A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data.

Post, L.J.G.; Roos, M.; Marshall, M.S.; van Driel, R.; Breit, T.M.

DOI
10.1093/bioinformatics/btm461

Publication date
2007

Published in
Bioinformatics

BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 22 2007, pages 3080–3087
doi:10.1093/bioinformatics/btm461

Databases and ontologies

A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data

Lennart J.G. Post[1,2], Marco Roos[1], M. Scott Marshall[1], Roel van Driel[2] and Timo M. Breit[1,*]

[1]Integrative Bioinformatics Unit and [2]Nuclear Organization Group, Swammerdam Institute for Life Sciences, University of Amsterdam, 1098 SM, Amsterdam, The Netherlands

Received on May 8, 2007; revised on August 10, 2007; accepted on September 5, 2007

Advance Access publication September 19, 2007

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: The numerous public data resources make integrative bioinformatics experimentation increasingly important in life sciences research. However, it is severely hampered by the way the data and information are made available. The semantic web approach enhances data exchange and integration by providing standardized formats such as RDF, RDF Schema (RDFS) and OWL, to achieve a formalized computational environment. Our semantic web-enabled data integration (SWEDI) approach aims to formalize biological domains by capturing the knowledge in semantic models using ontologies as controlled vocabularies. The strategy is to build a collection of relatively small but specific knowledge and data models, which together form a 'personal semantic framework'. This can be linked to external large, general knowledge and data models. In this way, the involved scientists are familiar with the concepts and associated relationships in their models and can create semantic queries using their own terms. We studied the applicability of our SWEDI approach in the context of a biological use case by integrating genomics data sets for histone modification and transcription factor binding sites.

Results: We constructed four OWL knowledge models, two RDFS data models, transformed and mapped relevant data to the data models, linked the data models to knowledge models using linkage statements, and ran semantic queries. Our biological use case demonstrates the relevance of these kinds of integrative bioinformatics experiments. Our findings show high startup costs for the SWEDI approach, but straightforward extension with similar data.

Availability: Software, models and data sets, http://www.integrativebioinformatics.nl/swedi/index.html

Contact: breit@science.uva.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

With the development of high-throughput biotechniques and the subsequent omics studies, exciting avenues of scientific exploration are opening up. Instead of being constrained to analyze a handful of genes or proteins per experiment, whole genomes and proteomes can be studied today. This allows biologists to investigate more complex processes that were not accessible before (Carroll *et al*., 2006; Lein *et al*., 2007; Souchelnytskyi, 2005; Spellman *et al*., 1998; van Steensel, 2005).

As became evident from the human genome project, once the technology limitations were lifted, the bottleneck rapidly shifted to the annotation of the produced DNA sequence data. Therefore, like the biotechniques, huge projects with numerous research groups collaborate to tackle complex issues such as annotating the human genome (The ENCODE Project Consortium, 2004). So on top of the omics data a growing layer of biological annotations is being produced. These data are made increasingly available through public web-accessible data stores like Ensembl and the UCSC Genome Browser. Because the data is distributed across the web, this raises new issues on data management, maintenance and usage. Biologists use these data as reference, but increasingly also for *in silico* data integration experiments. Integrating these heterogeneous data sets across different databases, however, is technically quite challenging, because one must find a way to extract information from a variety of search interfaces, web pages and APIs. To complicate matters, some databases periodically change their export formats, effectively breaking the tools that provide access to their data. At the same time, most omics databases do not yet provide computer-readable metadata and, when they do, it is not in a standard format. Hence, expert domain-specific knowledge from the user is required to interpret what the data actually represents before using it in integration experiments. This limits the practical scale and breadth of integration, given the variety and amount of data available from distributed resources.

The Semantic Web is designed to bring meaning to the raw data content by defining relationships between distinct concepts (http://www.w3.org/2001/sw/) using ontologies. This allows the sharing and processing of data by automated agents that can assist in the retrieval of relevant information and metadata (Roos *et al*., 2004). The Resource Description Framework (RDF) specification is a metadata model that forms the basis of the Semantic Web. The metadata model describes everything as a resource that can be linked to other resources by defining

*To whom correspondence should be addressed.

relationships as properties. Resources are described by making statements identifying the resource (the subject), its property (the predicate) and the value of the property (the object), e.g. MAPKAP-2, hasFunction, Kinase. The statements used in RDF are defined in RDF Schema (RDFS). RDFS describes the semantics and defines class and property hierarchies of the domain for which the RDF document is used.

To allow machine reasoning over the formalized knowledge of a domain, the W3C has developed a standard for a web based ontology language: OWL (http://www.w3.org/2004/OWL/), a language that builds upon RDF and RDFS. Essentially, an ontology is a formalization of a domain, defining concepts (i.e. collections of biological elements that share common properties) and the relationships between them, thus creating a common, controlled vocabulary that can be reasoned over in a well-defined manner. Applying ontologies to data involves populating the concepts with individuals, i.e. real-life entities. By defining ontologies for a field as complex as biology, one can eventually build a knowledge base that facilitates the exchange and interoperability of the data present in numerous available databases. Many biological ontology initiatives exist (http://obo.sourceforge.net/), with the Gene Ontology (GO) being the most widely adopted (Ashburner *et al.*, 2000). This allows these databases to transcend from data stores to knowledge stores. Thus, ontologies will greatly aid biological research by providing a structured approach to capturing knowledge in a computer-understandable way (Bodenreider and Stevens, 2006; Good and Wilkinson, 2006; Ruttenberg *et al.*, 2007; Strizh, 2006).

Because of the heterogeneity of life sciences data, the semantic web approach could be useful throughout the entire cycle of integrative bioinformatics experimentation. Figure 1 shows this cycle divided into five phases: problem definition, experimental design, data integration, data analysis and data interpretation. In the current study, we have applied the semantic web approach to the data integration phase of an example integrative bioinformatics experiment and evaluated its applicability in the context of the whole cycle. As biological use case, we set out to combine two genomics data sets from UCSC: data about a specific histone modification and data about transcription factor binding sites. Our approach using semantic web-enabled data integration (SWEDI) is based on semantic web technology for a model-based integration of data sets in the life sciences domain (Marshall *et al.*, 2006). We constructed three OWL biological knowledge models, one OWL technical knowledge model and two RDFS data models. We then transformed and mapped relevant data to the data models, linked the data models to the knowledge models using linkage statements and ran a semantic query. The analysis of the results of the biological use case demonstrated the relevance of these kinds of integrative bioinformatics experiments. Our findings are that the initial ‘startup’ costs for SWEDI are high, but that subsequent addition of (similar) data is straightforward.

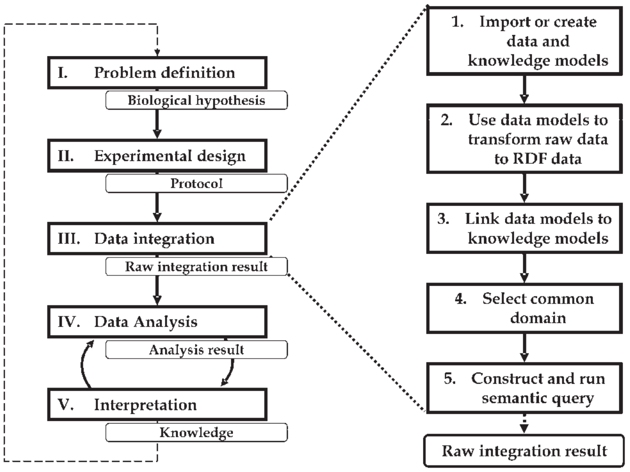

Fig. 1. Integrative bioinformatics experimentation cycle with five distinct phases. At the end of a phase, an outcome is generated that is input for the next phase. Our semantic data integration approach is applied to phase II, which can be further divided into five steps.

2 METHODS

2.1 Data

All data sets were downloaded from the UCSC genome browser website (http://genome.ucsc.edu/). We used H3K4me3 data from the ChIP-on-chip data set produced by the Sanger Institute (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=88704835&g=encodeSangerChip) for five human cell lines; GM06990, HeLaS3, HFL-1, K562 and MOLT-4. Each data set contains locations on the human genome where H3K4me3 is present plus the intensity score, which is an indication for the amount of H3K4me3 at that position (ftp://ftp.sanger.ac.uk/pub/encode/H3K4me3_GM06990_2/README) The Sanger data set is a ENCODE region-wide H3K4me3 analysis which comprises ~1% of the total human genome. These motif sequences are usually represented by position weight matrices of conserved TFBS types (cTFt). With these cTFt matrices, highly similar locations in the genome can be identified that are potentially conserved TFBS (cTFBS) for each cTFt. The cTFBS data was generated by UCSC using 410 binding matrices from the Transfac Matrix and Factor database v8.3 from Biobase (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=88704835&g=tfbsConsSites). In essence, the track gives information about cTFt and the predicted occurrence of associated cTFBS in the whole human genome. The Hidden Markov Model (HMM) data identifies hit regions in the Sanger data set using a two-state HMM analysis as performed by EBI (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=86022194&g=encodeSangerChipHits). We used HMM data for the three available human cell lines; GM06990, HeLaS3 and K562.

2.2 Semantic web technology

We created the data models, knowledge models and linkage statement files using Protégé 3.1.1 with OWL plug-in V2.1 (http://protege.stanford.edu/). To visualize the data sets within the knowledge models, we used Protégé to create individuals for each data set within the corresponding concept. To transform the tab-delimited data to RDF/XML data we used a version of Mapper (https://gforge.vl-e.nl/projects/mapper) that we modified for RDF output. Transformed data was loaded in Sesame v1.2.6 (http://www.openrdf.org/). Subsequently, SeRQL queries (Supplementary Material, Fig. S1) were constructed to find cTFBS that overlapped with H3K4me3 regions. The Sesame program was run on a server station with two Intel Xeon processors at 2.8 GHz equipped with 4 GB main memory.

3 RESULTS

We set out to analyze the applicability of our SWEDI approach for a specific phase of the integrative bioinformatics experimentation cycle by means of a biological use case. Although we only want to analyze one specific phase of this experimentation cycle, its indissoluble nature forces us to perform an experiment including all phases in order to determine the usability of SWEDI for life sciences research. We identified five phases in our experimental cycle: problem definition, experimental design, data integration, data analysis and interpretation (Fig. 1). At the end of each phase an outcome is generated that serves as input for the next phase. Our study focuses mainly on the application of SWEDI in the data integration phase.

3.1 Problem definition

We started by defining the biological hypothesis of our use case with the input from domain experts i.e. biologists. Figure 2 shows a cartoon representation of the hypothesis. A puzzling phenomenon in biology pertains to histone modifications. DNA is bound by histone octamers, called nucleosomes that package it inside the nucleus. Nucleosomes are built of eight histones, which can undergo post-translational chemical modifications at their N-terminal tail (Felsenfeld and Groudine, 2003). Different histone marks are associated with different cellular processes (Peterson and Laniel, 2004). It is believed that these histone marks act in concert to form a ‘histone code’ that defines the transcriptional state of the chromatin (Strahl and Allis, 2000). For instance, the presence of three methyl groups to the fourth amino acid, a lysine (K), of histone 3, named H3K4me3 (Turner, 2005) is believed to be a histone mark for active gene transcription (Schneider *et al.*, 2004).

Transcription factors regulate gene expression by (in)direct binding to specific regions in the genome. In its simplest view, one transcription factor with a DNA binding domain, recognizes and binds a specific DNA sequence upstream of a gene (i.e. transcription factor binding site, TFBS) to alter its transcriptional state (Fig. 2). For many transcription factors the associated TFBS sequence motifs they recognize have been identified (Matys *et al.*, 2003).

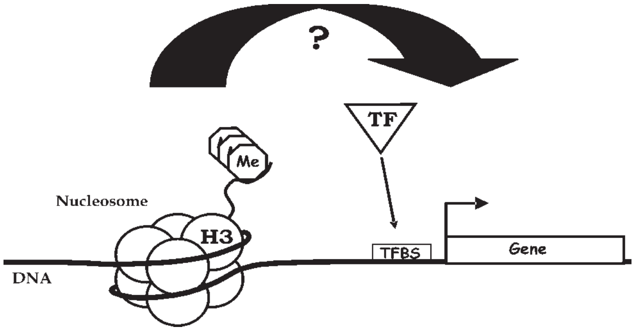

Fig. 2. Cartoon representation of the hypothesis of our biological use case. Nucleosomes are bound to DNA and consist of eight histones. One, histone 3 (H3), has a lysine on the fourth position of its N-terminal tail (K4), which can accept 3 methyl (me) groups and this modification H3K4me3 was shown to be a mark for active gene transcription. Generally, gene expression (arrow on gene) is regulated by transcription factors (TF) that recognize and bind to specific DNA sequences upstream of genes, i.e. transcription factor binding sites (TFBS). Our hypothesis predicts a relationship between histone modification H3K4me3 and gene expression regulation through TFBSs.

Because H3K4me3 is a histone mark for active gene transcription, we formulated a biological hypothesis that postulates a direct relationship between the presence of this histone mark and specific cTFBS or cTFt (Heintzman *et al.*, 2007). Although in essence this relationship is known in biology, it is a nice hypothesis to test our approach and possibly further interpret this relationship.

3.2 Experimental design

For this hypothesis, we identified relevant data sources about histone modification H3K4me3 and TFBS at the UCSC Genome Browser site, which stores various genome annotations concerning a number of different species including human. We used a data track about cTFt conserved among human, mouse and rat plus data tracks from the ENCODE project about ChIP-on-chip intensity scores of human H3K4me3. The ENCODE-H3K4me3 data tracks holds data of ~1% of the whole genome. We chose human cell line GM06990 from the Sanger Institute track together with the cTFBS data track from UCSC for SWEDI. For proof-of-principle, we decided to start with only one human cell line together with the cTFBS data track. Subsequently, we added similar data from four additional human cell lines, plus HMM-analyzed H3K4me3 data from three human cell lines to show extendibility.

In order to discover a relationship between the level of H3K4me3 modification and specific cTFBS or cTFt, we devised SWEDI, a model-based data integration approach, to integrate these genomics data sets. For this, we decided to model the domains that cover the minimal relevant biological and technical features associated with the data sets in a way that would allow future extension. This means several small models rather than one big, all-inclusive model. We also decided to use RDFS data models that capture the low-level metadata related to the data, to link the RDF data to our knowledge models in OWL.

3.3 Data integration

This is the actual phase in which we want to apply SWEDI. In our case we subdivided this phase into five steps: import or create models, transform raw data to RDF data, link data models to knowledge models, select common domain and construct and run semantic query (Fig. 1). Given the complexity of this phase, we will explain each consecutive step separately.

3.3.1 Importing or creating models The initial step in SWEDI is translating the hypothesis into a formalized, composite knowledge model that captures the domain-specific concepts and their relationships. This is done either by using an existing domain-specific ontology or creating a new ontology. Although an ontology like GO captures some of the desired concepts, as a knowledge model it is too restricted due to the limited types of relationships, only isA and partOf.

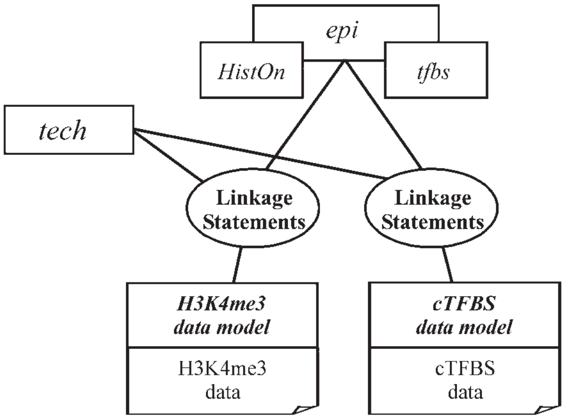

Fig. 3. Schematic overview of the OWL knowledge models, *epi* (epigenetics), *HistOn* (histone), *tfbs* (TFBS) and *tech* (technical); the RDFS data models, *H3K4me3* and *cTFBS*; and how they are linked.

Also, the level of granularity in the area of histones is limited. Since we could not find any suitable ontology for our use case, we constructed our own ontologies (Supplementary Material, Fig. S2).

Because our approach is meant to be scalable by allowing addition of more models and data, we purposely created distinct models that capture different aspects of the involved biology domain. Figure 3 shows our four OWL ontologies to model the domains that cover the data sets: an epigenetics model (*epi*), a histone model (*HistOn*), a TFBS model (*tfbs*) and a technical model (*tech*). The knowledge models were created with help of experts in the field of nuclear organization and peer-reviewed literature. We chose to express the knowledge models using OWL-DL in order to have enough expressiveness but still remain computationally efficient (http://www.w3.org/TR/owl-features/).

(i) *epi*, the largest model, was constructed to capture general concepts in the biology domain of histone modification from an epigenetics viewpoint. Epigenetics is about (heritable) DNA-related features other than the actual DNA sequence, such as DNA methylation, histone modification and chromatin structure. The model *epi* contains 78 concepts and 16 properties. Concepts range from amino acid modifications, sequences and chromosomes.

(ii) *HistOn*, the histone model, is more specific and covers histones and histone-related concepts like histone modifying proteins. It contains 17 concepts plus 19 properties and is nested within *epi*.

(iii) *tfbs*, the TFBS model covers TF, TFBS and related concepts like promoters, enhancers and repressors. It contains 14 concepts and 6 properties. *tfbs* is nested within *epi*.

(iv) *tech*, the technical model was created to cover abstract terms in the data sets such as experimental measurement scores and calculated *z*-scores. It contains seven concepts and four properties.

Together, these knowledge models capture all concepts essential for our use case, but they can be further expanded as needed.

We also constructed two RDFS data models that semantically capture the data sets we want to integrate. We based our data models on the database table schema of the data sets. The data models semantics is limited to describing the data file rows and columns.

(i) The *H3K4me3 data model* describes the H3K4me3 data track on UCSC and contains two objects and six properties.

(ii) The *cTFBS data* model describes the cTFBS data track on UCSC and contains two objects and nine properties.

Although we constructed the data models using Protégé/OWL, for our data models the expressiveness of RDFS is sufficient and computationally less expensive.

3.3.2 Transform raw data into RDF format We retrieved cTFBS and H3K4me3 data sets of human cell line GM06990 from UCSC and used an adapted version of Mapper with the associated data model to transform the tab-delimited flat file to RDF/XML. We chose to keep the data separated from the ontology so that we can describe the data using 'simple' RDF/XML.

A drawback of expressing data in RDF/XML is bloating of data size on disk. An approximate 15-fold increase in file size was observed when the cTFBS tab-delimited data set was transformed to RDF/XML and an approximate 18-fold increase for the H3K4me3 data sets. Although once loaded into memory, the XML bloat is no longer a problem, file storage on disk and file exchange are issues that will require attention.

3.3.3 Link data model to knowledge model The next step is to link the models that capture biological knowledge to the data models. Figure 4 shows how we started by populating the knowledge model with individuals representing each data set. We then linked the data models to the knowledge models by linking properties. Using the inference feature of RDFS, we declared that a property of the data model, for instance *chrom*, is a sub-property of the property *Chromosome_identifier* from the knowledge model. For each data set this resulted in two collections (i.e. files) that contain all linkage statements (Fig. 3). As such, the linkage statements function as a user-defined viewpoint of how a data model is related to the knowledge models, which allows defining queries in the more familiar terms of knowledge models.

We chose to keep our data independent of the knowledge models, with an explicit mapping in the form of the linking statements. This approach to linking also preserves the data supplier's naming scheme. We could have directly transformed the raw data files into RDF that includes our own OWL terms directly in the RDF version of the data. However, such an approach would shift control of linking to the import stage and subsequent changes to our knowledge models could require an entire new import process for any affected data to correct obsolete links embedded in the data.

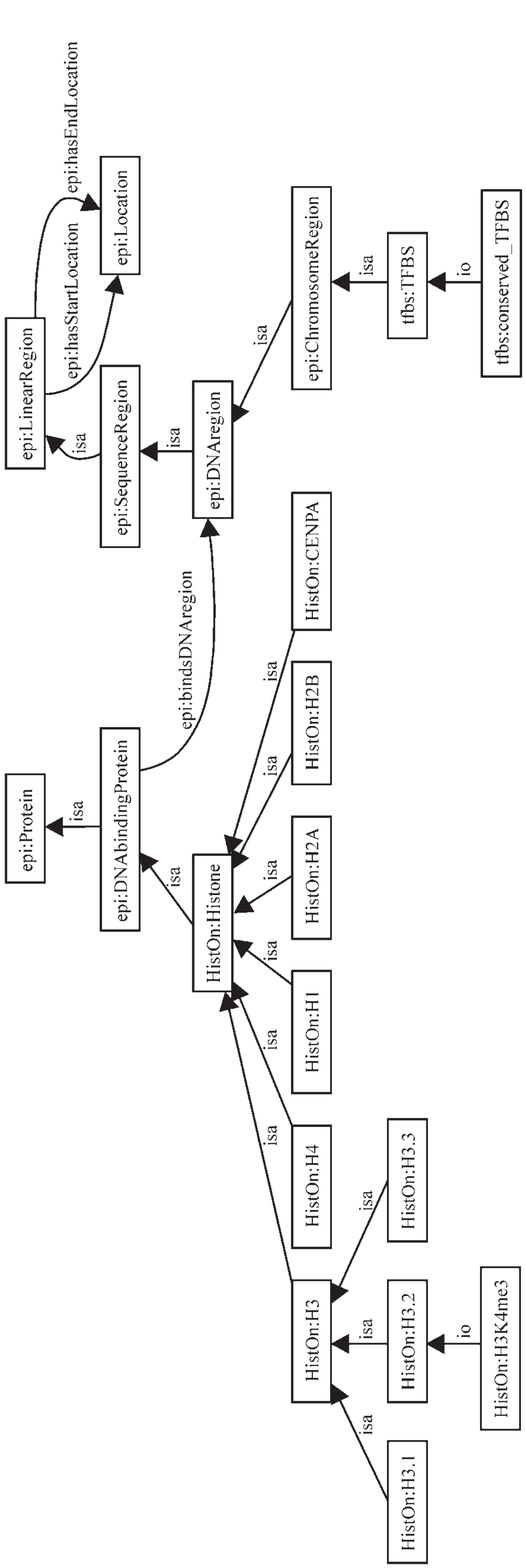

Fig. 4. Schematic overview on how the data is linked. Conserved transcription factor binding site (cTFBS) data and Histone 3 Lysine 4 tri-methylation (H3K4me3) data are integrated using our semantic models. The subsection of the model shows a path between the two data sets. The concept *ChromosomeRegion* is selected as the common domain to base the overlap query on.

3.3.4 Select common domain To determine the relationship between H3K4me3 and cTFBS or cTFt, we had to find a path between these concepts in our models and identify a domain for comparison. By selecting a common domain we could integrate these data sets. In our case, we chose *ChromosomeRegion* as the common domain, because both histone modification and cTFBS data have genomic positions coordinates in the form of chromosome number, start and end position.

3.3.5 Construct and run semantic query With a common domain identified, we created a query to test the relationship under question: 'Which DNA regions are bound by a H3K4me3 modified histone as well as a cTFt?' We constructed a SeRQL query (Supplementary Material, Fig. S1) that checks if H3K4me3 and cTFBS DNA regions are on the same chromosome within each other's start and stop positions. It states that two regions from each data set overlap when there is at least one base pair with a direct overlap. Overlap means identification of a cTFBS for each cTFt.

Initially, the query took around 45 h (wall time) to complete and returned 12 349 overlaps for GM06990 (Supplementary Material, Fig. S3). Restricting the whole-genome cTFBS data set (1 077 457) to cTFBS that are within an ENCODE region (13 779), dramatically reduced the query run time to ~30 min.

The raw integration results in fact are the proof-of-principle of SWEDI because it accomplishes data integration of heterogeneous data sets by means of semantic web technology. An important difference between the use of a traditional database and semantic web repositories is that the (meta)data model for the semantic web approach is described in a standardized language. Where RDFS and OWL are used, reasoning can be applied to the model.

3.4 Extension of data integration experiment

After achieving this proof-of-principle using data from just one cell line (GM06990), we evaluated the extensibility of SWEDI, since that is its main motivation. For this, we extended the experimental design with data from four additional human cell lines; HeLa, HFL-1, K562 and Molt4. By simply changing the identity of the histone modification data set and slightly adapting the SeRQL queries we were quickly able to achieve raw H3K4me3–cTFBS integration results for these cell lines; 13 350 overlaps for HeLa, 13 350 for HFL-1, 13 315 for K562 and 13 341 for Molt4.

We further extended our analysis with additional UCSC data: HMM-analyzed H3K4me3 data from cell lines: GM06990, HeLa and K562. The integration steps for these three data sets remained the same, with similar exception for the query step. The results were; 3289 overlaps for GM06990, 2134 for HeLa and 3273 for K562. Because HMM-analyzed data sets are much smaller than nonanalyzed data, the query took ~1.5 h (wall time) to complete for the whole genome cTFBS data set and ~2 min for cTFBS that are within an ENCODE region.

The UCSC Genome Browser contains an extensive number of genome annotation tracks that can potentially be integrated using our approach. There are 13 tracks containing 97 sub tracks that are almost identical to the data sets we used for our use case. These can be integrated almost directly, if we use the H3K4me3 data model to transform the tab-delimited data sets to RDF data. Also, the *HistOn* and *tfbs* knowledge models need to be populated with the new data sets and the SeRQL query needs to be changed to cover the new data sets. There are also 12 tracks containing 37 sub tracks that can be integrated requiring only minor concept additions to the tech model and/or a new data model in addition to the changes mentioned above (Supplementary Material, Table S1).

3.5 Data analysis and interpretation

Through SWEDI we have coupled H3K4me3 intensity scores to all cTFBS of each cTFt and we obtained raw integration results (Supplementary Material, Fig. S4). These results showed that the majority of cTFBS displayed a low H3K4me3 score. Applying a H3K4me3 score cutoff of >2 resulted in: 1382 overlaps for GM06990, 984 for HeLa, 1063 for HFL-1, 1303 for K562 and 739 for Molt4. An in-depth analysis and interpretation is beyond the scope of this article. A brief preliminary analysis and interpretation of the result can be found in Supplementary Material, Figure S5.

The analysis of the raw integration results from SWEDI on the HMM-analyzed H3K4me3 data from cell lines GM06990, HeLa and K562 showed essentially the same outcome as compared to the original H3K4me3 data (data not shown).

4 DISCUSSION

Along with the introduction of omic technologies in the life sciences came the need to handle, analyze and interpret biological data in a different, more formalized way. The semantic web approach enhances data exchange and integration by providing standardized formats such as RDF, RDFS and OWL, to achieve a formalized computational environment. In this study, we have investigated the potential of the semantic web concept for the purpose of data integration by computational experimentation in the life sciences. We focused on basic linkage of data to knowledge using semantic web based data and knowledge models.

Because of the complexity of biology, we adopted a strategy to formalize a (part of a) domain that is of interest to a specific (group of) scientist(s) by capturing the knowledge via a network of interrelated semantic models using ontologies as a controlled vocabulary. This allows a modular approach for data integration in which the individual scientists can use existing (general) models, potentially in combination with small specific models that they create themselves. This means that each scientist can interact with external data and knowledge models from their own perspective using a kind of ‘personal semantic framework’. In this way the involved scientists are familiar with the concepts and the relationships in the models they work with and can create semantic queries using their own terms. The external models should be either made by coordinating data-managing organizations (e.g. NCBI), or organized domain experts (e.g. FlyBase consortium).

After creating several necessary knowledge and data models, we were able to provide a proof-of-principle for our SWEDI approach. Multiple genomics data sets, which involved histone modification and TFBS, were successfully linked via a common domain. The integrated data in our biological use case resulted in some interesting biological observations that may lead to new hypotheses regarding the role of histone modification in gene expression regulation. With our approach, we established a type of formalization of the problem domain by creating a vocabulary in the form of knowledge models that describe the data and capture the domain knowledge. This promoted the transparency and reproducibility, as well as the easy extensibility of our experiments. However, more sophisticated tools (for example the OntoViz plug-in in Protégé) are needed for visualization of concepts and their relationships when more and more data sets are added. Also, our approach gives us flexibility in asking questions to the data sets. Although sites like UCSC Genome Browser have tremendous amounts of data and information, it is rather difficult to ask simple questions like: what is the overlap between any number of tracks concerning histone modifications, transcription factors and genes.

Although in this study we used a rather limited use case, we still could show that SWEDI is extendable by starting from just one cell line (GM06990) and adding similar H3K4me3 modification data sets from four additional human cell lines and three H3K4me3 recalculated data sets. The use of new data from the same site (UCSC Genome Browser) and from the same track, resulted in the re-use of the H3K4me3 data model, because this model describes the data on a very low level and the format is identical. Since the data model can be re-used, the linking between the data and knowledge model remains the same, as does the common domain. In contrast, the SeRQL queries do have to be altered slightly. The addition of extra data sets was facilitated by the great similarity between the data sets, but in essence any data set that contains at least a chromosome number, start position and end position can be integrated. Data models need to be created, if they cannot be re-used. Although we showed extensibility by adding similar data sets, it will be a challenge to add totally different data, e.g. data not related to genomic location.

There are also drawbacks to SWEDI. The main problem is that the initial setting-up costs are high, because there are hardly any adequate knowledge models available yet. This problem is inherent to formalizing a domain, but the applied semantic web technologies are still immature, so RDF data set manipulation is hard. Also, a future problem could rise when many, highly divergent personalized frameworks eventually have to be merged. This relates to the more general problem of ontology alignment (Euzenat and Valtchev, 2003). So it seems fair to assume that if any domain in biology wants to embrace this approach, it will take a community as well as a multidisciplinary effort to make and maintain something like a domain semantic framework. The effort can be compared to those from initiative such as NCBI, UCSC Genome Browser, FlyBase, WormBase, etc. An example in this context is the effort of the W3C-HCLS (Ruttenberg *et al.*, 2007) to recommend a standard scheme for the URIs that refer to commonly used bioconcepts. If widely adopted, such a URI scheme would have a normalizing effect that would greatly increase the ease of data integration and model sharing. This would be a first step in the direction of a domain semantic framework. Although the necessary consensus for a domain semantic framework would be a challenge to establish, it would only need to happen once. In contrast, consider the countless times that database schemas are re-invented for use with the same data but at different institutions. The scalability of SWEDI is also a matter of concern, as performance and provenance may become bottlenecks. With the sizes of our whole-genome data sets the queries took ~2 days to run, because the query engine is not optimized for our type of semantic web based query yet. As noted earlier, query optimizations can greatly improve the performance (Marshall *et al.*, 2006). Although we could have used SQL for better query performance, we would have to give up all the benefits of explicit use of our models in the query itself. Furthermore, in SWEDI data models are mapped to knowledge models using subPropertyOf statements that are stored in the data models. This poses a problem because in our scenario, we do not control these data models and so we cannot add mapping statements directly to these models. We therefore, had to store the linking statements in a separate file, thus keeping the data models intact, but adding an extra layer. Finally, choosing the common domain is done manually, which demands extensive domain knowledge. It would be extremely useful if methods could be developed that identify common domains automatically.

In the context of biological data and knowledge integration, numerous solutions have been developed to enable retrieval of data from heterogeneous distributed sources (Eckman *et al.*, 2003; Searls, 2005; Stein, 2003). The solutions range from monolithic, such as, SRS (Zdobnov *et al.*, 2002) with keyword indexes and hyperlinks, Kleisli/K2 (Ritter *et al.*, 1994) with a query language across databases as if they were one, data warehouses such as BioZon (Birkland and Yona, 2006), automated annotation systems such as PhosphaBase-$^{\text{my}}$Grid (Wolstencroft *et al.*, 2006), to BioMOBY which uses web services acting as portals to biological data (Wilkinson *et al.*, 2005). Perhaps the most widely used system is SRS, providing integration of over 400 databases. Our approach to data integration uses semantic models to provide a schema for integration. TAMBIS pioneered such an approach by creating a molecular-biology ontology as a global schema for transparent access to a number of sources including Swiss-Prot and Blast (Stevens *et al.*, 2000). Systems such as BACIIS (Miled *et al.*, 2003), BioMediator (Mork *et al.*, 2005) and INDUS (Caragea *et al.*, 2005) extend on this example. For instance, BioMediator uses a 'source knowledge base' that represents a 'semantic web' of sources linked by typed objects. The knowledge base includes a 'mediated schema' that can represent a user's domain of discourse. INDUS shows important similarities to our approach, offering an integrated user interface to import or create user ontologies, and creating ontological mappings between concepts and 'ontology-extended' data sources. In contrast to our approach, however, INDUS does not use semantic web formats such as OWL and RDF. While the syntactic step of our import is similar to that of YeastHub (Cheung *et al.*, 2005), our explicit linking of the semantic types to the syntactic types with RDFS moves the work of discovering semantics from the query to the model alignment stage. A different approach using Semantic Web methodologies to integrate gene data with phenotype data uses RDF graph analysis to prioritize candidate disease genes (Gudivada *et al.*, 2007).

With respect to the applicability of SWEDI in biology there is no doubt that semantic modeling is a necessity for biological knowledge bases (Ruttenberg *et al.*, 2007). Life sciences research today is all about data, information and knowledge management. Whereas previously the important domain information and knowledge resided mainly in the head of the responsible principal investigator and literature, we are moving into the era of data warehouses/repositories, information management systems and knowledge bases. For any life sciences research group, this means that they have to deal with these e-science issues if they want to stay competitive (Goble *et al.*, 2005; Rauwerda *et al.*, 2006). Furthermore, merely managing all resources will not help much. Once all resources are accessible, multidisciplinary skills such as data mining, data integration, data analysis, statistics, etc are needed. With SWEDI we advanced towards formalized resource management by semantic models. Even our limited SWEDI approach can be used to integrate a substantial number of data sets. However, at present SWEDI only covered data integration and not data analysis or interpretation. This means that with SWEDI we succeeded to integrate multiple genomics data sets. As always, once the technical bottleneck for data integration was lifted, it shifted immediately to data analysis and interpretation. These phases are not covered by SWEDI yet, and it takes quite an effort to extract relevant biological knowledge from raw integration results, as we experienced in our use case. However, the data and findings of any integrative bioinformatics experiment using SWEDI are by definition in a standardized format. This facilitates putting them in semantic web repositories, which subsequently increases their re-use by other members of the research community.

SWEDI is based on building OWL models confined within the scope of an experiment (Marshall *et al.*, 2006). OWL enables the linking of small models to form a larger semantic web, hence a 'bottom up' approach. This ensures freedom for scientists to compose and extend models to their specific needs, such as, new (hypothetical) concepts that have yet to reach the level of consensus necessary for consortia-managed ontologies. In contrast, OBO models are being built 'top-down' by consortia to encompass an extensive number of concepts. Integration could be enhanced through the use of upper-bio-ontologies (Grenon *et al.*, 2004; Rector and Rogers, 2004; Schulz *et al.*, 2006) and the anticipated problems of merging many and divergent personalized frameworks can be circumvented by linking personalized frameworks via upper-bio-ontologies. Upper-bio-ontologies can also facilitate the introduction of new domains to the created formalized computational environment. These upper ontologies increasingly offer guidance on how to categorize new concepts. Careful use of them, and best practices should simplify the alignment of semantic models.

Altogether our SWEDI approach is a first step towards a formalized computational environment for integrative bioinformatics experimentation. The modular nature of SWEDI in combination with the use of standardized semantic web formats ensures the extendibility and scalability of the approach. SWEDI can either be used to create bottom-up small personal semantic frameworks or (top–down) larger domain semantic frameworks. An important advantage of the use of semantic web repositories compared to relational databases is that their complete (meta)data models (i.e. data schemas) are described in a standardized language, which enhances transparency because they can be visualized and manipulated by nonproprietary tools. These schemas are also referenced by the query so that it is possible to examine any RDF query to discover precisely what it means, i.e. track data provenance. Where RDFS and OWL constructions are used, the corresponding reasoning can be applied to the data schemas, creating opportunities for innovative integrative bioinformatics experimentation. Furthermore, semantic web repositories can flexibly be accessed and modified, because many types of modifications require neither specialized knowledge about repository internals nor risky processes such as table migration. Finally, if upper ontologies and best practices are carefully applied in the smaller personal semantic frameworks, it will be possible to link them together into a functional semantic web.

ACKNOWLEDGEMENTS

We thank W. van Hage for assistance with RDFS features of Sesame, M. Jansen for helpful advice and R. Lingeman for statistical support. This work was carried out in the context of: the Virtual Laboratory e-Science project (www.vl-e.nl) supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and the ICT innovation program of the Ministry of Economic Affairs (EZ); and BioRange program of the Netherlands Bioinformatics Centre (NBIC) supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Birkland,A. and Yona,G. (2006) BIOZON: a hub of heterogeneous biological data. *Nucleic Acids Res.*, **34**, D235–D242.

Bodenreider,O. and Stevens,R. (2006) Bio-ontologies: current trends and future directions. *Brief Bioinform.*, **7**, 256–274.

Caragea,D. *et al.* (2005) Algorithms and software for collaborative discovery from autonomous, semantically heterogeneous, distributed information sources. *Discov. Sci. Proc.*, **3735**, 14.

Carroll,J.S. *et al.* (2006) Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.*, **38**, 1289–1297.

Cheung,K.H. *et al.* (2005) YeastHub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics*, **21** (**Suppl. 1**), i85–i96.

Eckman,B. *et al.* (2003) Data management in molecular and cell biology: vision and recommendations. *Omics*, **7**, 93–97.

Euzenat,J. and Valtchev,P. (2003) An integrative proximity measure for ontology alignment. In: Doan,A. *et al.* (eds.) *Proceedings of the Semantic Integration Workshop at ISWC-03*. CEUR-WS.

Felsenfeld,G. and Groudine,M. (2003) Controlling the double helix. *Nature*, **421**, 448–453.

Goble,C. *et al.* (2005) The semantic web and knowledge grids. *Drug Discov. Today: Technol.*, **2**, 225–233.

Good,B.M. and Wilkinson,M.D. (2006) The life sciences semantic web is full of creeps! *Brief Bioinform.*, **7**, 275–286.

Grenon,P. *et al.* (2004) Biodynamic ontology: applying BFO in the biomedical domain. *Stud. Health Technol. Inform.*, **102**, 20–38.

Gudivada,R.C. *et al.* (2007) A genome – phenome integrated approach for mining disease-causal genes using semantic web. *Health Care and Life Sciences Data Integration for the Semantic Web, Sixteenth International World Wide Web Conference (WWW2007) Workshops*.

Heintzman,N.D. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.

Lein,E.S. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.

Marshall,M.S. *et al.* (2006) Using semantic web tools to integrate experimental measurement data on our own terms. In Meersman,R. *et al.* (eds) *Workshop on Knowledge Systems in Bioinformatics (KSinBIT'06)*. Springer Verlag LNCS 4277, pp. 679–688.

Matys,V. *et al.* (2003) TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Miled,Z.B. *et al.* (2003) An efficient implementation of a drug candidate database. *J. Chem. Inf. Comput. Sci.*, **43**, 25–35.

Mork,P. *et al.* (2005) The multiple roles of ontologies in the biomediator data integration system. *Data Integration Life Sci. Proc.*, **3615**, 96–104.

Peterson,C.L. and Laniel,M.-A. (2004) Histones and histone modifications. *Curr. Biol.*, **14**, R546–R551.

Rauwerda,H. *et al.* (2006) The promise of a virtual lab in drug discovery. *Drug Discov. Today*, **11**, 228–236.

Rector,A.L. and Rogers,J. (2004) Patterns, properties and minimizing commitment: reconstruction of the GALEN upper ontology in OWL. In Gangemi,A. and Borgo,S. (eds), *Proceedings of the EKAW*04 Workshop on Core Ontologies in Ontology Engineering (CORONT)*. CEUR-WS, Northamptonshire.

Ritter,O. *et al.* (1994) Prototype implementation of the integrated genomic database. *Comput. Biomed. Res.*, **27**, 97–115.

Roos,M. *et al.* (2004) Future application of ontologies in e-Bioscience. *Position paper W3C Workshop on Semantic Web for Life Sciences*.

Ruttenberg,A. *et al.* (2007) Advancing translational research with the Semantic Web. *BMC Bioinformatics*, **8**, S2.

Schneider,R. *et al.* (2004) Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat. Cell Biol.*, **6**, 73–77.

Schulz,S. *et al.* (2006) Towards an upper level ontology for molecular biology. *AMIA Annu. Symp. Proc.*, pp. 694–698.

Searls,D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.*, **4**, 45–58.

Souchelnytskyi,S. (2005) Bridging proteomics and systems biology: what are the roads to be traveled? *Proteomics*, **5**, 4123–4137.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stein,L.D. (2003) Integrating biological databases. *Nat. Rev. Genet.*, **4**, 337–345.

Stevens,R. *et al.* (2000) TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics*, **16**, 184–186.

Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.

Strizh,I.G. (2006) Ontologies for data and knowledge sharing in biology: plant ROS signaling as a case study. *Bioessays*, **28**, 199–210.

The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, **306**, 636–640.

Turner,B.M. (2005) Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat. Struct. Mol. Biol.*, **12**, 110–112.

van Steensel,B. (2005) Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat. Genet.*, **37**, S18–S24.

Wilkinson,M. *et al.* (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The PlaNet exemplar case. *Plant Physiol.*, **138**, 5–17.

Wolstencroft,K. *et al.* (2006) Protein classification using ontology classification. *Bioinformatics*, **22**, e530–e538.

Zdobnov,E.M. *et al.* (2002) The EBI SRS server-new features. *Bioinformatics*, **18**, 1149–1150.