Rebholz-Schuhmann, D. (rebholz@ebi.ac.uk), (d.rebholz.schuhmann@gmail.com)

1359-6446/06/$ - see front matter © 2013 Published by Elsevier Ltd. http://dx.doi.org/10.1016/j.drudis.2013.10.024

Please cite this article in press as: Rebholz-Schuhmann, D. et al., A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources, Drug Discov Today (2013), http://dx.doi.org/10.1016/j.drudis.2013.10.024

toxicologists, to make use of existing data from the public domain and from local repositories and, thus, to improve their productivity. Indeed, the information from the gene to the phenotype must be readily accessible in an interoperable way to explore any complex disease fully [17–19].

However, data repositories are often focused on one type of entity only (e.g. proteins in UniProt Knowledgebase (UniProtKB)) or possibly two (e.g. drugs and their targets in Drugbank) [20,21]. A comprehensive information system for complex biomedical problems would require the integration of facts while, at the same time, considering large numbers of entities, as well as relevant data resources, data providers, or heterogeneous data from the scientific literature, to serve the research community efficiently.

Exploring the genetic causes and the pathogenesis of T2DM requires combining different data resources, such as functional annotations of proteins in UniProtKB as well as gene–disease associations (GDAs) from Online Mendelian in Men (OMIM), which are provided from specific tables or from the scientific literature, respectively [21]. This integration generates unnecessary extra work, since the data resources do not comply with standardized, transparent, or interoperable data formats [22,23]. Above all, facts from scientific manuscripts are still kept in monolithic electronic documents. A few attempts have been made to standardize and enrich such documents, but the facts are not yet delivered as structured data or as linked data into any public repository [24–27]. In particular, the use of data standards for scalable data resources (e.g. semantic web technologies, nanopublications and the Virtual Knowledge Broker) would improve the accessibility of information from the scientific literature significantly [28,29].

Sharing biomedical data with semantic web technologies

Semantic web technologies form a framework for public data exchange and data sharing, and serve as an alternative to proprietary relational databases; strong semantic support is part of the infrastructure. It enables the deliver of evidence from the literature as information pieces into publicly available biomedical data resources [30,31]. Following the 'Linked Data Principles' for the semantic web according to Berners-Lee, the first requirement is the use of universal resource identifiers (URIs) to label things or entities (e.g. for a protein or chemical entity, and also when they appear in the scientific manuscript) [21,32,33].

The next requirement is to combine names with web addresses (i.e. 'http://URIs'), which should lead to useful information in readily available representation standards. For scientific manuscripts, only the metadata of the documents has been exploited up to now, in contrast to the use of the scientific content itself [34–36]. Finally, links to further URIs should be provided to enable discovery (i.e. the entities, or 'things' in the document should be linked through URIs to publicly available biomedical data resources), so-called 'Semantic Enrichment'. Ideally, all entities from the document can be linked to a public data resource and the facts can be verified against the content from biomedical databases.

The Resource Description Framework (RDF) is a semantic web standard for access to public data. Each fact or statement is represented as a triple comprising a subject (e.g. 'P53'), a predicate (e.g. 'is-a') and an object (e.g. 'protein'), and, at best, all three parts are specified uniquely using web-based resources. RDF data triples enable semantic interoperability between resources and provide considerable advantages [37–39] including: (i) consistent reuse of content across distributed resources with well-specified concepts and relations [40,41]; (ii) better error handling through standardized and transparent data representations [28] and (iii) large-scale and seamless exploitation of data through the simplicity and generality of the data representation. Eventually, open standards also enable publishing companies to take part in the data integration and data distribution activities [24,42,43].

Sharing biomedical data in the semantic web

Integrating the literature with public data repositories, such as UniProtKB, requires that its content is structured in a formalized and standardized way: entities (e.g. *p53* gene) and concepts (e.g. transcription regulation activity) from the text have to be referenced through terminologies, ontologies and fact repositories to achieve interoperability [44–47]. For this goal, the Foundry for Open Biomedical Ontologies (OBO) determined principles enabling scientific data resources to communicate with minimum uncertainty (e.g. without ambiguity): for example, the Human Phenotype Ontology (HPO) uses cross-references to available ontologies, such as Gene Ontology (GO), Chemical Entities of Biological Interest (ChEBI), Foundational Model of Anatomy (FMA) and so on, to define entities logically [33,48–50].

The data repositories should link their data entries in a readable form to relevant information for interactive use [51,52]. In the biomedical domain, this has been achieved by assigning metadata information to experimental data, thus improving information retrieval: for example, transforming table data into a representation using triples integration of health-related data (e.g. in Chinese medicine or for the modelling of neurological receptors) [41,53–57]. For translational medicine in Alzheimer's disease, a fact repository of 350 million triples have been built using ontologies and their domain knowledge (in OWL) in a structured way using well-defined concepts [58,59]. Similarly, selected pathway repositories have been integrated, including Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome and BioCyc, together with EntrezGene [60–65].

Current solutions integrating the literature with semantic web technologies make use of metadata only, once it has been provided by the author or extracted from the textual content [36,66,67]. The existing solutions demonstrate the potential of semantic web technologies for the integration of data and services. By contrast, the proposed semantic web solutions do not sufficiently integrate facts from the scientific literature or demonstrate the requirements necessary for this goal.

Decomposing the biomedical scientific literature

Members from academia, life science and publishing companies have worked together in the Semantic Enrichment of the Scientific Literature (SESL) project to integrate public and proprietary data using semantic web technologies. The project has produced technical achievements in biomedical semantics in addition to explaining the wider perspective [68].

In total 638,088 scientific publications were contributed by the publishing companies involved (Elsevier, Nature Publishing Group, Oxford University Press and Royal Society of Chemistry). A further

Please cite this article in press as: Rebholz-Schuhmann, D. et al., A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources, Drug Discov Today (2013), http://dx.doi.org/10.1016/j.drudis.2013.10.024