Aravind Venkatesan

Application of Semantic Web Technology to Establish Knowledge Management and Discovery in the Life Sciences

Thesis for the degree of Philosophiae Doctor

Trondheim, February 2014

Norwegian University of Science and Technology
Faculty of Natural Sciences and Technology
Department of Biology

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Natural Sciences and Technology
Department of Biology

ISBN 978-82-326-0038-0 (printed ver.)
ISBN 978-82-326-0039-7 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2014:56

Printed by NTNU-trykk

TABLE OF CONTENTS

Abstract

The last three decades has seen the successful development of many high-throughput technologies that have revolutionised and transformed biological research. The application of these technologies has generated large quantities of data allowing new approaches to analyze and integrate these data, which now constitute the field of Systems Biology. Systems Biology aims to enable a holistic understanding of a biological system by mapping interactions between all the biochemical components within the system. This requires integration of interdisciplinary data and knowledge to comprehensively explore the various biological processes of a system.

Ontologies in biology (bio-ontologies) and the Semantic Web are playing an increasingly important role in the integration of data and knowledge by offering an explicit, unambiguous and rich representation mechanism. This increased influence led to the proposal of the Semantic Systems Biology paradigm to complement the techniques currently used in Systems Biology. Semantic Systems Biology provides a semantic description of the knowledge about the biological systems on the whole facilitating data integration, knowledge management, reasoning and querying.

However, this approach is still a typical product of technology push, offering potential users access to the new technology. This doctoral thesis presents the work performed to bring Semantic Systems Biology closer to biological domain experts. The work covers a variety of aspects of Semantic Systems Biology:

The Gene eXpression Knowledge Base is a resource that captures knowledge on gene expression. The knowledge base exploits the power of seamless data integration offered by the semantic web technologies to build large networks of varied datasets, capable of answering complex biological questions. The knowledge base is the result of the active collaboration with the Gastrin Systems Biology group here at the Norwegian University of Science and Technology. This resource was customised by the integration of additional data sets on users' request. Additionally, the utility of the knowledge base is demonstrated by the conversion of biological questions into computable queries. The joint analysis of the query results has helped in filling knowledge gaps in the biological system of study.

Biologists often use different bioinformatics tools to conduct complex biological analysis. However, using these tools frequently poses a steep learning curve for the life science researchers. Therefore, the thesis describes ONTO-ToolKit, a plug-in that allows biologists to exploit bio-ontology based analysis as part of biological workflows in Galaxy. ONTO-ToolKit allows users to perform ontology-based analysis to improve the depth of their overall analysis.

Visualisation plays a key role in aiding users understand and grasp the knowledge represented in bio-ontologies. To this end, OLSVis, a web application was developed to make ontology browsing intuitive and flexible.

Finally, the steps needed to further advance the Semantic Systems Biology approach has been discussed.

Acknowledgements

गुरुर्ब्रह्मा गुरुर्विष्णुर्गुरुर्देवो महेश्वरः ।

गुरुरेव परं ब्रह्म तस्मै श्रीगुरवे नमः ॥

Salutations to thee, the transcendental Guru who quells the darkness in us.

The studies described in this thesis were carried out at the Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway, during the years 2009-2013

I recognize that this research would not have been possible without the financial assistance of the Faculty of Natural Science and Technology, the Department of Biology at the Norwegian University of Science and Technology.

At the onset, I would like to express my gratitude to my supervisor, Prof. Martin Kuiper, whose expertise, understanding and patience, added considerably to my doctoral experience. I thank him for all the support he has provided me, both academically and outside the confines of the University. His guidance during the last four and half years is much appreciated.

A very special thanks goes out to Dr. Vladimir Mironov, without whose insight and motivation I could not have achieved my research objectives. It was under his tutelage that I developed the technical skills and domain knowledge needed to conduct my research work. His keen eye for details will always be remembered.

I would also like to thank Prof. Astrid Lægrid for the scientific discussions, feedback and for sharing her wide knowledge. Her enthusiasm and positive attitude was invaluable for the progress of this work.

Further I want to thank my co-authors and colleagues: Dr.Sushil Tripathi, Konika Chawla, Anders Øverby, Dr. Ward Blonde, Dr. Erick Antezana and Dr. Steven Vercruysse. Your co-operation, support and scientific inputs are much appreciated.

I would like to thank my past supervisors and colleagues at EBI, Dr. Helen Parkinson, Dr. James Malone and Tony Burdett for their inspiration.

I acknowledge the support and encouragement given by Dr. Sanjay Molur and Payal Bhojwani Molur.

I thank my friends Deeptanshu Dwivedi, Akshay Vijayashekar, Jørgen Skancke and Itziar Castanedo for their support during the highs and lows of this study. A special shout out to all my school mates Mahendra, Vidit, Nihit, Varun, Niren, Aswin, Raghuram, Debanjan, Karthik, Pramod, Supradeep and Manas.

I would like to thank my best friend and beloved wife Shalini, for being the pillar of support throughout my doctoral study. I thank her for being extremely supportive, helpful and patient these last months. It was her encouragement that kept me going! I thank my brother and sister-in-law Rangarajan and Rasika for believing in me as always and providing me with all the support in spite of the long distance. I would like to thank my cousin Raju for his support and motivation. I thank my extended family, Abhinav and parents-in-law for their support and faith in me.

I extend my sincere gratitude to Dr. Raghavan and Mr. Balasubramaniam for believing in me.

Last, but most importantly I would like to profusely thank my parents Lakshmi and Venkatesan. I am extremely grateful for their unconditional love and support, as always it has fuelled my enthusiasm and determination. I take this opportunity to dedicate my PhD work to them.

Trondheim, February 2014

Aravind Venkatesan

List of Papers

Paper I: Antezana E., **Venkatesan A.**, Mungall C., Mironov V. and Kuiper M. (2010) ONTO-ToolKit: enabling bio-ontology engineering via Galaxy. *BMC Bioinformatics*, **11**(12), S8.

Contribution: Implemented the ONTO-PERL software on to the GALAXY framework and demonstrated the utility of ONTO-Toolkit (Use Cases I and II). Involved in drafting of the manuscript.

Paper II: Blonde W., Mironov V., **Venkatesan A.**, Antezana E., Baets B. D. and Kuiper M. (2011) Reasoning with bio-ontologies: using relational closure rules to enable practical querying. *Bioinformatics*, **27**(11), 1562-1568.

Contribution: Involved in the investigation phase of the project: testing the five closure rules separately by recreating the inferred version of BioGateway; involved in developing queries to demonstrate the utility of the closures.

Paper III: Venkatesan A.*, Mironov V.* and Kuiper M. Towards an integrated knowledge system for capturing gene expression events. Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO), Graz, Austria, July 21-25, 2012.

Contribution: Involved in the data integration pipeline, this includes developing parsers for integration of data from various sources described in the paper and drafted the manuscript.

Paper IV: Venkatesan A.*, Tripathi S.*, Galdeano A. S., Blonde W., Mironov V., Lægrid A. and Kuiper M. (2013) Network candidate discovery using the Gene eXpression Knowledge Base *(manuscript to be submitted to BMC Bioinformatics)*.

Contribution: Contributed in the integration of additional data sets, designed the experiments and carried out the querying exercise and drafted the final manuscript.

Paper V: Vercruysse S., **Venkatesan A**. and Kuiper M. (2012) OLSVis: an Animated, Interactive Visual Browser for Bio-ontologies. *BMC Bioinformatics*, **13**(1), 116.

Contribution: Involved in testing the software, this included extensive survey of other tools that supported visualisation of bio-ontologies. Involved in designing use cases and drafted the manuscript.

*equal contributions

Abbreviations

BFO	Basic Formal Ontology
CCK2R	Cholecystokinin 2 receptor
CCO	Cell Cycle Ontology
CellML	Cell Markup Language
ChEBI	Chemical Entities of Biological Interest
CWA	Concept Web Alliance
DAS	Distributed Annotated System
DbTF	DNA binding transcription factors
DL	Description Logic
DNA	Deoxyribonucleic acid
DTD	Document Type Definition
EBI	European Bioinformatics Institute
ECOR	European Centre for Ontological Research
EFO	Experiment Factor Ontology
EMBL	European Molecular Biology Lab
GeXKB	Gene eXpression Knowledge Base
GeXO	Gene eXpression Ontology
GO	Gene Ontology
GOA	Gene Ontology Annotation
HCLSIG	Semantic Web Health Care and Life Sciences Interest Group
HKB	HyQue Knowledge Base
HTRI	Human Transcriptional Regulation Interactions
IAO	Information Artifact Ontology
INFOMIS	Institute of Formal Ontology and Medical Information Science
ISA	Investigation-Study-Assay
JSON	JavaScript Object Notation
KGML	KEGG Markup Language
KM	Knowledge Management
KUPKB	Kidney and Urinary Pathway Knowledge Base
LLD	Linked Life Data
MAGE-ML	Microarray Gene Expression – Markup Language
MBO	Molecular Biology Ontology
MI	Molecular Interaction
MIAME	Minimum Information About a Microarray Experiment
MIBBI	Minimum Information for Biological and Biomedical Investigations
NCBO	National Centre for Biomedical Ontology
NTNU	Norwegian University of Science and Technology
OBI	Ontology for Biomedical Investigations
OBO	Open Biomedical Ontologies
OLS	Ontology lookup Service
OWL	Web Ontology Language

RDF	Resource Description Framework
RDFS	RDF Schema
REST	REpresentational State Transfer
ReTO	Regulation of Transcription Ontology
ReXO	Regulation of Gene eXpression Ontology
RNA	Ribonucliec acid
SB	Systems Biology
SBML	Systems Biology Markup Language
SIO	Semanticscience Integrated Ontology
SOAP	Simple Object Access Protocol
SPARQL	SPARQL Protocol and RDF Query Language
SPARUL	SPARQL Update Language
SPIN	SPARQL Inferencing Notation
SSB	Semantic Systems Biology
SW	Semantic Web
SWRL	Semantic Web Rule Language
SWS	Semantic Web Service
ULO	Upper Level Ontology
URI	Unique Resource Identifier
VoID	Vocabulary of Interlinked Datasets
W3C	World Wide Web Consortium
WS	Web Service
WSMO	Web Service Modelling Ontology
XML	eXtensible Markup Language

Introduction

Chapter 1

This thesis articulates some advances made in research on knowledge management, a paradigm that lies within the domain of Bioinformatics. Bioinformatics can be broadly described as a field that is at the intersection of biology and computer sciences. The primary goal of this field is to develop and provide the necessary computational means to archive, handle and mine data to increase the understanding of the various processes in the biological system. To this end we currently witness the field of bioinformatics getting branched out into specific research areas such as biological data storage and exchange, data processing and mining, text mining, gene and protein feature identification, protein structure prediction, gene expression analysis, and protein-protein interactions, to name a few. In this sense bioinformatics has certainly transformed into a field underpinning the biomedical domain. Among the numerous research directions that bioinformatics has taken, knowledge management extends upon the biological data storing and exchange sub-domain. To put this thesis into perspective, it is essential to trace back the origin of bioinformatics and how all of what is currently being done in knowledge management came into existence. The role of this introductory chapter is thus to provide the readers with a brief account on the history of Bioinformatics, the status of data, need for data management and integration, the concept of knowledge management, and finally providing the foundation to be able to understand the described work.

Emergence of Bioinformatics

Molecular biology as we know it today began with the discovery of the structure of the deoxyribonucleic acid (or DNA). In the 1950s, Francis Crick and James D. Watson developed a model that accurately explained the structure of DNA [1]. Furthermore, Crick presented the "Central Dogma" model that described the relationship between the DNA, RNA (Ribonucleic acid) and proteins, thus forming the foundations for the modern molecular biology [2]. This advancement was succeeded in the early 1970s by the discovery and exploitation of the enzymatic toolbox with bacterial endonucleases (restriction enzymes) and other nucleic acid modification- or synthesizing enzymes. Together with the vast advances made in determining the base sequence of DNA, especially in the last decade of the previous century, this allowed the determination of virtually the complete sequence of the human genome a mere 50 years after the discovery of the structure of DNA. In parallel, by the 1960s computer systems became widely available to academic researchers. With the increase in biological data the early molecular biologists were working towards understanding the complexity of some fundamental concepts such as genetic information for proteins, factors influencing protein structure, molecular homology and biological pathways. These were research areas that could be enhanced by using computational methods. The initial approaches already combined computational and experimental data in enhancing the understanding of these fundamental concepts marking the birth of computational biology [3, 4]. However, The successful completion of the Human Genome Project at the beginning of this century marked a sharp increase in the amount of data generated, because advancements in sequencing technology made the determination of the full DNA sequence of any organism almost trivial and very affordable, and with the sequence of a genome in principle the full potential for all bio-molecules made by that organism became known [5]. Similar technical advancements are being made for virtually every bio-molecule type, which explains the massive amounts of data that are produced today. So indeed the trend has been to first manage and analyse genome sequences and now more into the area of massive data analysis, data management.

The need for efficient data management has proven to be a great driver for the development of advanced computer algorithms and applications. As noted above, computational data analysis has been embraced by the molecular biologist well before the start of the Human Genome Project. This period marked an increase in the influence of computer technology in life sciences, especially to harness the enormous amounts of genomic data. Thus, Bioinformatics gained importance and became an independent scientific discipline in its own right. Institutes such as European Molecular Biology Lab (EMBL) in Heidelberg, Germany, formed the very first departments exclusively devoted to bioinformatics. One of the domains covered by Bioinformatics is the design, development and maintenance of bio-molecular databases. These databases differ mainly in the type of data they provide such as sequence databases (e.g. NCBI GenBank [6], EMBL data library [7]), transcriptome databases (e.g. ArrayExpress [8], GEO [9]), protein databases (e.g. UniProtKB [10]), pathway databases (e.g. KEGG [11], Reactome [12]) and protein-protein interaction databases (e.g. IntAct [13], MINT [14], BIND [15]).

The emergence of the powerful genome-wide data production technologies both allowed and made necessary new approaches to analyse and integrate these data, which now constitute the field of Systems Biology (SB). SB aims for a holistic understanding of a biological system in contrast to the more traditional reductionist approach that focuses on individual components of the system [16]. One of the foundations of SB is the use of mathematical and computational modeling: simulation of the behavior of complex biological systems through mathematical and computational models. These models serve to integrate all relevant biological data about a process or system, a comparison of the process and the model's behavior allows an assessment of the model's accuracy. The simulations ultimately help biologist predict the behavior of a system under new conditions which paves the way to new hypotheses that can be validated experimentally (Figure 1).

Perspectives on data in the Life Sciences

In the year 2001, business analyst Doug Laney of Gartner, Inc., USA, reported the trends in data explosion in the e-commerce domain. He categorised data growth into three factors, namely Volume, Variety and Velocity [17]. These factors have since been considered as the foundation for the concept of 'Big Data'. A closer look at the trend of data growth in the life science domain reveals similarities with the 'Big Data' concept. In this case, the Volume and Velocity factors can be attributed to mainly two aspects; a) the affordability in conducting high-throughput genome-wide technologies, the efforts taken in conducting these experiments and generating data is trivial when compared to the time spent in its integration and analysis; b) publishing data over the web has become fairly trivial.

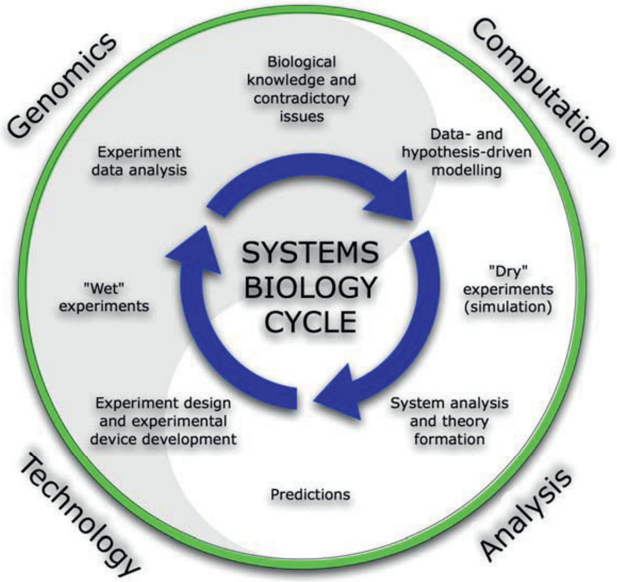

Genomics
Computation
Technology
Analysis
Biological knowledge and contradictory issues
Data- and hypothesis-driven modelling
"Dry" experiments (simulation)
System analysis and theory formation
Predictions
Experiment design and experimental device development
"Wet" experiments
Experiment data analysis
SYSTEMS BIOLOGY CYCLE

Figure 1: This schematic diagram depicts the Systems Biology cycle. Systems Biology is an integrative biology that requires a multidisciplinary effort, incorporating experimental design and data generation (‘wet’ lab) with knowledge extraction, data analysis, mathematical modelling and simulations (‘dry’ lab). These efforts are performed in a consecutive an iterative way with each cycle improving the quality of a system model.

Consequently, multiple databases exist for each type of data (i.e. ‘omics’ data) with a possible overlap or slight variation in its coverage. The number of databases in this domain outnumbers other data-intensive disciplines; currently, there are more than 1500 online databases covering various sub-domains that include nucleotide and protein sequence, protein-protein interactions and metagenomic databases [18]. Another factor that adds to these numbers is the existence of secondary and tertiary databases. These sources derive data from primary sources to provide a partial solution for data integration, however, in most cases these sources remain autonomous and disconnected [19, 20]. Furthermore, the ‘Variety’ factor is very distinct to the life science community. The biological system is a dynamic and complex one with constant interactions between different biomolecules. The rapid categorisation of these biomolecules creates the variety and complexity in the type of data being generated. As a result, the variety of data is followed by varied data representation formats. This is due to the fact that in the life sciences there is an absence of data standardisation. As articulated in the previous section, with the advent of SB, researchers are attempting to combine and overlay various data sets to understand various events in a biological system. Given the extreme heterogeneity of data models, formats and its volume, data interoperability and integration has become a challenging task.

To deal with this growing necessity to integrate data, an assortment of technologies have been developed and explored over the years mainly varying in their architecture and the level of automation. Some of the technological efforts made towards data integration are as follows:

- **Cross-database referencing:** Connecting data sets through cross-database indexing is achieved by providing cross references i.e. a data entry (a web page) of a data source is referenced to its corresponding entry in another data source. This "point and click" approach is most widely used by data providers and helps user surf through data entries by following the corresponding hyperlinks. This method hinges on ontological terms and identifier mappings that requires co-operation between the service providers. Although it is a quick method that aids users to browse through varied data sets, issues such as maintenance, updates and accuracy of the links are limiting factors. Moreover, this approach merely interlinks different data entries and the data integration part has to done by the user or a software application. This method is light on computation aspect and is incapable of answering complex biological question.

- **Data warehousing:** The data warehousing approach was adopted to provide a one-stop solution for data access. This method follows the Extract-Transform-Load (ETL) approach which is importing and translating data from various sources and transforming them into a pre-defined data model. In most cases the data warehouse is hosted by a third party, typically created to serve a particular sub-domain. Some popular data warehouses include Atlas [21], BioWarehouse [22], BioMart [23], and Columbia [24]. This approach has several advantages such as single point access as opposed to accessing different websites; elimination of network bottlenecks for querying; data maintenance undertaken by one party hence benefiting from data control [25]. However, when the nature of data is taken into account (specially the Velocity and Variety factors), there are a number of limitations such as low adaptability to changing data structures (source data), large up-front time investment required for data integration, technical challenges with respect to the infrastructure, data provenance and maintenance issues.

- **Data federation:** The data federation approach mainly depends on the query flexibility i.e. instead of transforming data into one store (like in data warehouse) applications are developed to execute queries to fetch data from their source and provide a unified view to the user. Popular initiatives that represent this approach include the Distributed Annotated System (DAS) [26], BIOZON [27] and Kleisli [28]. This approach presents an alternative to data warehousing method as the users have access to updated information with no data replication and is comparatively inexpensive. However, data federation is dependent on the network connectivity due to the geographical locations of the data sources affecting query efficiency and time consumed for fetching data.

- **Service architectures:** The Web Service (WS) method has evolved into a popular option for data integration in bioinformatics. This approach came into existence to evade the issues brought by both data warehousing and data federation. In this decentralised approach data providers agree to open their data via webservices. These are designed to make computers communicate with each other over the Web. In this technology, machine interoperability is afforded by describing the data in the Web Service Description Language (WSDL, XML-based language). With WS protocols such as SOAP (Simple Object Access Protocol) and REST (REpresentational State Transfer) data integration is supported by developing Web API (Application Programming Interface) as a tool for programmatic access to the source data. The potential of WS is well exploited by projects like Taverna [29], Enrez utilities Web Service [30] and KEGG Web Services [31] that uses WSs to build user defined workflows to integrate data. The BioCatalogue [32] serves as a curated catalogue of WS with the life sciences domain. This catalogue can be used to search for available WSs to build workflows on Taverna. However, the main challenge in this approach is the availability of WSs and the success also depends on the data exchange formats.

The heterogeneity and fragmentation of data sources has resulted in a significant gap between the generated of data and the knowledge being extracted from these data. To bridge this gap, a seamless data integration approach is required to support complex queries over the varied resources; facilitating data analysis, hypothesis generation and experimental design. However, this rapid development in data production has indeed increased awareness towards structured collection and quality control that should be adopted by the scientific community. This challenge was taken up by grassroots movements like the MIAME consortium (Minimum Information About a Microarray Experiment) [33], which put together recommendations for experiment design and data recording in the area of transcriptome analysis. This initiative has been extended by the global research community, including the Norwegian University of Science and Technology (NTNU) [34] to cover aspects of quality control and quality assurance. Similar initiatives have been taken up in virtually every area of data production. For instance, the Minimum Information for Biological and Biomedical Investigations (MIBBI) [35] is a web-based project that provides a common platform for ongoing 'minimum information' initiatives to promote collaborative and integrative development of data standardisation. There are two key aspects to the MIBBI project, a) the MIBBI portal, this provides access to information on a wide range of MIBBI-affiliated projects such as project scope and developmental status; b) MIBBI Foundry, modelled on the OBO Foundry (refer section: Bio-ontologies) to promote standard practices in producing orthogonal minimum information modules. Additionally, the Investigation-Study-Assay (ISA) [36] framework is an open source framework that is built on the 'Investigation' (the project context), 'Study' (a unit of research) and 'Assay' (analytical measurement) metadata categories to promote data exchange and integration. This is facilitated by the ISA-file format (ISA-Tab) that provides an extensible, hierarchical structure that enables the representation of studies that employs a particular technology or a combination of technologies. Also, the framework is supported by a software suite that aids in lowering the barriers for the users to implement ISA-Tab for the regularisation in experimental metadata management, and consistent curation.

Towards effective knowledge management in the Life Sciences

Having outlined the nature of data in the life sciences and the challenges concerning its integration, this section describes the concept of knowledge management and how the bioinformatics community has adopted elements of this concept to meet the data integration woes.

Knowledge management is a broadly defined concept varying from one domain to the other. For instance, knowledge management in the business domain would mainly deal with management of business activities such as business policies, assets and risk assessments. In comparison, knowledge management in bioinformatics deals with management of what is understood about the various components of a system of interest. Implementation of knowledge management practices may vary from a technology-driven approach to a cultural and behavioristic approach [37]. However, irrespective of the domain, to manage knowledge it is important to understand the three fundamental aspects: data, information and knowledge. The definitions of these fundamental blocks seem to be constantly changing [38] and it is beyond the scope of this work to dwell deep into the precise definition of these terms. However, Floridi [39] presents a detailed discussion on differences between these terms. From the perspective of knowledge management in bioinformatics the following definitions have been proposed and are implied throughout this thesis:

- **Data:** unstructured observations made to understand an event within a biological system using trusted empirical methods. For instance, output from a high-throughput experiment consists of a matrix with numbers.
- **Information:** a collection of (processed) data that makes decisions easier i.e. numbers along with metadata.
- **Knowledge:** it is the outcome of placing the information within context i.e. interpreted information for a thorough understanding of the underlying meaning.

Generally speaking, the evolution of the concept of data is synchronous with the advances made in the field of information technology. From the proposed definition it is clear that a set of data warrants interpretation with respect to a given context which subsequently creates knowledge. The task of interpretation was traditionally carried out by scientists but with the outpour of large–scale data; interpretation is highly dependent on computer systems. Hence, to allow interpretation from a multitude of data sources requires sophisticated computer technologies that integrates data and bridges the gap between data and its underlying meaning. This process of systematically capturing and structuring data to enable the understanding of a system is called Knowledge Management (KM). Thus, efficiently managed knowledge in principle will make data analysis easier aiding in more efficient decision making. An effective KM system will enable the hierarchical information flow, having a significant effect on the performance of research groups both in academia and life science companies [40, 41, 42].

Efficient KM promotes exchange and reusability of information, mainly hinging on two fundamental aspects:

- **Knowledge Representation:** Knowledge representation plays a crucial role in the facilitation of processing and sharing knowledge between people and application systems. Knowledge representation entails development of formal languages that enables modelling of the entities (both physical and abstract entities) in a domain through modelling guidelines that provide a standard, agreed upon by domain experts. Additionally, knowledge representation languages should adopt a common syntax that is reusable and enables parsing of data in a semantically unambiguous manner [20]. These conditions will support intelligent inferencing of facts over a given domain or sub-domain and facilitate processing of information in a computational environment.

- **Data Integration:** To effectively capture knowledge, data needs to be combined from various sources. The different approaches taken towards data integration have been discussed in the previous section. These approaches failed to deliver as KM platforms due to the vagueness in their approach, lack of semantics and reusability.

One of the main data exchange formats used in bioinformatics is the XML with numerous variants developed to support data exchange for specific sub-domains, this includes CellML [43], SBML [44], KGML [45] and MAGE-ML [46], to name a few. Documents represented in XML are based on a nested tree structure which consists of attribute-value pair; the Document Type Definition (DTD) or the XML schema enforces constraints on how the data needs to be nested, defining the grammar of the document. XML serves well as a data exchange format, especially when two applications are aware of the data being exchanged. Given the volatile nature of biological data, the structure of data and its relationship is bound to change. The inclusion of the new data elements should be carried out by making an extension in the schema. For a domain with constantly changing data this process is very cumbersome. Furthermore, overlaying different data sets requires mapping different sub-domain models and it is nontrivial using XML. This issue of interoperability is not due to the lack of agreement between various interested parties using these XML variants but in the architecture itself, mainly due to the rigidity in the schema. Matching two models would require reengineering of the models to accommodate mappings between the concepts and relationships; additionally the mapping must be defined in the schema [47]. Moreover, XML schema is limited to the syntax and does not specify the semantics of the content, lacking granularity with respect to semantics between the nested nodes [48]. Alternatively, in recent times the JSON (JavaScript Object Notation) format is being widely used as a data exchange format. The format is more human readable and imposes far less overhead for computer application to parse and extract data. This is mainly because the JSON data structure translates directly into the native data structure common to most of the programming languages.

Bio-ontologies:

In recent times, knowledge representation and knowledge management have gained popularity and growth in the life sciences, an important trend towards this direction is the adoption of ontologies. The concept of ontologies stems from a branch of philosophy that engages in the study of being. This approach dates back to the days of ancient philosophers such as Plato and Aristotle [49], to provide a definitive and systematic classification of entities of nature and reality in general. For the first time in 1913 philosophers, Rudolf Gockel and Jacob Lorhard coined the term 'ontology' for this method of classifying entities [50]. Furthermore, Quine [51] formalised the ontology as a tool to understand scientific theories in the language of first order logic. Gradually, ontology development became a part of the information sciences domain; where it was used to create vocabularies (concepts and their relations) for a domain of discourse, yielding a framework for knowledge sharing, reusability and reasoning for various streams of research [49, 52]. One of the main factors that fuelled research on ontology application was the rapid increase in data over the Web. To integrate data, improve machine interoperability and data analysis required a conceptual scaffold. This is especially visible in the life science domain. Bio-ontologies have certainly helped in capturing the semantics of entities and their interrelationships within biology, thereby reducing conceptual ambiguity, increasing re-usability and computational automation that aids knowledge discovery.

Initially, Ocelot - a knowledge representation system was developed to organise information with similar properties in the form of classes [53] for EcoCyc – a species specific genome/pathway database. Similar knowledge systems were developed to organise biological data such as Molecular Biology Ontology (MBO) [54], RiboWeb [55], TAMBIS [56] and PharmGKB [57]. Figure 2, shows the time line for the growth of ontologies in molecular biology and related fields marking the first phase in the development of ontologies in the life science domain. Currently, one of the main ontologies used in the life sciences is the Gene Ontology (GO) [58], which facilitated the unambiguous annotation of biomolecules with terms specifying molecular functions, cellular components and biological processes. The controlled vocabulary terminology, the relational rules and more importantly the unique IDs provided in GO are incorporated by the database source, providing unique and cross-domain common entry points in the description of the gene products, the Gene Ontology Annotation (GOA) [59] project highlights this effort. This aids the users (life scientists) in further investigation of a gene of interest, enriching the knowledge related to that gene. The success of GO gave rise to the establishment of the Open Biomedical Ontologies (OBO) [60] consortium, who among others provide a set of foundational principles to structure the further co-ordinated development of bio-ontologies (e.g. ontology orthogonality: different ontologies in the foundry should not overlap). The OBO foundry now constitutes a set of 130 domain-specific candidate ontologies, which are becoming widely accepted as a reference by the life science community. Additionally, there have been a number of institutes established with the focus of supporting the development of bio-ontologies. Among others, a few notable ones are as follows:

The Institute of Formal Ontology and Medical Information Science (INFOMIS) [61]: The institute is located at the University of Saarbrücken, Germany. The centre was founded in 2002 bring specialists from fields such as Philosophy, Information Science and Medicine, focusing on theoretical research on formal and applied ontologies in the bio-medical domain. The institute contributed significantly for the development of GO.

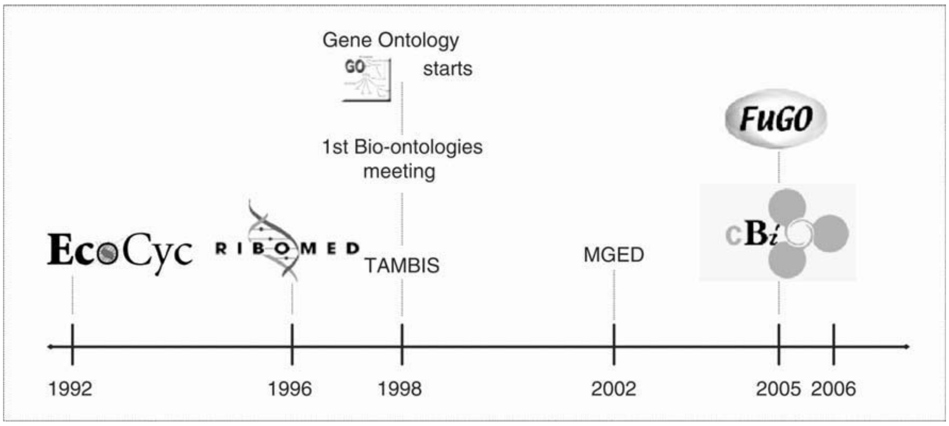


Figure 2: Shows the time line of the appearance of bio-ontology/ontology-like knowledge sharing systems. Reprinted from Bodenreider and Stevens, 2006 [62].

The European Centre for Ontological Research (ECOR) [63]**:** The center was founded in 2004 at the University of Saarbrücken, Germany. The center mainly focuses on design, development and implementation of ontology for varied domains. The centre is involved in developing ontologies for the life science domain in collaboration with INFOMIS.

The National Centre for Biomedical Ontology (NCBO) [64]**:** The centre founded created in 2006, as a part of the National Centre for Biomedical Computing, USA. NCBO has been actively involved in the development ontologies to promote semantic interoperability of knowledge and data. NCBO BioPortal, is one such effort towards this endeavour.

European Bioinformatics Institute (EBI): One of the flagship institutes for the Bioinformatics domain is actively involved in the development of ontologies and knowledge bases to provide interoperable resources. Ontologies such as the Chemical Entities of Biological Interest (ChEBI) [65], Experimental Factor Ontology (EFO) [66] and the Ontology for Biomedical Investigations (OBI) [67] were developed under the auspices of EBI.

Ontologies can be classified according to the degree of conceptualisation which includes:

- **Upper-level ontology:** Ontologies that describes general concepts which are independent of a particular domain. Their applicability is in providing support to a large number of ontologies. The Basic Formal Ontology (BFO) [68] is a widely used upper level ontology in a number of sub-domains within the life sciences.

- **Domain ontology:** The knowledge represented in this type of ontology serves a particular domain by providing vocabularies about concepts and their relationships governing the domain such as GO.

- **Application ontology:** These ontologies are typically used to define concepts for a particular use case. For instance, EFO is used to represent concepts and sample variables from gene expression experiments and Cell Cycle Ontology (CCO) [69], an ontology that captures knowledge related to the cell cycle processes.

Furthermore, ontologies can range from simple taxonomies (with 'is-a' hierarchy among concepts) to highly interconnected networks including constraints associated with concepts and relations (cardinality constraints and axioms). This is afforded by the type of ontology language (knowledge representation language) that is chosen for ontology development and plays an important role in the machine readability of the ontology. For instance, ontologies adhering to formal logic such as Description Logic (DL) and expressed in the Web Ontology Language (OWL) render a higher degree in machine-readability when compare to ontologies developed in the OBO format, designed to be more human readable.

Bio-ontologies have become a cornerstone for knowledge representation and management in life sciences. Currently, numerous applications exploit the advantages offered by bio-ontologies. Tools such as Cytoscape [70] and ONDEX [71] use ontologies to integrate and visualise diverse data sets to build various biological networks. A number of plug-ins have been developed to provide ontology driven data analysis, for instance BiNGO [72] utilises the GO vocabulary to determine the overrepresentation of GO terms among a set of genes; similarly PiNGO [73] utilises user-defined target GO categories to identify and classify unknown genes in a network; and finally RDFScape [74], a tool that brings Semantic Web (refer section: Semantic Web Technology) functionalities to Cytoscape for the utilisation of ontology-based knowledge to enhance biological analysis.

Semantic Web Technology:

Bio-ontologies have been successfully used as method to unambiguously represent domain knowledge. However, as discussed earlier (see section: Towards effective knowledge management of Life Science resources), efficient knowledge management additionally requires the adoption of effective data integration methodologies. This entails efficient integration of the disparate data sources, represented in a machine-readable format.

The emergence of the Semantic Web Technology (SW) is starting to have a significant impact on knowledge integration, querying, and knowledge sharing in the life science domain. Conceptualised by Tim Berners-Lee in 2001 [75], SW intends to convert the current World Wide Web which consists of hyperlinked pages into a Web of Knowledge that is machine comprehensible. The SW depends on a set of web technologies specifically designed to facilitate automated machine interoperability. It promises to meet the challenge of integrating and querying highly diverse and distributed resources. Systems based on SW would provide sophisticated frameworks to manage and retrieve knowledge. This is achieved by the use of well established web technologies and a number of essential components added on top. The architecture of the SW involves a hierarchical assembly (Semantic Web Stack, Figure 2) of various formats and technologies where each layer exploits the capabilities of the layer below providing a formal description of concepts and relationships within a given domain. The bottom layers in the Semantic Web Stack consists of technologies that are widely used in the current Web, thus SW is built on the basis of these technologies and the middle layer (RDF onwards) consists of technologies that have been standardised specifically to support semantic web applications. The various components of the Semantic Web Stack are as follows:

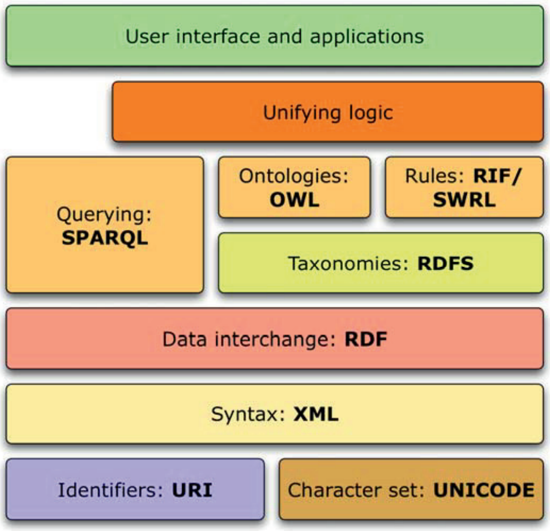


Figure 3: *Semantic Web Stack*

Unicode: This is a standard developed for the consistent representation of text expressed in the world's various writing systems. The usage of unicode aids the semantic web applications in bridging documents expressed in different human language systems.

Unique Resource Identifier (URI): A URI is a series of characters used to identify an abstract or a physical entity. URIs are used in Semantic Web based systems to describe a resource and its components, enabling interactions over the Web using the Hypertext Transfer Protocol (HTTP). The URI is one of the fundamental aspects of SW; they are more generic when compared to the Unique Resource Locator (URL) and the Unique Resource Name (URN). URIs are formed in a way that it can be used as a URL (locator) or URN (name space) or as both. Recently, the Internationalised Resource Identifier (IRI) was proposed as a generalisation of the URI [76], accommodating all Unicode characters which includes characters from other languages such as Chinese and Japanese.

Extensible Markup Language: The e**X**tensible **M**arkup **L**anguage **(XML)** is a widely used language to exchange data over the internet. XML provides an elementary structure in describing data with documents consisting of nested sets of open and closed tags with corresponding data indexed using labels. The restrictive syntax rules of XML are highly suited for the Semantic Web and provide a scaffold in data representation.

Resource Description Framework (RDF): RDF is a data modelling language that provides a framework to describe a resource and its relationship with other resources in the form of *triples*, i.e. Subject-Predicate-Object. A set of such triples forms a RDF graph that is made machine-readable using the XML serialisation (Figure 3). Furthermore, the World Wide Web Consortium (W3C) has proposed additional non-XML serialisations for RDF, this includes Turtle [77], Notation 3 [78], N-Triples [79] and JSON-LD [80]. These formats make the underlying RDF triples more user-friendly and perform better while parsing. The RDF graphs can be stored in repositories called Triple stores or RDF stores. They are categorised under NoSQL databases, where one could store and query the graphs via a query language, SPARQL (also known as SPARQL endpoints). In recent times, a number of efficient scalable triple stores have been developed equipped with SPARQL functionality, e.g. Openlink Virtuoso [81], Apache Jena [82] and 4store [83].

RDF Schema (RDFS): RDF provides flexibility and machine interoperability in data representation. However, to create a Web of Knowledge, systems need to have reasoning capability. This requires advanced logics capable of capturing the complexity of data. RDF technology does not provide logical meaning for the data descriptions. Hence, to bring about an improvement in this area the RDFS was developed. RDFS is an ontology language that facilitates simple inferences through hierarchical specification of classes, sub-classes, properties and sub-properties. RDFS vocabulary is abundantly used in RDF graphs, however, it has not added much in terms of advancing reasoning due to the limitations in its expressivity. It does not support commonly required features such as union, negation and disjunction.

SPARQL: The **S**PARQL **P**rotocol and **R**DF **Q**uery **L**anguage is a query language for RDF. It offers the developers and end users a way to retrieve and manipulate data stored in RDF format. It is considered as one of the key technologies of semantic web, as it allows users to write unambiguous queries consisting of triple patterns, conjunctions, disjunctions, and optional patterns. Furthermore, as an extension to SPARQL, the SPARQL/Update (SPARUL) has been developed. It is a declarative data manipulation language that offers the ability to insert, delete and update RDF data held within a triple store. With the latest updated version of SPARQL, SPARQL v1.1, more features have been added to the language for instance, SPARQL 1.1 provides its users with the capability of query federation i.e. the ability to launch SPARQL queries to different RDF stores.

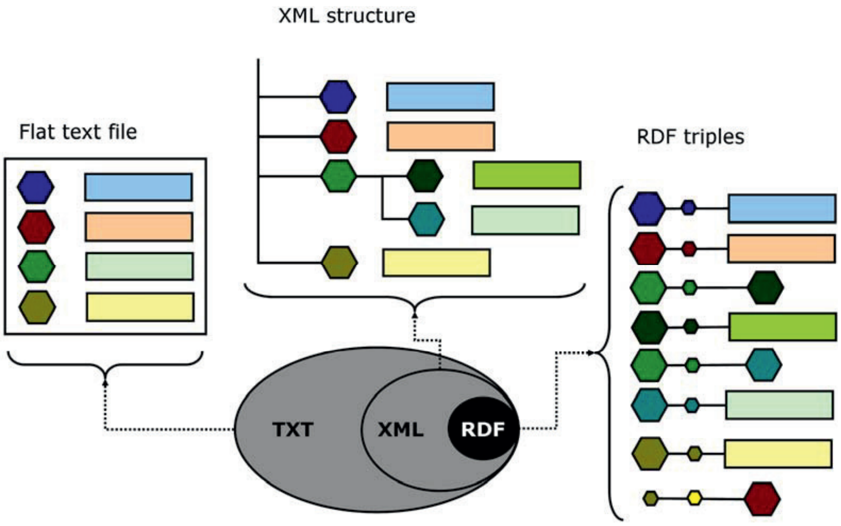


Figure 4: Evolution of data representation formats: The large hexagons and rectangles represent the data elements, attributes and their corresponding values, and the small hexagons represents the relation between the attribute-value pair. Reprinted from Deus et al., 2008 [84].

Web Ontology Language (OWL): OWL goes beyond RDF and RDFS in its expressiveness by providing logical constructs to the description of classes. OWL was developed based on already existing languages such as OIL and DAML made compatible with RDF. W3C provided three specifications for OWL (also known as OWL 1): OWL-Lite, OWL-DL and OWL-Full. OWL-DL was developed based on a well understood fragment of Description Logics (DL, www.dl.kr.org), which guarantees computability. As a consequence, a number of DL reasoners, developed by the artificial intelligence community, were made available for deployment in the Semantic Web for: a) consistency checking, b) reasoning, c) classification and d) querying. FaCT++, Pellet and RACERPro are among the most commonly used DL reasoners. Although, research on DL has kept the hope alive for realising fully automated reasoning on bio-ontologies. The application of reasoners on fully fledged large, integrated bio-ontologies failed due to scalability issues. In contrast, OWL-Lite developed as a sublanguage of OWL-DL is less expressive and provides better performance for reasoning tasks. However, with the recent release of OWL 2 significant work has gone into making reasoning more tangible, offering several DL-based sublanguages such as OWL 2 DL, OWL 2 EL and OWL 2 RL [85]. Particularly, OWL 2 EL appears to be a promising alternative within the life science domain [86, 87].

Rule Languages: The Semantic Web Rule Language (SWRL) [88] was initially developed in 2004. This proposed language for SW could be used to specify logical inferencing. SWRL is based on a combination of RuleML [89] and OWL-DL or OWL Lite. However, SWRL has not become a W3C recommendation due to practical implementation hurdles. Alternatively the more recently developed SPARQL Inferencing Notation (SPIN) [90], submitted to W3C in 2011, provides standards for implementing rules on SW models expressed in SPARQL. The implementation of SPIN rules is straightforward, as it uses already established methods such as SPARQL CONSTRUCTS and SPARQL Update requests.

Since its inception SW has gained steady acceptance among the life science community. Several projects have been undertaken to demonstrate the potential of SW, some notable initiatives are as follows:

- **Bio2RDF** [91]: an open-access SW knowledge base that provides a mashup over 19 different data sets that includes the Gene Ontology, OMIM, Reactome, ChEBI, BioCyc and KEGG.

- **BioGateway** [92]: a SW resources that integrates the entire set of OBO Foundry ontologies (including both accepted and candidate OBO ontologies), the complete collection of annotations from the Gene Ontology Annotation (GOA) files, fragments of the NCBI taxonomy and SWISS-PROT and IntAct. This project marked the fusion of SW to Systems Biology (refer section: Semantic Systems Biology).

- **Linked Life Data** [93]: a semantic data integration platform developed by Ontotext as part of the Large Knowledge Collider (LarKC) project. The platform interconnects data sets from the Pathway and Interaction KB (PIKB), PubMed, KEGG, IntAct, MINT, Entrez-Gene, and the SKOS representation of OBO ontologies.

- **KUPKB** [94]: a SW knowledge base that integrates high-throughput experiment data on kidney and urine. This includes data from sources such as NCBI Gene, UniProt and KEGG.

- **HyQue** [95]: a SW tool for the evaluation of hypotheses on molecular events pertaining to the galactose gene network in *Saccharomyces cerevisae*. The hypotheses submitted by the user are validated using the HyQue Knowledge Base (HKB). HKB contains data on various molecular events like protein-protein interaction, regulation of transcription and gene expression.

These initiatives have certainly helped to demonstrate the advantages of the SW technologies, including a richer knowledge representation, streamlined data integration and efficient querying. In particular, the graph-based data model of RDF makes it a compelling choice to model knowledge and integrate data from multiple sources. It has become the cornerstone for data integration across computing platforms due to its flexibility and its suitability to represent concepts in the biomedical domain. Through the query language SPARQL, users are provided with the capability of simultaneously querying and integrating results from multiple RDF graphs. With properly designed RDF graphs the querying is very robust [96] and in principle it is even possible to query multiple RDF stores. This has also encouraged primary data providers to publish their data in RDF. Currently, the EBI has setup the *RDF platform* [97] aimed at providing RDF representation of EMBL-EBI resources i.e. BioModels, BioSamples, ChEMBL, Expression Atlas, Reactome and UniProt. Additionally, the utilisation of SW has been pushed further by combining it with WS called the Semantic Web Service (SWS) framework. SWS is aimed to provide solutions for the challenges faced in a WS setup by adding semantics to WS standards for automation of data processing and reasoning. Ontologies such as OWL-S [98] and the Web Service Modelling Ontology (WSMO) [99] and the capabilities offered by RDF are used to model various aspects of WS such as service interfaces and structures, enabling automation of discovery and invocation of services. SWS frameworks such as SSWAP [100], SADI [101] and BioSemantic[102] have provided a proof of concept towards describing WS with semantic annotations by the use of ontologies. Considering that the data within the life sciences will further increase and make the domain more data intensive, SWS will adopt the advances made in high-performance computing such as cloud computing [103, 104] to setup SWS pipelines.

Semantic Systems Biology:

As discussed earlier SB seeks to explain the various biological phenomena through a web of interactions between all the biochemical components within biological systems. Obviously, this requires integration of interdisciplinary data and knowledge to comprehensively explore the various biological processes of a system. To this end, SW has been proposed as a part of the work flow in SB [72, 105]. Antezana et al. [106] have taken a significant step in fusing SW and SB, as their initiative has resulted in the proposal of a platform called Semantic Systems Biology (SSB) [107]. SSB provides a semantic description of the knowledge about the biological systems on the whole facilitating data integration, knowledge management, reasoning and querying (which plays a key role in hypothesis generation). Figure 4 is a schematic representation of the work flow in SSB. Firstly, biological knowledge is extracted from disparate resources and integrated into a knowledge base. Through inferencing and querying of the integrated data hypotheses are generated based on which new experiments could be designed. Through experimentation the hypotheses could be validated and further added to the knowledge base completing the cycle of SSB. Essentially, SSB is similar to SB in that it constitutes a cyclical process (Figure 1) that revolves around hypothesis generation, which in SB is driven by computational modelling and in SSB by computational reasoning and querying. Furthermore, the impact of SW in the life sciences has urged the World Wide Web Consortium to establish a special interest group, the Semantic Web Health Care and Life Sciences Interest Group (HCLSIG) [108]. The interest group is currently involved in the development and support the use of SW, a special task force has been setup to explore the application of SW for Systems Biology [109]. SW as an area of research is very active and is continuously improving (e.g.: SPARQL 1.1, OWL 2 and SPIN) and these initiatives makes SW a compelling aspect of the Semantic Systems Biology.

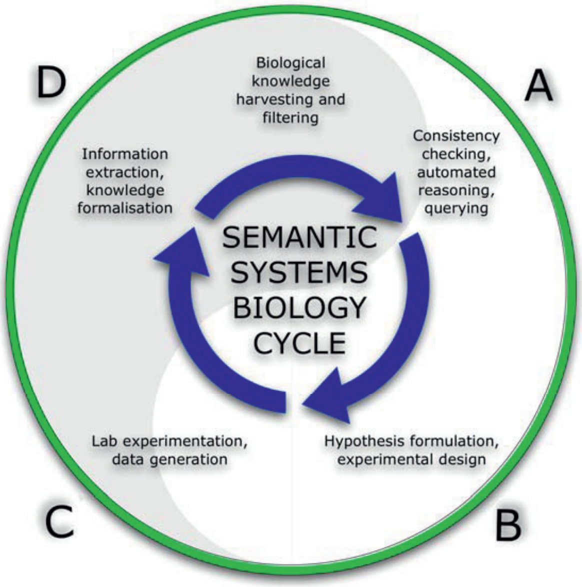

D
Biological knowledge harvesting and filtering
A
Information extraction, knowledge formalisation
Consistency checking, automated reasoning, querying
SEMANTIC SYSTEMS BIOLOGY CYCLE
Lab experimentation, data generation
Hypothesis formulation, experimental design
C
B

Figure 5: The Semantic Systems Biology cycle. The cycle starts with gathering and integrating biological knowledge into a semantic knowledge base; (A) Data is then checked for consistency, and made available for querying and automated reasoning; (B) This yields hypotheses about particular functions of biological components that may be used to design experiments; (C) The experiments generates new data and might also verify hypotheses. (D) The new findings are integrated into the knowledge base, thereby enhancing the quality of the knowledge base and allowing a new cycle of hypothesis building and experimentation.

Objectives of the Study:

The definitive aim of bioinformatics research is to automate data analysis and generate hypotheses ultimately reducing the knowledge discovery cycle for the biologists. To that end, the Semantic Web provides solutions to make seamless data integration a reality, bridging the data-knowledge gap. The various initiatives that provide semantic web solutions have greatly facilitated this process by increasing the critical mass of integrated data. However, these resources are still typical products of a technology push, offering potential users access to the new technology. As the Semantic Web community wrestles to provide better solutions to handle the data deluge in the life science community, it is also important that the technology is brought closer to biologists. To promote the advantages of the semantic web and to encourage the end-users to utilise these resources as part of their daily research activities requires the lowering of boundaries to adopt the new approach. For instance, exploiting various ontologies to performing ontology-driven data analysis could enhance the process of new hypothesis generation that can drive new experimental design. Correspondingly, tools are required to allow easy manipulation ontologies and further link them to other tools that allow further analysis. Furthermore, the semantic web knowledge bases currently provide a SPARQL endpoint as means to access the integrated data but these endpoints are more machine-friendly. In some cases, the knowledge bases are equipped with faceted browser interfaces or sample queries suitable for a quick search, such as retrieving a local neighborhood for a particular term (e.g.: ontological term, protein or gene identifiers). Although, this helps the user to get acquainted with the data housed in these resources, hypotheses generation warrants formulation of complex queries. This requires at the minimum a moderate knowledge of SPARQL. It is evident that the SPARQL technology may be intimidating to the biologists resulting in semantic web resources not being exploited completely. Therefore, attempts have to be made to pursue various joint initiatives with the intended target audience to mobilise 'user-pull' as an essential source for knowledge base design ideas and to develop real world use cases to generate hypotheses.

Therefore, this study was aimed at investigating approaches to bring the Semantic Systems Biology closer to the user community. The definitive objectives of this study were:

1. To use or develop tools that provides a user friendly interface allowing biologists to conduct ontology-based analysis.

    - **Paper I:** *ONTO-ToolKit: enabling bio-ontology engineering via Galaxy* – ONTO-ToolKit is a tool suite for Galaxy [110] that allows biologists to manipulate OBO ontologies and opens up the possibility to perform further analyses by using other tools available within the Galaxy environment.

2. Contribute to the development of a knowledge base that captures knowledge of a domain of discourse. Further, it supports intuitive and complex queries to generate hypotheses.

- **Paper II:** *Reasoning with bio-ontologies: using relational closure rules to enable practical querying* – This paper demonstrates reasoning on bio-ontologies as a semi-automated process by using Metarel, a vocabulary that specifies relation semantics, to apply reasoning on a large semantic web knowledge bases.

- **Paper III:** *Towards an integrated knowledge system for capturing gene expression events* – This paper outlines the development of a semantic web based knowledge system, the Gene eXpression Knowledge Base (GeXKB). The knowledge base is capable of answering complex biological questions with respect to gene expression and facilitates the generation of new hypothesis.

- **Paper IV:** *Network candidate discovery using the Gene eXpression Knowledge Base* – This paper describes the further enhancement of GeXKB and demonstrate the potential of this resource in the identification of candidate network components (proteins) that may be considered for network extension. The network components were generated by the formulation of SPARQL queries based on set of biological questions launched to GeXKB.

3. Contribute to better visualisation of ontologies that makes ontology browsing intuitive and flexible.

   - **Paper V:** *OLSVis: an Animated, Interactive Visual Browser for Bio-ontologies* – OLSVis is a web application that provides a user-friendly environment to visualise bio-ontologies from the OLS repository. It broadens the possibilities to investigate and select ontology subgraphs through a smooth visualisation method.

Reference

[1] Watson, J. D. and Crick, F. H. C. (1953) A Structure for Deoxyribose Nucleic Acid. *Nature,* **171**, 737-738.

[2] Crick, F. (1970) Central Dogma of Molecular Biology. *Nature,* **227**, 561 – 563.

[3] Hagen, J. B. (2000) The origins of bioinformatics. *Nature Reviews Genetics*, **1**(3), 231-236.

[4] Ouzounis, C. A. and Valencia, A. (2003) Early bioinformatics: the birth of a discipline—a personal view. *Bioinformatics*, **19**, 2176-2190.

[5] Gisler, M., et al. (2010) Exuberant Innovation: The Human Genome Project. Swiss Finance Institute Research; Paper No. 10-12. Available at SSRN: http://ssrn.com/abstract=1573682

[6] Benson, D. A., et al. (2010) GenBank. *Nucleic Acids Res*, **38**(1), D46-D51.

[7] EMBL data library: http://www.ebi.ac.uk/ena/

[8] Brazma, A., et al. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res, **31**(1), 68-71.

[9] Edgar, R., et al. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**(1), 207-10.

[10] Apweiler, R., et al. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, **32**(1), D115-D119.

[11] Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**(1), 27-30.

[12] Vastrik, I., et al. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biology*, **8**(3), R39.

[13] Kerrien, S., et al. (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, **40**, D841-D846.

[14] Ceol, A., et al. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res,* **38**, D532-9.

[15] Bader, G. D., et al. (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Res*, **31**(1), 248-250.

[16] Kitano, H. (2002) Systems biology: a brief overview. *Science*, **295**(5560), 1662-1664.

[17] Laney, D. (2001) 3D Data Management: Controlling Data Volume. *Velocity, and Variety, Application Delivery Strategies published by META Group Inc*.

[18] Fernández-Suárez, X. M. and Galperin, M. Y. (2013) The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Res*, **41**, D1-D7.

[19] Goble, C. and Stevens, R. (2008) State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*, **41**(5), 687-693.

[20] Antezana, E., et al. (2009) Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief. in Bioinformatics*, **10**(4), 392-407.

[21] Shah, S. P., et al. (2005) Atlas - a data warehouse for integrative bioinformatics, *BMC Bioinformatics*, **6**, 34.

[22] Lee, T. J., et al. (2006) BioWarehouse: a bioinformatics database warehouse toolkit, *BMC Bioinformatics*, **7,** 170.

[23] Haider, S., et al. (2009) BioMart Central Portal--unified access to biological data. *Nucleic Acids Res*, **37**, W23-27.

[24] Trissl, S., et al. (2005) Columba: an integrated database of proteins, structures, and annotations, *BMC Bioinformatics*, **6**, 81.

[25] Zhang, Z., et al. (2011) Data integration in bioinformatics: Current efforts and challenges. *Bioinformatics-Trends and Methodologies,* 41-56.

[26] Dowell, R. D., et al. (2001) The distributed annotation system. *BMC bioinformatics*, **2**(1), 7.

[27] Birkland, A. and Yona, G. (2006) BIOZON: a hub of heterogeneous biological data, *Nucleic Acids Res*, **34**, D235-242.

[28] Davidson, Susan B., et al. (2001) K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. *IBM systems journal* **40**(2), 512-531.

[29] Oinn, T., et al. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045-3054.

[30] Entrez Utilities Web Service at NCBI:
http://eutils.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html

[31] KEGG Web Services: http://www.genome.jp/kegg/soap/

[32] Goble, C. A., et al. (2009) BioCatalogue: a curated Web Service registry for the Life Science community. Available from *Nature Precedings.*

[33] Brazma, A., et al. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics,* **29,** 365-371.

[34] Beisvåg, V., et al. (2011) Contributions of the EMERALD project to assessing and improving microarray data quality. *BioTechniques*, **50,** 27-31.

[35] Taylor, C. F., et al. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol,* **26,** 889-896.

[36] Susanna-Assunta, S., et al. (2012) Toward interoperable bioscience data. *Nature genetics,* **44**(2), 121-126.

[37] Barclay, R. O. and Murray, P. C. (1997) What is knowledge management. *A Knowledge praxis*.

[38] Zins, C. (2007) Conceptual approaches for defining data, information, and knowledge. *J. Assn. Inf. Sci. Technol*, **58**(4), 479-493.

[39] Floridi, L. (2007) Semantic conceptions of information. In E. N. Zalta, editor, The Stanford Encyclopedia of Philosophy. *Spring.*

[40] Hodgson, J. (2001) The headache of knowledge management. *Nature Biotechnology*, **19**, BE44-BE46.

[41] Szalay, A. and Gray, J. (2006) 2020 Computing: Science in an exponential world. *Nature*, **440**(7083), 413-414.

[42] Attwood, T., et al. (2009) Calling International Rescue: knowledge lost in literature and data landslide! *Biochem. J.*, **424**, 317-333.

[43] Lloyd, C. M., et al. (2004) CellML: its future, present and past. *Progress in biophysics and molecular biology*, **85**(2), 433-450.

[44] Hucka, M., et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**(4), 524-531.

[45] KEGG Markup Language: http://www.kegg.jp/kegg/xml/

[46] Spellman, P. T., et al. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome biology*, **3**(9), research0046.

[47] Decker, S., et al. (2000) The semantic web: The roles of XML and RDF. *Internet Computing, IEEE*, **4**(5), 63-73.

[48] Wang, X., et al. (2005) From XML to RDF: how semantic web technologies will change the design of 'omic' standards. *Nature biotechnology*, **23**(9), 1099-1103.

[49] Smith, B. (2003) Ontology. *The Blackwell guide to the philosophy of computing and information*, 153-166.

[50] Ingarden, R. (1964) *Time and Modes of being,* translated by Helen R. Michejda, Springer, Ill: Charles Thomas. Translanted extracts from a masterly four volume work in realist ontology entitled *The problem of the Existence of the World.*

[51] Quine, W. V. O. (1953) On what there is. *From a logical point of view*, *9*.

[52] Gruber, T. R. (1991) The role of common ontology in achieving sharable, reusable knowledge bases. *KR*, **91**, 601-602.

[53] Karp, P. D., et al. (2002) The ecocyc database. *Nucleic acids research*, **30**(1), 56-58.

[54] Schulze-Kremer, S. (1997) Adding semantics to genome databases: towards an ontology for molecular biology. In *Ismb,* **5**(272), 5.

[55] Altman, R. B., et al. (1999) RiboWeb: An ontology-based system for collaborative molecular biology. *Intelligent Systems and Their Applications, IEEE*, **14**(5), 68-76.

[56] Stevens, R., et al. (2000) TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics*, **16**(2), 184-186.

[57] Klein, T. E., et al. (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. *Pharmacogenomics J*, **1**(3), 167-170.

[58] Ashburner, M., et al. (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25-9.

[59] Barrell, D., et al. (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res*, **37**, D396-D403.

[60] Smith, B., et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, **25**(11), 1251–1255.

[61] Institute of Formal Ontology and Medical Information Science: http://www.ifomis.org/

[62] Bodenreider, O. and Stevens, R. (2006) Bio-ontologies: current trends and future directions. *Briefings in bioinformatics*, **7**(3), 256-274.

[63] The European Centre for Ontological Research: http://www.ecor.uni-saarland.de/

[64] National Centre for Biomedical Ontology: http://www.bioontology.org/

[65] Degtyarenko, K., et al. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, **36**(1), D344-D350.

[66] Malone, J., et al. (2010) Modeling Sample Variables with an Experimental Factor Ontology. *Bioinformatics*, **26**(8), 1112-1118.

[67] Brinkman, R. R., et al. (2010) Modeling biomedical experimental processes with OBI. *J Biomed Semantics*, **1**(1), S7.

[68] Basic Formal Ontology: http://www.ifomis.org/bfo

[69] Antezana, E., et al. (2009) The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process. *Genome Biol*, **10**(5), R58.

[70] Cytoscape: http://www.cytoscape.org/

[71] Köhler, J., et al. (2006) Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, **22**, 1383-90.

[72] Maere, S., et al. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**(16), 3448-3449.

[73] Smoot, M., et al. (2011) PiNGO: a Cytoscape plugin to find candidate genes in biological networks. *Bioinformatics*, **27**, 1030-1.

[74] Splendiani, A. (2008) RDFScape: Semantic Web meets systems biology. *BMC bioinformatics*, **9**(4), S6.

[75] Berners-Lee, T. and Hendler, J. (2001) Publishing on the semantic web. *Nature*, **410**, 1023-4.

[76] Duerst, M. and Suignard, M. (2005) Internationalized Resource Identifiers (IRIs). RFC 3987.

[77] Turtle syntax: http://www.w3.org/TR/turtle/

[78] Notation 3 syntax: http://www.w3.org/TeamSubmission/n3/

[79] N-Triples syntax: http://www.w3.org/TR/rdf-testcases/#ntriples

[80] JSON-LD: http://www.w3.org/TR/2013/CR-json-ld-20130910/

[81] Openlink Virtuoso: http://virtuoso.open-linksw.com/

[82] Apache Jena: http://jena.apache.org/

[83] 4store: www.4store.org

[84] Deus, H. F., et al. (2008) A Semantic Web management model for integrative biomedical informatics. *PloS one*, **3**(8), e2946.

[85] OWL 2 Profiles: http://www.w3.org/TR/owl2-profiles/

[86] Hoehndorf, R., et al. (2011) Integrating systems biology models and biomedical ontologies. *BMC systems biology*, **5**(1), 124.

[87] Hoehndorf, R., et al. (2012) Semantic integration of physiology phenotypes with an application to the Cellular Phenotype Ontology. *Bioinformatics*, **28**(13), 1783-1789.

[88] Horrocks, I., et al. (2004) SWRL: A Semantic Web Rule Language Combining OWL and RuleML, W3C Member Submission, W3C.

[89] Boley, H. and Wagner, G. (2001) Design rationale of RuleML: A markup language for semantic web rules," in *Proc. Semantic Web Working Symposium* (I. F. Cruz, S. Decker, J. Euzenat, and D. L. McGuinness, eds.), (Stanford University, California), 381–402.

[90] Knublauch, H., et al. (2011) SPIN - Overview and Motivation. W3C Member Submission, W3C.

[91] Belleau, F., et al. (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems.*Journal of biomedical informatics*, **41**(5), 706-716.

[92] Antezana, E., et al. (2009) BioGateway: a semantic systems biology tool for the life sciences. *BMC bioinformatics*, **10**(10), S11.

[93] Momtchev, V., et al. (2009) Expanding the pathway and interaction knowledge in linked life data. *Proc. of International Semantic Web Challenge*.

[94] Jupp, S., et al (2011) Developing a kidney and urinary pathway knowledge base. *Journal of biomedical semantics*, **2**(2), S7.

[95] Callahan, A., et al. (2011) HyQue: evaluating hypotheses using Semantic Web technologies. *Journal of biomedical semantics*, **2**(2), S3.

[96] Vladimir, M., et al. (2012) Gauging triple stores with actual biological data. *BMC Bioinformatics*, **13**(1), S3.

[97] EBI RDF Platform: http://www.ebi.ac.uk/rdf/

[98] Martin, D., et al. (2004) OWL-S: Semantic markup for web services. *W3C member submission*, **22**, 2007-04.

[99] Roman, D., et al. (2005) Web service modeling ontology. *Applied ontology*, **1**(1), 77-106.

[100] Gessler, D. G., et al. (2009) SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services. *BMC bioinformatics* **10**(1), 309.

[101] Wilkinson, M. D., et al. (2010) SADI, SHARE, and the in silico scientific method. *BMC Bioinformatics*, **11**(12), S7.

[102] Wollbrett, J., et al. (2013) Clever generation of rich SPARQL queries from annotated relational schema: application to Semantic Web Service creation for biological databases. *BMC bioinformatics*, **14**, 126.

[103] Bateman, A. and Wood, M. (2009) Cloud computing, *Bioinformatics*, **25**, 1475.

[104] Stein, L. D. (2010) The case for cloud computing in genome informatics. *Genome Biol*, **11**, 207.

[105] Chen, H., et al. (2012) Semantic web meets integrative biology: a survey. *Briefings in Bioinformatics*, **14**(1), 109-125

[106] Antezana, E., et al. (2009) BioGateway: a semantic systems biology tool for the life sciences. *BMC Bioinformatics,* **10**(10), S11.

[107] Semantic Systems Biology: http://www.semantic-systems-biology.org/

[108] HCLSIG: http://www.w3.org/wiki/HCLSIG/SysBio

[109] Semantic Web for Systems Biology: http://www.w3.org/wiki/HCLSIG/SysBio

[110] Giardine, B., et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, **15**, 1451-5.

Paper I

Chapter 2

Antezana *et al. BMC Bioinformatics* 2010, **11**(Suppl 12):S8
http://www.biomedcentral.com/1471-2105/11/S12/S8

BMC Bioinformatics

PROCEEDINGS **Open Access**

ONTO-ToolKit: enabling bio-ontology engineering via Galaxy

Erick Antezana[1*], Aravind Venkatesan[1], Chris Mungall[2], Vladimir Mironov[1], Martin Kuiper[1]

From The 11th Annual Bioinformatics Open Source Conference (BOSC) 2010
Boston, MA, USA. 9-10 July 2010

Abstract

Background: The biosciences increasingly face the challenge of integrating a wide variety of available data, information and knowledge in order to gain an understanding of biological systems. Data integration is supported by a diverse series of tools, but the lack of a consistent terminology to label these data still presents significant hurdles. As a consequence, much of the available biological data remains disconnected or worse: becomes misconnected. The need to address this terminology problem has spawned the building of a large number of bio-ontologies. OBOF, RDF and OWL are among the most used ontology formats to capture terms and relationships in the Life Sciences, opening the potential to use the Semantic Web to support data integration and further exploitation of integrated resources via automated retrieval and reasoning procedures.

Methods: We extended the Perl suite ONTO-PERL and functionally integrated it into the Galaxy platform. The resulting ONTO-ToolKit supports the analysis and handling of OBO-formatted ontologies via the Galaxy interface, and we demonstrated its functionality in different use cases that illustrate the flexibility to obtain sets of ontology terms that match specific search criteria.

Results: ONTO-ToolKit is available as a tool suite for Galaxy. Galaxy not only provides a user friendly interface allowing the interested biologist to manipulate OBO ontologies, it also opens up the possibility to perform further biological (and ontological) analyses by using other tools available within the Galaxy environment. Moreover, it provides tools to translate OBO-formatted ontologies into Semantic Web formats such as RDF and OWL.

Conclusions: ONTO-ToolKit reaches out to researchers in the biosciences, by providing a user-friendly way to analyse and manipulate ontologies. This type of functionality will become increasingly important given the wealth of information that is becoming available based on ontologies.

Background

Bio-ontologies are artefacts used to represent, build, store, and share knowledge about a biological domain by capturing the domain entities and their interrelationships. Bio-ontologies have become an important asset for the life sciences. They not only provide a controlled, standard terminology (to support annotations for instance); a variety of tools are available to exploit these ontologies, making them one of the cornerstones for biological data analysis. The Gene Ontology (GO) [1] is probably the best known bio-ontology. One of the most common uses of the GO is to perform term enrichment [2,3] on a gene set. The GO website lists over fifty such tools [4]. In addition, the life sciences community began to utilise other available ontologies (such as the Plant Ontology [5]) as well as to develop their own bio-ontologies to support other biology or technology domains. A recent example is the Ontology of Biomedical Investigations (OBI [6]), a community effort to build an ontology describing the different elements of a biomedical investigation (e.g. protocols, instruments, reagents, experimentalists). The Open Biomedical Ontologies (OBO) foundry [7] suggests a set of principles to guide the development of ontologies, for instance the

* Correspondence: erick.antezana@bio.ntnu.no
[1]Department of Biology, Norwegian University of Science and Technology (NTNU), Høgskoleringen 5, N-7491 Trondheim, Norway
Full list of author information is available at the end of the article

'orthogonality principle' designed to prevent overlapping ontologies. Most of the bio-ontologies gathered by the OBO foundry are represented in the OBO format [8], which has became the *lingua franca* to build bio-ontologies. An increasing number of bio-ontologies is being developed in the more expressive Web Ontology Language (OWL) that allows for advanced automated reasoning [9,10]. Automated reasoning, performed on OWL-formatted ontologies via the so-called reasoners (such as HermiT [11]), allows bio-ontologists to perform various tasks such as classification (also known as subsumption), which enables the process of making explicit the relations that were hidden (i.e. implicitly captured), and in general provides help to ensure the consistency of an ontology.

Several open source tools are available to deliver native support for bio-ontology manipulation (BioPerl [12], ONTO-PERL [13], BioRuby [14], BioPython [15]). We have previously published ONTO-PERL, a suite of Perl tools supporting the management of ontologies represented in OBO format (OBOF). ONTO-PERL is a full-blown API to manipulate bio-ontologies in OBOF. It offers a set of scripts supporting the typical ontology manipulation tasks, which can be used from the command line. Useful as this API may be for bioinformaticians or expert ontologists, biologists may find it intimidating to use. To accommodate their easy use, working with ontologies has for instance been facilitated by the setting up of ontology portals [16,17]. These applications can be directly linked to knowledge systems that store information in local infrastructures, thus taking advantage of the ontological scaffold (generally, hierarchical and partonomical relationships) through mappings between the ontology components (terms and relationships) and actual data. The linking of ontologies and biological data is proving to be a successful stepping stone towards ontology-based knowledge discovery platforms [18]. Those platforms may eventually become important tools in the quest for new hypotheses that can drive experimental design.

To further improve the repertoire of tools available to biologists to handle and analyze the knowledge available through ontologies we have turned to Galaxy [19], a web-based environment that integrates various types of tools to handle biological data. Galaxy's development is strongly targeted towards end-users who have limited computational skills (including many molecular biologists), so that they may easily perform analysis or have their favourite command line tool integrated. A tracking of the history of analyses, support for building workflows and data sharing are among Galaxy's most appealing features.

We used Galaxy to construct ONTO-ToolKit, which is an extension of the ONTO-PERL software that we developed previously. ONTO-PERL consists of a collection of Perl modules that enable the handling of OBO-formatted ontologies (like the Gene Ontology). With these modules a user can for instance manipulate ontology elements such as a Term, a Relationship and so forth, or employ scripts to carry out various typical tasks (such as format conversions between OBO and OWL (*obo2owl*, *owl2obo*).

ONTO-ToolKit allows exploiting the ONTO-PERL functionality within the Galaxy environment. Galaxy not only provides a user friendly interface to manipulate OBOF ontologies, it also offers the possibility to perform further biological (and ontological) analyses by using other tools provided within the Galaxy platform. In addition, ONTO-ToolKit provides tools to translate OBOF ontologies into Semantic Web formats such as RDF (Resource Description Framework) and OWL.

Methods

The functionalities of ONTO-PERL are enabled as tools in Galaxy through a set of tool configuration files (XML files), or 'wrappers'. These files contain execution details of the tool, e.g. path to the script, the arguments and the output format. Table 1 lists the functionalities provided by ONTO-PERL that are useful to understand the relationship between various biological components. The script *get_ancestor_terms.pl*, for instance, retrieves all the ancestor terms for a particular term id from a given OBO ontology. Furthermore, through *obo2owl.pl* and *obo2rdf.pl* scripts users can convert their data (OBOF) into OWL and RDF, respectively. A schematic representation of how ONTO-PERL is embedded as ONTO-ToolKit in Galaxy is given in Figure 1. A detailed description of installing ONTO-ToolKit is available at http://bitbucket.org/easr/onto-toolkit/wiki/Home.

Results

We illustrate the use of ONTO-ToolKit through three ontology-analysis use cases. In use case I we have analysed the relationship between terms from the Cell Cycle Ontology (CCO), an application ontology that we described previously [20]. In use case II we carried out an analysis combining ONTO-ToolKit functionality with other tools available in Galaxy, and in use case III we have demonstrated how a workflow was built to analyse gene sets with GO and *S. pombe* annotations.

Use case I: "Investigating similarities between given molecular functions"

The first use case illustrates the functionality of ONTO-ToolKit in identifying the ontology terms linking a pair of molecular function terms. A user might be interested to search for the most specific ancestor term that is shared by two molecular functions, to see if these

Table 1 Examples of ONTO-PERL functionalities

Scripts	Functionality
get_ancestor_terms.pl	Collects the ancestor terms (list of IDs) from a given term (existing ID) in the given OBO ontology.
get_child_terms.pl	Collects the child terms (list of term IDs and their names) from a given term (existing ID) in the given OBO ontology.
get_descendent_terms.pl	Collects the descendent terms (list of IDs) from a given term (existing ID) in the given OBO ontology.
get_subontology_from.pl	Extracts a sub-ontology (in OBO format) of a given ontology having the given term ID as the root.
get_obsolete_terms.pl	Finds all the obsolete terms in a given ontology.
get_parent_terms.pl	Collects the parent terms (list of term IDs and their names) from a given term (existing ID) in the given OBO ontology.
get_relationship_types.pl	Finds all the relationship types in a given ontology.
get_root_terms.pl	Finds all the root terms in a given ontology.
get_term_synonyms.pl	Finds all the synonyms of a given term name in an ontology.
get_terms.pl	Finds all the terms in a given ontology.
get_terms_by_name.pl	Finds all the terms in a given ontology that have a given string in their names.
obo2owl.pl	OBO to OWL translator.
obo2rdf.pl	OBO to RDF translator.
obo_trimming.pl	This script trims a given branch of OBO ontology.
obo2cco.pl	Converts an ontology into another one which could be integrated into CCO.
obo2tran.pl	OBOF into RDF translator. The resulting file has (full) transitive closure
obo2xml.pl	OBO to XML translator (CCO scheme).
go2owl.pl	Gene Ontology (in OBO) to OWL translator.
goa2rdf.pl	Generates a simple RDF graph from a given GOA file
owl2obo.pl	OWL to OBO translator.
obsolete_term_id_vs_def_in_go.pl	Obsolete terms vs. their definitions
obsolete_term_id_vs_name_in_go.pl	Obsolete terms vs. their names
term_id_vs_term_def.pl	Gets the term IDs and term definitions of a given ontology.
term_id_vs_term_name.pl	Gets the term IDs and term names of a given ontology.
term_id_vs_term_namespace.pl	Gets the term IDs and its namespaces in a given ontology
get_list_intersection_from.pl*	Collects common OBO terms from a given set of lists containing OBO terms
get_intersection_ontology_from.pl*	Provides an intersection of the given ontologies (in OBO format)

The left column lists ONTO-PERL scripts available in ONTO-ToolKit, with their functionality described in the right column. *: Scripts written specifically for ONTO-ToolKit and included in the ONTO-ToolKit download package.

functions fall into the same biological category. As a primary step all ancestor terms pertaining to the molecular function term IDs defined in a query are retrieved. In a next step a comparison is made between the two sets of ancestor terms for their relatedness. Figure 2 shows a schematic depiction of this use case, with retrieval of individual ancestor terms and checking for the most specific terms shared by the two molecular functions specified in the query. It is noteworthy that such a step will always result in a set of shared upper-level terms (as all molecular function terms are linked to the root), but obviously the relationship will be more specific if their shared terms are positioned further away from the root of the ontology, where information is more fine-grained. To implement this concept, the *S. pombe*-specific CCO was chosen along with the two molecular function term IDs (*CCO:F0000391* – 6-phosphofructokinase activity; *CCO:F0000759*–glucokinase activity). The analysis consisted of several steps. Firstly, using the *get_ancestor_terms* functionality two queries were used to fetch the ancestor terms for each of the two term IDs (see Figure 3). This resulted in two sets of ancestor terms and annotations associated with the terms. The intersection of these two sets was determined using the *get_list_intersection_from* function yielding one set of specific terms (see Figure 4) and corresponding annotations allowing the assessment of the relatedness of the initial terms.

Figure 5 shows the set of ancestor terms for the two terms of the query. For both the terms (*CCO:F0000391, CCO:*F0000759) ten ancestor terms were retrieved (see Supplementary file). Furthermore, the most specific common terms for the two molecular function term IDs were retrieved (see Figure 6). This list (Additional file 1) contained nine terms that were common, with various degrees of specificity, to both the molecular function

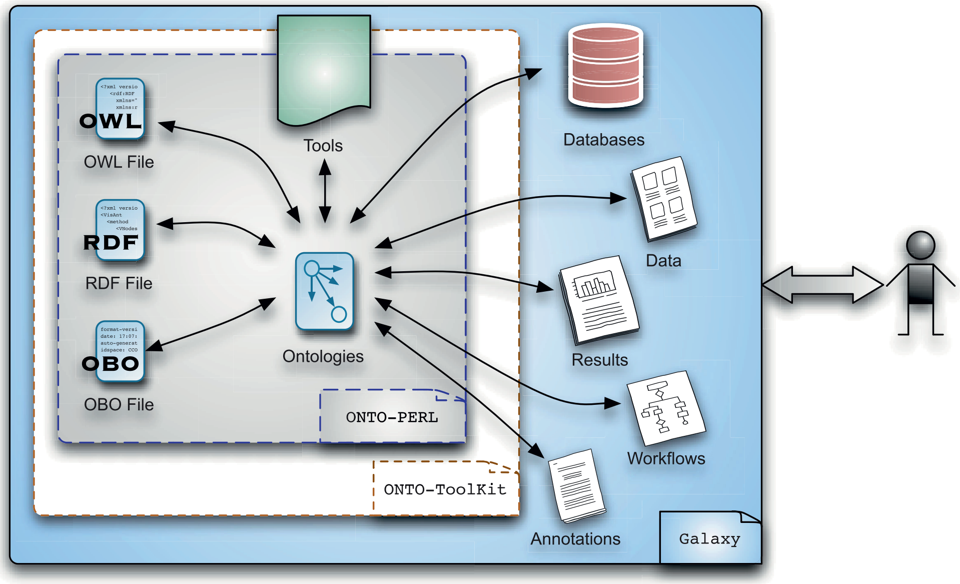


Figure 1 Schematic representation of ONTO-PERL, ONTO-ToolKit and Galaxy. The ONTO-ToolKit suite of tools provides a support within the Galaxy framework to analyse and manipulate OBO-formatted ontologies. ONTO-ToolKit relies on the functionality enabled by ONTO-PERL to handle bio-ontologies and to enable operations (such as format conversions from OBO to OWL) that could in turn produce results that might be further analysed and exploited through other tools (such as workflows or statistical analyses) provided in the Galaxy environment.

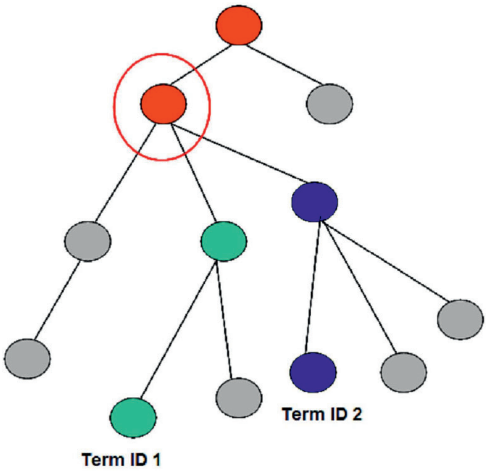
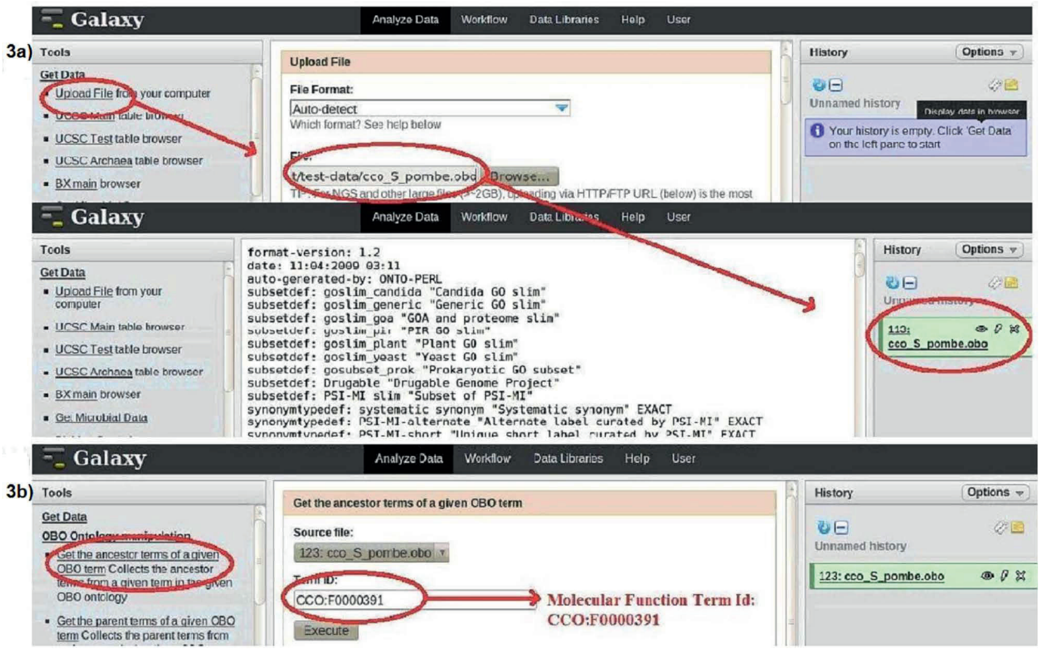


Figure 2 Schematic diagram of use case I The nodes and the edges represent a section of an ontology, with the higher nodes representing terms with general descriptions, and the nodes further down in the graph depicting terms with higher specificity. The nodes in green and blue represent the terms associated with the molecular function term id 1 and term id 2, respectively. The red nodes represent the terms shared by search terms, with the most specific term encircled in red.

terms. The most specific terms shared between them were: *CCO:F0004123 – carbohydrate kinase activity, CCO:F0003345 – phosphotransferase activity, CCO: F0003344 – transferase activity*. These results suggest that the two chosen terms are related, and additional ancestral terms make it clear that the two molecular function terms both describe functions of the glycolytic pathway in *S. pombe*.

Use case II: "Identifying shared terms for a pair of proteins"

Use case II illustrates how ONTO-ToolKit can be used in combination with other functionalities available in Galaxy. A user might be interested in identifying the functional relatedness of two proteins, as described by their GO annotations. To assess this, two lists of GO Terms associated with the two proteins need to be retrieved and then matched to determine their intersection. The example uses the *H. sapiens proteins JUN* (UniProt ID: P05412) and *FOS* (UniProt ID: P01100). Their UniProt IDs were used to query the BioMart [21] central server from within Galaxy to retrieve lists of *JUN* and *FOS* GO terms and annotations (see Additional file 1). In the second step, the ONTO-ToolKit function *get_list_intersection_from* was used to obtain all the


3a)
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Tools
Get Data
Upload File from your computer
UCSC Main table browser
UCSC Test table browser
UCSC Archaea table browser
BX main browser
Get Microbial Data
History
Options
Unnamed history
Display data in browser
113:
cco_S_pombe.obo
3b)
OBO Ontology manipulation
Get the ancestor terms of a given OBO term Collects the ancestor terms from a given term in the given OBO ontology
Get the parent terms of a given OBO term Collects the parent terms from
123: cco_S_pombe.obo
Molecular Function Term Id:
CCO:F0000391

Upload File

File Format: Auto-detect

Which format? See help below

File: t/test-data/cco_S_pombe.obo Browse...

TIP: For NGS and other large files (>~2GB), uploading via HTTP/FTP URL (below) is the most

Your history is empty. Click 'Get Data' on the left pane to start

```
format-version: 1.2
date: 11:04:2009 03:11
auto-gererated-by: ONTO-PERL
subsetdef: goslim_candida "Candida GO slim"
subsetdef: goslim_generic "Generic GO slim"
subsetdef: goslim_goa "GOA and proteome slim"
subsetdef: goslim_pir "PIR GO slim"
subsetdef: goslim_plant "Plant GO slim"
subsetdef: goslim_yeast "Yeast GO slim"
subsetdef: gosubset_prok "Prokaryotic GO subset"
subsetdef: Drugable "Drugable Genome Project"
subsetdef: PSI-MI slim "Subset of PSI-MI"
synonymtypedef: systematic synonym "Systematic synonym" EXACT
synonymtypedef: PSI-MI-alternate "Alternate label curated by PSI-MI" EXACT
synonymtypedef: PSI-MI-short "Unique short label curated by PSI-MI" EXACT
```

Get the ancestor terms of a given OBO term

Source file: 123: cco_S_pombe.obo

Term ID: CCO:F0000391

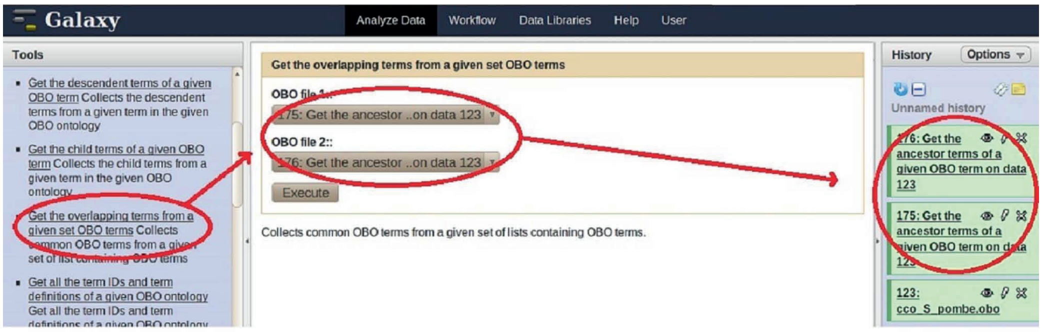
Execute

Figure 3 Screenshot of use case I implementation – step 1. Details of use case I analysis in the Galaxy user interface. **3a:** Method to upload the chosen obo ontology (CCO *S. pombe*). The uploaded ontology can be browsed, a feature available in Galaxy (encircled on the right); **3b:** Demonstration of the method to query the uploaded ontology using the *get_ancestor_term* function with the chosen term ID as the argument.

annotations shared between *JUN* and *FOS* (see Additional file 1). The results show the four GO terms (GO:0010843, GO:0070412, GO:0060395, GO:0007179) common between these two transcription factors.

Use case III: "Performing term enrichment using an ontology subset"

Use case III shows how ONTO-ToolKit can be used to create interdependent workflows (see Figure 7). Here a researcher may wish to analyze an *S. pombe* gene expression dataset using a subset of GO. The dataset contains a set of genes that have a high likelihood of being differentially expressed, and the researcher wants to know if this gene set has an overrepresentation of GO terms that are annotated to a specific biological process. As this type of analysis considers all GO terms sequentially, running this analysis on the whole GO may result in insignificant P-values due to the large


Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Tools
Get the descendent terms of a given OBO term Collects the descendent terms from a given term in the given OBO ontology
Get the child terms of a given OBO term Collects the child terms from a given term in the given OBO ontology
Get the overlapping terms from a given set OBO terms Collects common OBO terms from a given set of list containing OBO terms
Get all the term IDs and term definitions of a given OBO ontology Get all the term IDs and term definitions of a given OBO ontology
History
Options
Unnamed history
176: Get the ancestor terms of a given OBO term on data 123
175: Get the ancestor terms of a given OBO term on data 123
123: cco_S_pombe.obo

Get the overlapping terms from a given set OBO terms

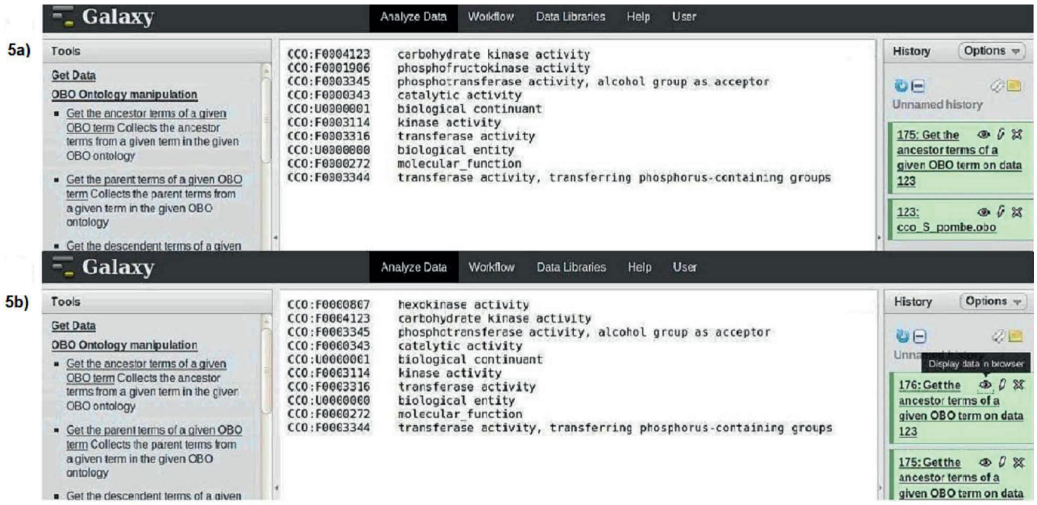
OBO file 1:: 175: Get the ancestor ..on data 123

OBO file 2:: 176: Get the ancestor ..on data 123

Execute

Collects common OBO terms from a given set of lists containing OBO terms.

Figure 4 Screenshot of step 2 in use case I Illustration of use case I, step 2. The Galaxy interface shows the use of the *get_overlapping_terms* function to intersect the two sets of terms obtained in step 1.


5a)
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Tools
Get Data
OBO Ontology manipulation
History
Options
Unnamed history
175: Get the ancestor terms of a given OBO term on data 123
123: cco_S_pombe.obo
5b)
Display data in browser
176: Get the ancestor terms of a given OBO term on data 123
175: Get the ancestor terms of a given OBO term on data

- Get the ancestor terms of a given OBO term Collects the ancestor terms from a given term in the given OBO ontology
- Get the parent terms of a given OBO term Collects the parent terms from a given term in the given OBO ontology
- Get the descendent terms of a given

```
CCO:F0004123    carbohydrate kinase activity
CCO:F0001906    phosphofructokinase activity
CCO:F0003345    phosphotransferase activity, alcohol group as acceptor
CCO:F0000343    catalytic activity
CCO:U0000001    biological continuant
CCO:F0003114    kinase activity
CCO:F0003316    transferase activity
CCO:U0000000    biological entity
CCO:F0000272    molecular_function
CCO:F0003344    transferase activity, transferring phosphorus-containing groups
```

```
CCO:F0000867    hexokinase activity
CCO:F0004123    carbohydrate kinase activity
CCO:F0003345    phosphotransferase activity, alcohol group as acceptor
CCO:F0000343    catalytic activity
CCO:U0000001    biological continuant
CCO:F0003114    kinase activity
CCO:F0003316    transferase activity
CCO:U0000000    biological entity
CCO:F0000272    molecular_function
CCO:F0003344    transferase activity, transferring phosphorus-containing groups
```

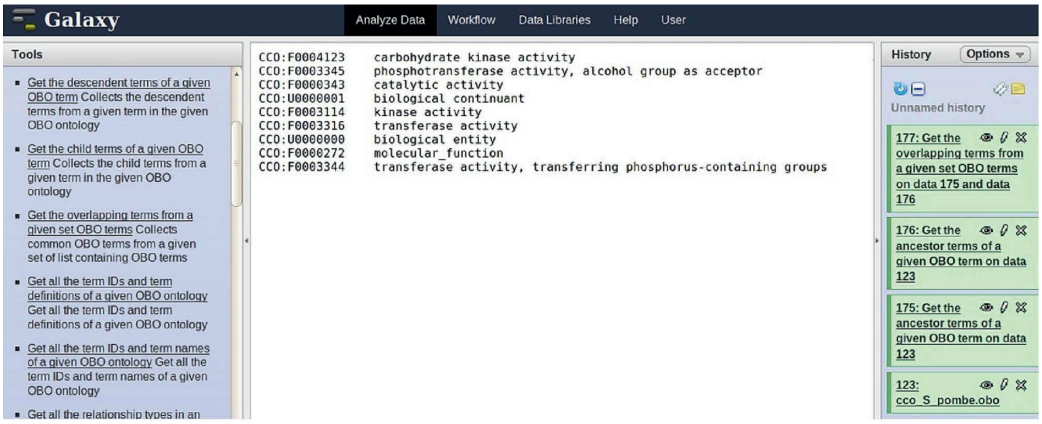
Figure 5 Ancestors term list – use case I Illustration of use case I, results. Panel 5a shows the results obtained for the term ID *CCO: F0000391*. Panel 5b shows ancestor terms for the term ID *CCO:F0000759*.

hypothesis space. This may be remedied by reducing this hypothesis space – for example, by considering only the role of these genes in the cell cycle.

This workflow starts by fetching an ontology and a set of gene associations, in this case, the Gene Ontology and the *S. pombe* annotations. The next step is to use the *get_descendant_terms* function (the converse of the *get_ancestor_terms* function described above) to extract a subset of the ontology (in this case, it is configured to extract all descendants of the term "cell cycle"). To get the corresponding annotations an annotation mapping function is used to get all annotations corresponding to this sub ontology. This cell cycle specific annotation file is fed into the GO TermFinder [3] enrichment tool, along with a user-supplied gene set. This workflow can be reused multiple times (for example, to re-check results with the latest ontology and annotations), and can be shared between Galaxy users.

Discussion

A coherent integration of public, online information resources is still a major bottleneck in the post-genomic era. Bioinformatics databases are especially difficult to


Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Tools
History
Options
Unnamed history
177: Get the overlapping terms from a given set OBO terms on data 175 and data 176
176: Get the ancestor terms of a given OBO term on data 123
175: Get the ancestor terms of a given OBO term on data 123
123: cco_S_pombe.obo

- Get the descendent terms of a given OBO term Collects the descendent terms from a given term in the given OBO ontology
- Get the child terms of a given OBO term Collects the child terms from a given term in the given OBO ontology
- Get the overlapping terms from a given set OBO terms Collects common OBO terms from a given set of list containing OBO terms
- Get all the term IDs and term definitions of a given OBO ontology Get all the term IDs and term definitions of a given OBO ontology
- Get all the term IDs and term names of a given OBO ontology Get all the term IDs and term names of a given OBO ontology
- Get all the relationship types in an

```
CCO:F0004123    carbohydrate kinase activity
CCO:F0003345    phosphotransferase activity, alcohol group as acceptor
CCO:F0000343    catalytic activity
CCO:U0000001    biological continuant
CCO:F0003114    kinase activity
CCO:F0003316    transferase activity
CCO:U0000000    biological entity
CCO:F0000272    molecular_function
CCO:F0003344    transferase activity, transferring phosphorus-containing groups
```

Figure 6 Intersection of ancestor terms – use case I Use case I results. The main panel shows the intersection of the two sets of ancestor terms of the terms of the query.

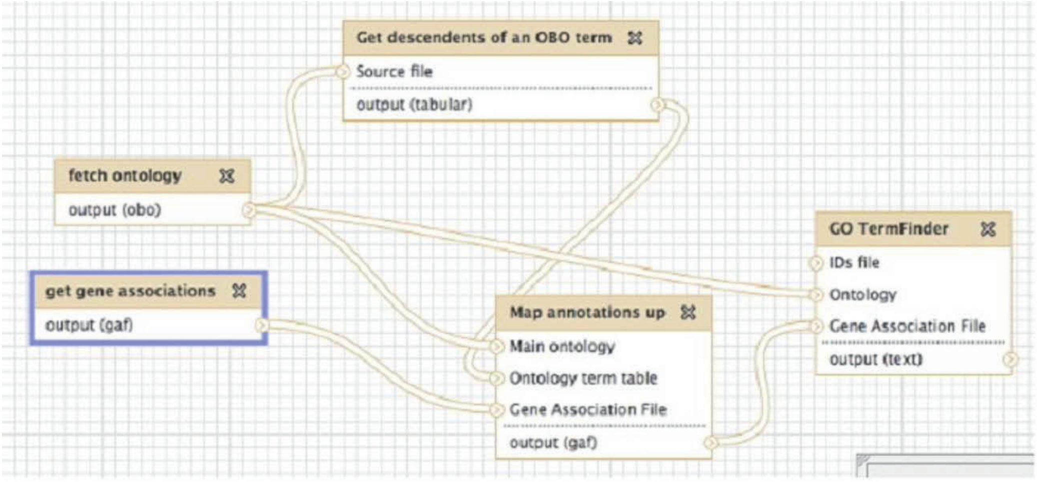


Figure 7 Example of workflow in Galaxy The boxes depict functions and intermediate workflow steps; the arrows indicate how these functions are connected.

integrate because they are often complex, highly heterogeneous, dispersed and incessantly evolving [22-24]. Moreover, consensus naming conventions and uniform data standards are often lacking. Nevertheless, the need for efficient procedures to integrate data is only increasing, due to the growing popularity of integrative biology and systems biology: approaches that need a variety of data from multiple sources to build computational models in order to understand biological systems behaviour.

Bio-ontologies can greatly facilitate this integration process [25] because they provide a scaffold that allows computers to automate parts or the whole of the integration process [26]. Setting up an integrative platform that can support an advanced data analysis based on bio-ontologies typically requires the establishment of an environment that enables access both to the many public biological databases that contain curated information, and to the various bio-ontologies. Moreover, such an integrative environment must enable the sharing of the information at any time with all contributors to the data curation process. In addition to curated databases, vast amounts of literature-independent data are being generated by high-throughput genome-wide analyses and accumulated in various databases. These databases represent another resource of context to infer biological function and to assess relations between biological entities. To obtain a powerful structuring and synthesis of all available biological knowledge it is essential to build an efficient information retrieval and management system. This system requires an extensive combination of data extraction methods, data format conversions, ontology-based analysis support and a variety of information sources. Ultimately, such an integrated and structured knowledge base may facilitate the use of computational reasoning for analysis of biological systems, an approach that we have named Semantic Systems Biology [26].

ONTO-ToolKit offers functionality that allows a biologist to exploit the increasingly abundant information supported by ontologies. The Gene Ontology Consortium is participating in the development of ONTO-ToolKit as an integration platform for performing many GO based workflows, replacing existing functionality in AmiGO [27] and expanding the range of tools to be used. For example, it is possible to extract all experimental annotations for the clade Mammalia, generate a slim (subset) from this set, or to fetch all annotations belonging to a pre-defined ontology subset. Annotations extracted in this way can also be used in term enrichment analyses using GO TermFinder [3]. Term enrichment analysis on ontology subsets reduces the number of terms that are considered for the overrepresentation analysis, making the analysis more sensitive.

Platforms such as Galaxy are aimed to overcome the barriers in global data processing, and its flexibility offers ample opportunity to identify and implement new ways to fill the gaps in data visualisation and analysis. We have explored Galaxy's use to implement data analysis techniques based on bio-ontologies. Bioinformatics data resources are constantly updated, *i.e.* by automated, software-mediated annotation or manual curation processes that depend on human intervention. Ontologies provide a means of improving the annotation process and to semantically represent the knowledge contained in biological databases in an unambiguous way. ONTO-ToolKit builds on this trend by enabling the manipulation of bio-ontologies within an integrative platform, which in turn allows analysis results to become the entry-point for further biological data analysis.

Conclusions

We presented several use cases to illustrate how the functionality of ONTO-PERL can be combined with the functionality of other tools in Galaxy. We have shown how the functionality of ONTO-PERL can be used to identify all the ancestor terms of a pair of ontology terms, or to simply retrieve all the terms shared by two proteins in order to assess their potential biological relatedness. We have extended and used ONTO-ToolKit to build a workflow to dynamically extract a subset of GO, map annotations to this subset, and then perform term enrichment analysis. With this we have shown that ONTO-ToolKit constitutes a useful extension to the functionalities available in Galaxy, by adding a variety of ontology-based analysis approaches that can improve the depth of the overall analysis because it builds on an increasing wealth of annotation and curation results.

Availability

ONTO-ToolKit can be obtained from its project page [28] or from the Galaxy Tool Shed [29]. ONTO-ToolKit is distributed under an Open Source License: GNU General Public License [30]. ONTO-ToolKit provides access to the latest obo2owl conversion code that implements the new proposed OBO Foundry mapping to OWL [31]. Once the ontology is converted to OWL, there are a number of OWL processing tools available, including Pellet [32], and ontology processing via the Thea library [33]. OntoToolkit, including the workflow example mentioned in use case III, is also available online [34].

Additional material

Additional file 1: This file contains all the additional results referred to in the description of the use cases I and II.**Subsection I:** Use case I - Lists the ancestor terms for CCO:F0000391.**Subsection II:** Use case I - Lists the ancestor terms for CCO:F0000759.**Subsection III:** Use case I - Lists the overlapping terms generates as part of step 2.**Subsection IV:** Use Case II - GO terms associated with JUN (Uniprot ID: P05412)**Subsection V:** Use Case II - GO terms associated with FOS (Uniprot ID: P01100)**Subsection VI:** Use Case II - Intersection of GO terms associated JUN and FOS

List of abbreviations used

OBO: Open Biomedical Ontologies; OBOF: Open Biomedical Ontologies Format; OBI: Ontology of Biomedical Investigations; GO: Gene Ontology; CCO: Cell Cycle Ontology; RDF: Resource Description Framework; OWL: Web Ontology Language; XML: eXtensible Markup Language.

Acknowledgements

We thank the Galaxy developers for the support they provided while implementing ONTO-ToolKit, and two anonymous reviewers who greatly helped to improve our manuscript

This article has been published as part of *BMC Bioinformatics* Volume 11 Supplement 12, 2010: Proceedings of the 11th Annual Bioinformatics Open Source Conference (BOSC) 2010. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/11?issue=S12.

Author details

[1]Department of Biology, Norwegian University of Science and Technology (NTNU), Høgskoleringen 5, N-7491 Trondheim, Norway. [2]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA.

Authors' contributions

EA implemented the ONTO-PERL extensions, the ONTO-ToolKit tools and steered the project. AV implemented use cases I and II. CM implemented the workflow example. VM and MK provided expertise in biological data management. All the authors have contributed to and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 21 December 2010

References

1. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al*: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32**:D258-261.
2. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks.** *Bioinformatics* 2005, **21**:3448-3449.
3. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes.** *Bioinformatics* 2004, **20**:3710-3715.
4. [http://www.geneontology.org/GO.tools.shtml#micro].
5. Jaiswal P, Avraham S, Ilic K, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Stevens P, Vincent L, Ware D, Zapata F: **Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages.** *Comp. Funct. Genomics* 2005, **6**:388-397.
6. [http://obi-ontology.org].
7. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nat Biotechnol* 2007, **25**:1251-1255.
8. [http://www.geneontology.org/GO.format.obo-1_4.shtml].
9. Blake JA, Bult CJ: **Beyond the data deluge: data integration and bio-ontologies.** *J Biomed Inform* 2006, **39**:314-320.
10. Wolstencroft K, Stevens R, Haarslev V: **Applying OWL Reasoning to Genomic Data.** New York: Springer;Semantic Web. Edited by Baker CJ, Cheung KH 2007:225-248.
11. [http://hermit-reasoner.com/].
12. [http://www.bioperl.org/wiki/Main_Page].
13. Antezana E, Egaña M, De Baets B, Kuiper M, Mironov V: **ONTO-PERL: an API for supporting the development and analysis of bio-ontologies.** *Bioinformatics* 2008, **24**:885-887.
14. Goto N, Prins P, Nakao M, Bonnal R, Aerts J, Katayama T: **BioRuby: Bioinformatics software for the Ruby programming language.** 2010, doi:10.1093/bioinformatics/btq475.
15. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ: **Biopython: freely available Python tools for computational molecular biology and bioinformatics.** *Bioinformatics* 2009, **25**:1422-1423.
16. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA: **BioPortal: ontologies and**

integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009, **37(Web Server issue)**:W170-173.

17. Côté R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H: **The Ontology Lookup Service: bigger and better.** *Nucleic Acids Res* 2010, **38(Suppl)**:W155-160.
18. Antezana E, Blondé W, Egaña M, Rutherford A, Stevens R, De Baets B, Mironov V, Kuiper M: **BioGateway: a semantic systems biology tool for the life sciences.** *BMC Bioinformatics* 2009, **10(Suppl)**:S11.
19. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A: **Galaxy: a platform for interactive large-scale genome analysis.** *Genome Res* 2005, **15**:1451-1455.
20. Antezana E, Egaña M, Blondé W, Illarramendi A, Bilbao I, De Baets B, Stevens R, Mironov V, Kuiper M: **The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process.** *Genome Biol* 2009, **10**:R58.
21. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart – biological queries made easy.** *BMC Genomics* 2009, **10**:22.
22. Cannata N, Merelli E, Altman RB: **Time to organize the bioinformatics resourceome.** *PLoS Comput Biol* 2005, **1**:e76.
23. Brooksbank C, Quackenbush J: **Data standards: a call to action.** *OMICS* 2005, **10**:94-99.
24. Philippi S, Kohler J: **Addressing the problems with life-science databases for traditional uses and systems biology.** *Nat Rev Genet* 2006, **7**:482-488.
25. Stevens R, Goble CA, Bechhofer S: **Ontology-based knowledge representation for bioinformatics.** *Brief Bioinform* 2000, **1**:398-414.
26. Antezana E, Kuiper M, Mironov V: **Biological knowledge management: the emerging role of the Semantic Web technologies.** *Brief Bioinform* 2009, **10**:392-407.
27. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, AmiGO Hub, Web Presence Working Group: **AmiGO: online access to ontology and annotation data.** *Bioinformatics* 2009, **25**:288-289.
28. [http://bitbucket.org/easr/onto-toolkit/wiki/Home].
29. [http://community.g2.bx.psu.edu/].
30. [http://dev.perl.org/licenses/gpl1.html].
31. [http://berkeleybop.org/~cjm/obo2owl/obo-syntax.html].
32. Sirin E, Parsia B, Grau B, Kalyanpur A, Katz Y: **Pellet: A practical OWL-DL reasoner.** *Web Semantics* 2007, **5**:51-53.
33. [http://www.semanticweb.gr/TheaOWLLib/].
34. [http://yuri.lbl.gov:8080/].

doi:10.1186/1471-2105-11-S12-S8
Cite this article as: Antezana *et al.*: **ONTO-ToolKit: enabling bio-ontology engineering via Galaxy.** *BMC Bioinformatics* 2010 **11**(Suppl 12):S8.

Paper II

Chapter 3

Bioinformatics Advance Access published April 5, 2011

Reasoning with bio-ontologies: using relational closure rules to enable practical querying

Ward Blondé [1,*], Vladimir Mironov [2], Aravind Venkatesan [2], Erick Antezana [2], Bernard De Baets [1] and Martin Kuiper [2]

[1]Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, 9000 Gent, Belgium
[2]Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

Associate Editor: Prof. Martin Bishop

ABSTRACT

Motivation: Ontologies have become indispensable in the Life Sciences for managing large amounts of knowledge. The use of logics in ontologies ranges from sound modelling to practical querying of that knowledge, thus adding a considerable value. We conceive reasoning on bio-ontologies as a semi-automated process in three steps: 1) defining a logic-based representation language; 2) building a consistent ontology using that language; and 3) exploiting the ontology through querying.

Results: Here, we report on how we have implemented this approach to reasoning on the OBO Foundry ontologies within BioGateway, a biological RDF knowledge base. By separating the three steps in a manual curation effort on Metarel, a vocabulary that specifies relation semantics, we were able to apply reasoning on a large scale. Starting from an initial 401 million triples, we inferred about 158 million knowledge statements that allow for a myriad of prospective queries, potentially leading to new hypotheses about for instance gene products, processes, interactions or diseases.

Availability: SPARUL code, a query endpoint and curated relation types in OBO Format, RDF and OWL 2 DL are freely available at http://www.semantic-systems-biology.org/metarel.

Contact: ward.blonde@ugent.be

1 INTRODUCTION

Life Sciences researchers become more and more acquainted with ontologies that support the management of knowledge in their research domains. Many initiatives on biomedical knowledge management have evolved into large Knowledge Bases (KB). Some of these consist of an ontology with a rich semantical content, like the Foundational Model of Anatomy (FMA) (Rosse *et al.*, 2003) —supporting anatomical aspects, SNOMED CT (Truran *et al.*, 2010) —for medical and clinical terms, the Gene Ontology (GO) (Ashburner *et al.*, 2000) —containing cellular information for gene description and the NCBI Taxonomy (Sayers *et al.*, 2010) —holding a classification of living organisms. Other KBs hold a large body of similarly formatted knowledge, like UniProt (UniProt Consortium, 2010) —collecting valuable information about proteins and the Gene Ontology Annotations (GOA) (Barrell *et al.*, 2009) —annotating gene products using the cellular information in GO. Public ontology repositories such as the BioPortal at NCBO (Noy *et al.*, 2009), the Ontology Lookup Service (OLS) (Cote *et al.*, 2008) and BioGateway (Antezana *et al.*, 2009a) make ontologies better accessible for scientists through visualisations, browse menus and search facilities.

Biologists are beginning to accept formal languages and ontologies as instruments to reach consensus while modelling the knowledge of their interest (Antezana *et al.*, 2009b). The large amounts of data that are generated with high-throughput methods call for such a framework (Taylor *et al.*, 2008). In general, a sound framework consists of a common syntax (the symbols and language constructs used), a common semantics (the meaning of the symbols) and common modelling practices (describing how to use the language). Ontology, a domain in philosophy that has long been trying to describe reality, often with the use of logics, is strongly stimulated by its fusion with computer science. Theories that have been developed for over 2000 years can now be applied in automated systems (Petrie, 2009). Since the last decade of the previous century, Description Logics (DL) have been developed as decidable fragments of First Order Logic, and some of them with the purpose of efficient reasoning (Baader *et al.*, 2003).

Another important evolution in computer science with respect to ontologies is the emergence of the Semantic Web and the use of Linked Data (Shadbolt *et al.*, 2006). The Semantic Web is viewed as a stack of languages and technologies that make knowledge and data on the internet computer intelligible. This stands in stark contrast to the current World Wide Web, which consists of human readable websites on the internet. The bottom layer of the Semantic Web stack consists of IRIs (Internationalized Resource Identifier, availabe at http://www.w3.org/2004/11/uri-iri-pressrelease) that can identify anything, like for instance biomedical concepts. On top of IRIs there is a layer dealing with the syntax, called XML (Extensible Markup Language, available at http://www.w3.org/XML/), in turn followed by a layer of RDF (Resource Description Framework, available at http://www.w3.org/RDF/), which is useful for querying and inferring graph-based representations (the linked data). Most of the KBs mentioned above are also provided in RDF. The top layer is OWL (Web Ontology Language, available at http://www.w3.org/TR/owl-features/), used for expressing the

[*]to whom correspondence should be addressed

meaning of knowledge and data (Horrocks, 2003). In October 2009, OWL upgraded to OWL 2, which distinguishes several DL-based sublanguages (profiles), like OWL 2 DL, OWL 2 EL and OWL 2 RL (OWL 2 Profiles, available at http://www.w3.org/TR/owl2-profiles/). By their RDF/XML syntax and by using IRIs, OWL ontologies are computer-manageable, syntactically sound and they provide an unambiguous meaning to well-identified concepts. All these technologies are the most important current standards at our disposal for the implementation of computer-readable ontologies.

Bio-ontologies are meant to be accessible to humans. The investment that biologists put in ontology development is driven by the need to build clear and sound models from the knowledge they have conceptualised collectively. Ontologists have the task to coordinate these efforts into a useful and manageable integrated framework. A consortium that has taken up this challenge is the OBO Foundry, fostering the Open Biomedical Ontologies (OBO) (Smith *et al*., 2007), which should all follow a set of 10 design principles (available at http://www.obofoundry.org/crit.shtml). So far 6 OBO ontologies (GO, CHEBI, PATO, PRO, XAO and ZFA) have been adopted by the OBO Foundry, while the others remain candidates under review. Many of the OBO ontologies that were developed in the more human-readable OBO Format have been translated into OWL. The Basic Formal Ontology (BFO) (Grenon *et al*., 2004) is developed as an upper level ontology that can integrate all the OBO ontologies. These scientific initiatives —OBO and BFO— are involved in the development of common modelling practices for ontologies across different scientific communities.

Query systems make bio-ontologies accessible to humans. OWL reasoners sustain such a query system on the Semantic Web, but they are very slow and demand too much memory to operate on large KBs, if they work at all. The SPARQL query language (SPARQL, available at http://www.w3.org/TR/rdf-sparql-query/) for RDF performs much better. RDF is perfectly suited for connecting large amounts of knowledge, however, it was not engineered for reasoning purposes.

With the construction of BioGateway (Antezana *et al*., 2009a) we have shown that biomedical resources (OBO ontologies, GOA annotations, UniProtKB/Swiss-Prot and the NCBI Taxonomy) can be interconnected in a single RDF store on the basis of common IRIs, and queried with SPARQL. BioGateway was given some minimal reasoning support for queries through Perl-operated inferences for transitivity of the *is_a* and *part_of* relation types in the OBO ontologies. In (Blondé *et al*., 2009) we presented Metarel, a controlled vocabulary for the semantics of relations in RDF, that is very well suited to create inference rules in conjunction with the RDF update language SPARQL/Update (SPARQL 1.1 Update, available at http://www.w3.org/TR/sparql11-update/). Metarel can provide a meaning to a relation between classes as a knowledge statement that takes the basic triple form subject-relation-object. It can be used by simply loading *metarel.rdf*, a meta-ontology for relations, together with KB-derived graphs in a single RDF store.

In this paper we show that semi-automated reasoning on bio-ontologies is possible for a set of closure rules in RDF, with the use of Metarel and SPARQL/Update. We augmented the query system behind BioGateway with inferences from these closure rules, thus further integrating the biomedical resources incorporated in BioGateway. Fully automated OWL reasoning, even on a single OBO ontology, is currently found challenging (e.g. the Sequence Ontology (Holford *et al*., 2010)) or even vexing (e.g. the Cell Cycle Ontology (Antezana *et al*., 2009)). By using Metarel as a representation framework for the logical reasoning, we were able to keep an RDF representation that is uncomplicated on both the syntactic level and on the semantic level.

2 DESCRIPTION LOGICS IN THREE STEPS

Description Logics research has kept the hope alive for realising fully automated reasoning on bio-ontologies. This research promises that any ontology that is modelled in a DL language makes unambiguous sense and that an automated reasoner can answer any logical question about the ontology correctly. However, applying the fully fledged reasoning approach on large, integrated bio-ontologies has proven to be overambitious for two main reasons. First of all, the developers of bio-ontologies, often more experienced in biology than computer science, do not succeed to model all the available knowledge into the rigid language constructs of logics. Consequently, bio-ontologies are full of glitches concerning their logic-based rules (Good and Wilkinson, 2006). Secondly, even if a large bio-ontology succeeds to pass the computational proof of consistency, current automated reasoners are not fast enough for answering queries. Although computer performance continues to increase, the amount of knowledge and data in bioinformatics has been growing even faster. Another hurdle is that being computationally consistent gives us no guarantee that the ontology is actually meaningful and correct.

Even an ontology with imperfections can be useful by providing sensible answers to many real life questions. In order to better exploit the available ontologies, we need an approach that benefits from DL as much as possible, without insisting on the exclusive use of DL at all stages in constructing a practical query system.

We approach the enabling of large-scale reasoning in three steps: 1) define a logic-based representation language; 2) build a consistent ontology; and 3) create inferences for enabling queries. DL reasoning is very useful in the first two steps and has proven already useful for consistency checking of smaller units. However, it is still problematic to implement DL in a query system on a very large scale.

We accomplished the third step for the ontologies in BioGateway by capitalizing on the prior work (by others and ourselves) with respect to the steps one and two. We minimally adjusted this prior work by a manual curation effort that was restricted to the relation types (types of relations like *is part of*, *is located in*) that were used. This curation effort implies a certain feedback from the last step to the previous steps. Certain language constructs, like defined classes, domains and ranges or number restrictions on relations, may turn out to be very expensive in terms of query time. Alternatively a relation type used in a given ontology may turn out to create masses of useless inferences. By using Metarel as a semantic framework, and SPARQL/Update as inference tool, we had ample flexibility to engage in a trial and error process to create only those inferences that were useful and necessary. This is a practical alternative to the ambitious approach of DL to execute reasoning as a one-step process without any flexibility for optimisation or feedback.

3 REASONING ON BIO-ONTOLOGIES

3.1 All-some relations between classes

Biological knowledge consists almost always of relations between classes (groups) of different individual biological entities. When we express knowledge about cells, proteins or organisms, for instance, we are not referring to a single cell that we observe under a microscope, or a particular mouse that was injected yesterday. We rather refer to classes of many entities that behave in similar ways, and these classes are what we name and identify. In comparison, for example the geographical knowledge domain is strikingly different, as the Atlantic Ocean, New York and Bermuda are large and significant enough to be referred to as individuals with a (usually capitalised) proper name and a proper identifier.

In queries about biological knowledge, we need a logical semantics for relations between classes. The all-some interpretation is the most prominent example to illustrate this (Smith *et al.*, 2005). When we relate the classes 'p53-protein' and 'tumour suppression' with the *has function* relation, it has to mean that all p53-proteins have some tumour suppression as function. This way of using relations provides a very powerful system to infer sound statements of biological knowledge.

Let us give an example using two statements that have the all-some interpretation: 'every p53-protein *is* some protein' and 'every protein *is encoded by* some gene'. From these two we can derive logically that 'every p53-protein *is encoded by* some gene'. The inferred statement is sound and it may be the basis for further conclusions.

All the millions of biological classes and the relations between them can be represented as a large network or graph. Queries can be constructed by defining a pattern or subgraph that must match one or several segments of the larger netwerk. Imagine we want to find all the objects that are encoded by a gene and that have some tumour suppression function. Then the search pattern will consist of two triples and one subject that we are interested in: 'my subject *is encoded by* gene' and 'my subject *has function* tumour suppression'. This pattern should match sections of the network, with the middle part of the triples fitting to the relations and the binding elements to the biological classes. The subject 'p53-protein' is a possible answer to the query.

We want to use this example to demonstrate the importance of reasoning. What happens if nobody bothered to add 'every p53-protein *is encoded by* some gene' explicitly? This absence would prohibit finding 'p53-protein' among the list of answers, although this statement follows logically from the two statements described above. It appears that in a good knowledge system all sound statements that are implicit should be made explicit by logical inference, thus augmenting the explicit knowledge in the system by pre-computing. A complete inference of implicit knowledge can be referred to as a 'closure'.

3.2 Five closure rules for inferring all-some relations

We propose five closure rules for inferring knowledge statements concerning relations between biological classes with an all-some interpretation. These five rules together provide the foundation for the reasoning in step 3 on the current state-of-the-art OBO ontologies and on annotations with OBO ontologies. Annotations of biological subjects imply that an ontology relation and an ontology term are used in the second and third parts of a knowledge statement that is represented as a triple.

Let A, B and C be classes and R, S and T be relation types. For instance, with A = 'p53-protein', R = '*is encoded by*' and B = 'gene', the knowledge statement $A\ R\ B$ means 'Every p53-protein *is encoded by* some gene'.

1. **Reflexivity**
   A reflexive closure infers the knowledge statements $A\ R\ A$, where R is a reflexive relation type. For instance, 'every body *is part of* some body'.
2. **Transitivity**
   A transitive closure infers the knowledge statements $A\ R\ C$, when the knowledge statements $A\ R\ B$ and $B\ R\ C$ exist and R is a transitive relation type. For instance, 'every kidney *is located in* some body' follows from 'every kidney *is located in* some abdomen' and 'every abdomen *is located in* some body'.
3. **Priority over subsumption**
   The priority over subsumption infers the knowledge statement $A\ R\ C$, if A is a subclass of B and the knowledge statement $B\ R\ C$ exists, or if the knowledge statement $A\ R\ B$ exists and B is a subclass of C. For instance, 'every API5-protein *regulates* some cell death' follows from 'every API5-protein *regulates* some apoptosis' and 'every apoptosis *is* some cell death'.
4. **Super-relations**
   A knowledge statement $A\ S\ B$ is inferred if S is a super-relation of R and the knowledge statement $A\ R\ B$ exists. For instance, 'every API5-protein *regulates* some apoptosis' follows from 'every API5-protein *negatively regulates* some apoptosis'.
5. **Chains**
   A knowledge statement $A\ R\ C$ is inferred if the knowledge statements $A\ S\ B$ and $B\ T\ C$ exist and R holds over a chain of S and T. The relation types R, S and T do not need to be all different. For instance, 'every API5-protein *negatively regulates* some apoptosis' follows from 'every API5-protein *participates in* some anti-apoptosis' and 'every anti-apoptosis *negatively regulates* some apoptosis'.

Such closure rules for all-some relations between classes follow directly from rules expressed for (chains of) relations between instances, which are common for Description Logics and OWL. Indeed, if all instances from class A are related to some instances of class B and all instances of class B are related to some instances of class C, then all instances of class A are connected by a chain of two instance relations to some instances of class C. The language features corresponding to the closure rules within step 3 (DL: chains as role constructors; global reflexivity for atomic roles, transitivity for atomic roles and role inclusions as role axioms; existential restrictions of atomic concepts by a role as concept constructors; and concept inclusions with atomic concepts on the left-hand side as terminological axioms) are a subset of those in OWL 2 EL and OWL 2 DL, which warrants efficient reasoning in a decidable semantics.

The semantics of OBO implies additional rules for inferring new knowledge statements. A prominent asset is the use of classes that are logically defined from primitive classes (DL: atomic concepts) through necessary and sufficient conditions. Such defined classes are used in most DL languages, like OWL DL, OWL EL and OWL RL. OBO ontologies have mostly primitive classes with natural language definitions, although logical definitions through intersections of classes are also used. However, the rules in step 3 treat an all-some relation between classes only as a necessary condition, which is not enough for a logical definition.

Other features in OBO that do not appear in step 3 are domains, ranges, symmetry, union, disjointness and functionality of relation types, and union and disjointness of classes. It is inherent to the idea introduced above (separating reasoning in three steps) that rules for these features were applied already in step 2 (building a consistent ontology).

Step 3 uses language features that can express knowledge more compactly (DL: logic entailment) and avoids the reasoning problems associated with consistency checking for logically defined classes (DL: satisfiability). If for instance the relation type *precedes* was given the range *process* and some annotator or ontology engineer erroneously creates a *precedes* relation to a logically defined class that is disjoint from *process*, then a reasoner should detect this problem in step 2.

An issue not mentioned here is the treatment of individuals (DL: assertional axioms), because they are currently not used in OBO ontologies, nor in the biomedical KBs that are annotated with OBO ontology classes. The individual geographical entities present in OBO's environmental ontology Gazetteer are modelled as singleton classes. In order to model and treat them as individuals, the five rules would need to be complemented with some extra rules. Inverses of relation types, which do not have logical consequences on all-some relations between classes, might also be of use in this extension.

4 METHODS

The large-scale inference of biological knowledge statements was achieved with RDF tools, operating on a merger of Metarel and BioGateway. The merging was relatively straightforward, as the ontologies in BioGateway consisted of the simple triple form subject-relation-object. We curated all the relation types that were used in the OBO ontologies, both candidates and adopted ones, assembling them in a relation ontology called *biorel.obo*. Subsequently, we translated *biorel.obo* into OWL 2 DL and merged it as an RDF graph with *metarel.rdf*. This resulted in the relation graph *biometarel.rdf* for use in BioGateway. Finally, we inferred new knowledge statements as RDF triples by running SPARQL/Update queries over both *biometarel.rdf* and the existing RDF graphs in BioGateway, thereby executing the above-described reasoning approach.

4.1 Manual curation of the relation types

Most relation types in BioGateway originate from the OBO ontologies. All OBO ontologies exist in BioGateway as RDF graphs, providing the opportunity to transform the relational information available in BioGateway with RDF tools. However, standard RDF conversion tools do not properly translate all information embedded in OBO ontologies to RDF, so the work was initiated with the original OBO files for a more expressive translation. All Typedef sections for relation types were separated from all OBO files and through a process of manual curation this long list was reduced to a single valid, consistent OBO file. Text sorting operations and spreadsheets were used to compare and select the best annotated and authoritative relation type entries among the duplicates. In this manner 833 relation type entries were reduced in a consistent, single-person effort to 365 unique, curated relation types. The resulting OBO file, *biorel.obo*, is available for download at http://www.semantic-systems-biology.org/metarel/biorel.

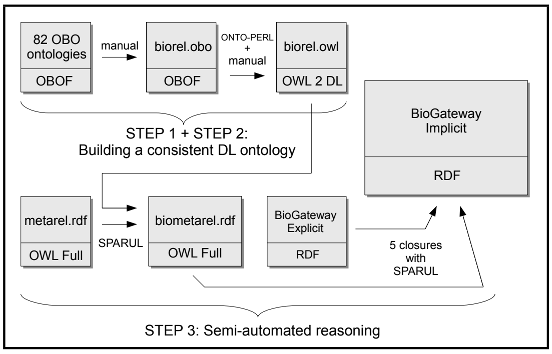


Fig. 1. A practical implementation of the three-step process for reasoning with bio-ontologies through management of relation semantics. A consistent, validated *biorel.owl* in OWL 2 DL contains all the relation types. It is the starting point for applying 5 important closure rules with a basic RDF tool like SPARQL/Update (SPARUL).

The most crucial step in our curation process, central to the decision of using the Metarel/RDF framework, was to make a consistent interpretation of the relation types as either object properties (relation types between instances) or all-some relation types between classes. Every relation type used between two terms in an OBO file was interpreted as a metarel:AllSomeClassRelationType and the corresponding relation type in the Typedef section as an owl:ObjectProperty. Relation types that were annotated as 'metadata-tags' in the Typedef section were always interpreted as an owl:AnnotationProperty. This interpretation is entirely consistent with the current standardised practices of translation between OBO and OWL DL (OboInOwl) (Aitken *et al.*, 2008), but it is not consistent with the original interpretation that is still commonly held by many OBO ontology developers. Judging from definitions and tags in OBO's Typedef section, the relation types are most often still viewed as class relation types.

As a consequence, some tags that were introduced in an ad hoc manner to solve this ambiguity, like 'inverse_of_at_instance_level' and 'instance_level_is_transitive', were replaced by the standard variants that are captured better by OBO translation tools.

Six relation types, like *has taxonomic rank*, *is valid for taxon* and *is extinct*, have OBO's metadata-tag because they cannot be given an interpretation as object properties. We added this tag to *is integral part of* for the same reason. Its semantics is clear as it can always be written as a combination of the two all-some relation types *is part of* and *has part* in opposite directions: if 'A *is integral part of* B', then every A *is part of* some B and every B *has part* some A. It is interpreted as a *metarel:InvertibleRelationType*, which enables some additional closure rules. However, it does not fit in the general system and it needs to be translated to an annotation property for validation in OWL 2 DL.

A consistent naming system was created, by giving every relation type a name that contains a verb in the third person singular. For instance the name *after* was replaced by *exists after*, as this seemed to be the intended meaning.

Rules that were only poorly formulated as informal comments were upgraded to sound logic. For instance, the comment that any *starts at end of* implies an *is preceded by* was easily translated to OBO's logic by modelling the former as a subproperty of the latter. The new OBO tag 'holds_over_chain' for creating property chains was exploited to its fullest extent and it was added in several cases. For instance, *is directly preceded by* holds over a chain of *has start* and *is end of*. The 'transitive_over' tag became superfluous through the use of 'holds_over_chain'.

One informally asserted rule stated: "Gt influences P & Gt variant_of G => G influences P". This is a chain rule with one object property in the inverse direction. Interpreting the object properties as all-some relation types between classes we will have *is variant of* from the some-side to the all-side, which does not result in a sound rule on the class level. Indeed, as every Gt is a variant of some entity, it would follow that every entity (everything) influences some P. Implementing such a rule for classes would corrupt the whole knowledge base. The rule was translated to a formal chain rule by using an inverse relation. As no inverse was tagged for *is variant of*, the following choice was made: *is influenced by* holds over a chain of *is influenced by* and *is variant of*. This chain goes from the all-side to the some-side and it retains the intended semantics.

Transitivity was added for all the object properties that were tagged as the inverse of a transitive object property. For instance *is preceded by* was provided with transitivity by the transitivity of *precedes*.

Apart from the names, some dozens of OBO tags for the semantics of relation types had to be altered. Contradictions were nowhere found and the intended semantics could always be retrieved by informal comments and by the way the relation types were actually used in the ontologies.

4.2 Translation to the Semantic Web

The use of the available Semantic Web tools for inferring and querying requires a translation to a Semantic Web language. The current standards are OWL and RDF. BioGateway, an RDF store, does not contain any of the OWL profiles. By using Metarel/RDF as a target framework for the translation, we are, however, still using the standards, because Metarel is valid OWL Full and apart from using class relation types, Metarel is fully compatible with the language constructs used in OWL. Moreover, Metarel being valid OWL Full is technically equivalent with it being valid RDF (http://www.w3.org/TR/owl-ref/). Unlike OWL Full, however, Metarel can connect the class relation types that reside in the RDF of BioGateway with the object properties in Biorel.

We translated *biorel.obo* first to *biorel.owl* using ONTO-PERL (Antezana *et al*., 2008) and adjusted the translated file with some manual curation, which resulted in a valid OWL 2 DL ontology file for the relation types. We added also the chains of relation types, a feature novel in OWL 2 that is in the process of being included in the OboInOwl translation standard. In principle, *biorel.owl* should contain all the expressivity for the rules in Section 3.2, even in RDF.

For our purposes, *biorel.owl* was not practically useful yet, because it contains only object properties, while BioGateway contains only class relation types. We uploaded *biorel.owl* into the relation meta-graph *metarel.rdf* alongside the other RDF ontologies in BioGateway. The merged graph is called *biometarel* and it is this graph that is used for the reasoning process. With SPARUL updates in biometarel we could connect the object properties of biorel with the class relation types of BioGateway and propagate the semantic rules, like transitivity and chains, to the level of classes.

4.3 Inferring new knowledge statements

Each of the five rules that are required for a query system, as discussed in Section 3.2, corresponds to a single SPARUL/Update query type (see Figure 2) . These update queries range over biometarel and the ontology graphs in BioGateway. They need to be operated in a recursive loop until there is no new knowledge statement left that can be inferred.

This practice implies that the entailment of the inferred triples is fully materialised on a hard disk and when this is executed on BioGateway with OpenLink Virtuoso 5.0.8, it shows an acceptable performance. It takes approximately 20 hours to produce 158 million inferred knowledge statements, which is reasonable compared to an uploading time of 5 hours for the 401 million original triples. We would like to point out that not all triples are knowledge statements with a meaningful relation type as a connector. Many triples are used for asserting names, synonyms, definitions, textual annotations, literature references, etc. As these triples are often more verbose, we will call them *verbose triples* as opposed to knowledge statements. For the Gene Ontology we get the following numbers: 54 718 explicit knowledge statements, 643 384 verbose triples and 2 031 247 newly inferred knowledge statements. This implies a multiplication factor of 38,12 for the number of knowledge statements, but a multiplication factor of only 3,90 for the total number of triples. For the complete BioGateway we have a multiplication factor of only 1,39.

```
BASE   <http://www.semantic-systems-biology.org/>
PREFIX metarel:<http://www.metarel.org/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

INSERT INTO GRAPH <human_disease_tc> {
  ?Class ?RelType ?Class.
}
WHERE {
  GRAPH <human_disease_tc> {
    ?Class rdf:type ?Type.
  }
  GRAPH <biometarel> {
    ?RelType rdf:type metarel:ReflexiveRelationType.
    ?RelType rdf:type metarel:ClassRelationType.
  }
}
```

Fig. 2. The update query for inferring reflexive relations in the human disease graph. A PERL script parametrises the 5 SPARQL/Update query types with about 2000 graph names. All the inferences from such small graphs are merged later in the large SSB_tc graph through other update queries. Contrary to OWL, Metarel uses reflexivity and relation types that fit directly between classes.

The relatively low multiplication factor and the high percentage of verbose triples clearly show that a full materialisation does not pose many extra storage-related problems for bio-ontologies. It makes no sense to start the reasoning process in a temporal memory only after a query is launched. It takes 20 hours to generate all the informative knowledge statements, whereas the majority of typical biologically relevant queries take no more than some seconds to produce an answer. A quick response is absolutely required in the knowledge exploration phase that precedes a more systematic investigation of a new hypothesis. Therefore the materialisation of inferred triples is the preferred practice for bio-ontologies.

5 RESULTS

We inferred about 158 million knowledge statements through semi-automated reasoning within BioGateway. The inferences are almost always sound within the intuitive system of all-some relations (an exception: plural forms in ‘Every *Mammalia* is some *cellular organisms*’, following from an incompatible system for defining terms in the NCBI Taxonomy) and they can be accessed directly for any term through BioGateway’s most basic lookup query (results for API5 in supplementary material). Most of the inferences are rather trivial if they are considered as a single statement, however, their effect becomes clear to those who are querying the knowledge base. Without the inferences, certain queries either simply return only a fraction of the answers potentially available in the knowledge sources, or they require a lot of very specific knowledge on the architecture of the ontologies in the KB to retrieve full results. The reasoning process would need to be done by the query builder and the resulting queries would become huge and slow.

By precomputing all the inferences, the hardest part of the reasoning process happens only once in a single although substantial computational effort, but the results are stored and are available for all subsequent queries. The query builder can now concentrate solely on the intended meaning of the relation types that are used, instead of reconstructing this meaning by his query. He can query over the explicit knowledge statements as well as over the implicit ones.

Queries on the RDF graphs with inferred knowledge statements are now short and the answers are more informative and complete. Imagine a cancer researcher who investigates the ASPP1-proteins (Apoptosis stimulating of p53-protein 1) and she finds in direct manual annotations from GOA that proteins of this type are located in the nucleus and in the cytoplasm and that they participate in the processes 'induction of apoptosis' and 'negative regulation of cell cycle'. Now she would like to see which other proteins fulfil these conditions within mammals.

This query involves just a pattern of knowledge statements in the triple form:

- my_subject *is located in* cytoplasm
- my_subject *participates in* apoptosis
- my_subject *participates in* negative regulation of cell cycle
- my_subject *has source* mammal

The query will return all the classes of biological entities (proteins in this case) that can, within BioGateway, be inferred to be located in the cytoplasm, to participate in apoptosis, etc., instead of searching only through the knowledge statements that were once annotated explicitly by an ontology engineer or a manual curator. 'Mammal' is too generic to be chosen as an annotation for a source species. Also 'cytoplasm', 'apoptosis' and 'negative regulation of cell cycle' have many sublocations and subprocesses that are often choosen for annotations. The query returns 36 types of proteins that actually fulfil all the conditions, but just 29, 29 and 17 respectively if only the explicit annotations are queried for 'cytoplasm', 'apoptosis' and 'negative regulation of cell cycle'. We still get 13 protein types for explicit annotations on all these three conditions, but none for direct annotations on 'mammal'. The relatively high numbers of explicit annotations are due to the fact that they are abundant and redundant in meaning, though taken together still incomplete.

We investigated the necessity of each of the five closure rules separately by recreating the inferred version of BioGateway five times and omitting one of the rules during each recreation. We detected that many inferences followed from several different closure rules, however, for all 5 recreations of the implicit KB, some of the inferences were missing. Four specific biological queries in BioGateway illustrate the practical relevance of each of the closure rules:

- **Query 1:** Which are all the biological processes in which a given protein (dnaJ in Chlamydophila felis Fe/C-56) is involved, which are all the other proteins that participate in these biological processes and which cellular locations were annotated for these other proteins? (Bio4 in BioGateway)

- **Query 2:** Which are the proteins that have both the nucleus and the endoplasmatic reticulum as inferred locations, compared and ordered for all the organisms in the KB? (Bio5 in BioGateway)

- **Query 3:** What are the subparts of liver parenchyma?

- **Query 4:** Which are the developmental stages preceding the unfertilised egg stage, and that are themselves preceded by oogenesis stage S6 (the stage during which follicle cell division ceases)?

The SPARQL translations of these queries and the answers to the queries can be found as supplementary material to this paper. For each query we counted the number of answers rendered by either the KB with only the explicit knowledge, on the KB with all the additional inferred knowledge and in each of the KBs that lacked one specific type of closure. The results can be viewed in Table 1.

To demonstrate that the additional answers also make biological sense we will analyse the queries and the corresponding parts of the KB. Query 1 asks for proteins that are involved in the same biological process as a given process. This means that a protein involved in a subprocess is also a good answer. The query asks generically for proteins in the same process and/or the subprocesses, but without the reflexive closure proteins annotated with the exact same process are disregarded (79 answers). Without any closure we get only the proteins annotated with the subprocesses on the level immediately below the original process, but not subprocesses of subprocesses (2 answers). Query 2 fails to return proteins that are annotated with sublocations of the nucleus and the endoplasmic reticulum when either the priority over *is_a* or the chain closure is omitted. This depends on the particular engineering of the Gene Ontology. We find almost exclusively the *is_a* relation type below the nucleus and the endoplasmic reticulum, with only nuclear part *is part of* nucleus and endoplasmic reticulum part *is part of* endoplasmic reticulum, for instance 'germ cell nucleus *is a* nucleus' and 'ARC complex *is a* nuclear part'. The priority over *is_a* propagates *is located in* over all these *is_a*'s, but we need a specific chain rule to propagate *is located in* over *is part of*. The priority over *is_a* generates extra answers for annotations with terms like 'germ cell nucleus' (616 answers). But as no protein was annotated with nuclear part nor endoplasmic reticulum part, we get only the explicit annotations on nucleus and endoplasmic reticulum if the priority over *is_a* is omitted (593 answers). Only if both the chain closure and the priority over *is_a* are in place, proteins with annotations in the hierarchy below nuclear part and endoplasmic reticulum part are retrieved (738 answers). Query 3 requires the transitive closure of *is part of* for finding the subparts of 'liver parenchyma'. Without any closures only 'liver lobule', 'portal lobule' and 'portal triad' are retrieved (3 answers), but not the 6 more specific terms like 'bile canaliculus', which are subparts of the liver lobule and the portal triad. Reflexivity acknowledges that a liver parenchyma is also part of itself (4 answers). Query 4 asks for a series of developmental stages. The ontology of developmental stages uses *starts at end of* as a relation type to connect subsequent stages. *Starts at end of* is a subrelation of the transitive relation type *is preceded by*. That is why only answers are found if both the transitive closure and the super-relation closure are implemented (19 answers).

The results show that every query executed in BioGateway that uses any of the 365 relation types in *biorel.obo* benefits from the reasoning process that has created the inferences. The answers to such a query are complete and they correspond to the logical meaning of the relation types as intended by the ontology engineers. This meaning no longer needs to be simulated in the queries.

	Exp.	Imp.	R1	R2	R3	R4	R5
Query 1	**2**	118	**79**	118	118	118	118
Query 2	**593**	738	738	738	**593**	738	**616**
Query 3	**3**	10	**9**	**4**	10	10	10
Query 4	**0**	19	19	**0**	19	**0**	19

Table 1. The number of answers to queries compared on explicit knowledge (Exp.) and implicit knowledge (Imp.), and on partial closures where reflexivity (R1), transitivity (R2), priority over subsumption (R3), super-relations (R4) and chains (R5) were omitted respectively.

6 DISCUSSION

Bio-ontologies and the Semantic Web are two important evolutions for knowledge management in the Life Sciences. They provide a logical framework, universal identifiers and tools for the integration of knowledge. However, in order to become really useful for Life Sciences researchers, both pillars need to mature further.

As the amount of biomedical knowledge keeps growing exponentially, the scalability of Semantic Web tools should be a main concern. Slow queries and memory overflows form a real obstacle for the exploitation of KBs. With this work, we have chosen to enable efficient querying with the most basic semantic features, instead of hampering the query system with advanced, fully automated reasoning.

The large and diverse possibilities of querying RDF demands better browsing and visualising tools to make the technology more accessible to biologists. Some specific tools for browsing BioGateway are under construction (available at http://www.semantic-systems-biology.org/biogateway/sparql-viewer), but parameterizing and reworking the SPARQL code is still the best option for acquiring all the expressivity of the SPARQL query language. However, the direct relations between classes used in BioGateway may help overcome some of the current shortcomings pertaining to reasoning and browsing.

On the side of the development of bio-ontologies, more efforts are required: ontology engineers should reuse other bio-ontologies to avoid duplication, create appropriate relations, provide identifiers, synonyms, definitions and cross-references. The Semantic Web architecture is perfectly suited for exploiting an orthogonal, cross-linked set of bio-ontologies. BioGateway, with logical inferences in place, can be used to identify the glitches in OBO ontologies.

As a future work, we plan to include more data in BioGateway, like biological pathways, and to test the biological usefulness of the inferences with in-depth queries on very specific research questions. For example the inferences on GO and the NCBI Taxonomy will allow to compare gene functions across species and kingdoms.

7 CONCLUSION

Many different ontology engineers have collaborated in the coordinated development of more than 80 OBO ontology files. We have brought consistency to the stack of relation types in these files by gathering all the relation types in Biorel and translating them to OWL 2 DL. After merging the OWL translation of Biorel with Metarel in the RDF store BioGateway, we could infer 158 million previously hidden knowledge statements from the explicitly asserted knowledge in the OBO ontologies, GOA annotations for about 2000 species, UniProtKB/Swiss-Prot and the NCBI Taxonomy. The inferred knowledge statements can be used for biological hypothesis generation through querying. The success of our methodology is due to the soundness of OBO ontologies, the use of Semantic Web tools and the semi-automated approach of reasoning.

Our work shows that a small set of simple rules for bio-ontologies results in efficient practices for reasoning and querying. As many researchers are involved in building bio-ontologies, more restrictive guidelines and principles for building bio-ontologies are required in order to obtain more uniformity and reach more convergence for knowledge management in the Life Sciences.

REFERENCES

Aitken,S., Chen,Y., Bard,J. *et al.* (2008) OBO Explorer: an editor for open biomedical ontologies in OWL, *Bioinformatics*, **24(3)**, 443-444.

Antezana,E., Blondé,W., Egaña,M. *et al.* (2009) BioGateway: a semantic systems biology tool for the life sciences, *BMC Bioinformatics*, **10(10)**, S11.

Antezana,E., Egaña,M., Blondé,W. *et al.* (2009) The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process, *Genome Biol.*, **10(5)**, R58.

Antezana,E., Kuiper,M., Mironov,V. (2009) Biological knowledge management: the emerging role of the Semantic Web technologies, *Brief. Bioinform.*, **10(4)**, 392-407.

Antezana,E., Egaña,M., De Baets,B. *et al.* (2008) ONTO-PERL: An API for supporting the development and analysis of bio-ontologies, *Bioinformatics*, **24(6)**, 885-887.

Ashburner,M. *et al.* and Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology, *Nat. Genet.*, **25(1)**, 25-29.

Baader,F. *et al.* (2003) The Description Logic Handbook: Theory, Implementation, and Applications, *Cambridge University Press*.

Barrell,D., Dimmer,E., Huntley,R. *et al.* (2009) The GOA database in 2009-an integrated Gene Ontology Annotation resource, *Nucleic Acids Res.*, **37(Sp. Iss. SI)**, D396-D403.

Blondé,W., Antezana,E., De Baets,B. *et al.* (2009) Metarel: an Ontology to support the inferencing of Semantic Web relations within Biomedical Ontologies, *Proceedings of the International Conference on Biomedical Ontologies (ICBO)*, 79-82.

Cote,R. *et al.* (2008) The Ontology Lookup Service: more data and better tools for controlled vocabulary queries, *Nucleic Acids Res.*, **36(S)**, W372-W376.

Good,B., Wilkinson,D. (2006) The Life Sciences Semantic Web is full of creeps!, *Brief. Bioinform.*, **7(3)**, 275-286.

Grenon,P., Smith,B., Goldberg,L. (2004) Biodynamic ontology: Applying BFO in the biomedical domain, *Studies in Health Technology and Informatics*, **102** of *Ontologies in Medicine*, 20-38.

Holford,M. *et al.* (2010) Using semantic web rules to reason on an ontology of pseudogenes, *Bioinformatics*, **26(12)**, i71-i78.

Horrocks,I., *et al.* (2003) From SHIQ and RDF to OWL: The making of a web ontology language, *J. Web Semant.*, **1(1)**, 7-26.

Noy,N., Shah,N., Whetzel,P. *et al.* (2009) BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucleic Acids Res.*, **37(S)**, W170-W173.

Petrie,C. (2009) The Semantics of “Semantics”, *IEEE Internet Comput.*, **13(5)**, 94-96.

Rosse,C., Mejino,J. (2003) A reference ontology for biomedical informatics: the Foundational Model of Anatomy, *J. Biomed. Inform.*, **36(6)**, 478-500.

Sayers,E. *et al.* (2010) Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res.*, **38**, D5-D16.

Shadbolt,N., Hall,W., Berners-Lee,T. (2006) The Semantic Web revisited, *IEEE Intell. Syst.*, **21(3)**, 96-101.

Smith,B. *et al.* (2005) Relations in biomedical ontologies, *Genome Biol.*, **6(5)**, R46.

Smith,B., Ashburner,M., Rosse,C. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nat. Biotechnol.*, **25(11)**, 1251-1255.

Taylor,F. *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project, *Nat. Biotechnol.*, **26(8)**, 889-896.

The UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010, *Nucleic Acids Res.*, **38**, D142-D148.

Truran,D. *et al.* (2010) SNOMED CT and its place in health information management practice, *Health. Inf. Manag. J.*, **39(2)**, 37-39.

Paper III

Chapter 4

Towards an integrated knowledge system for capturing gene expression events

Aravind Venkatesan[†] Vladimir Mironov[†*] and Martin Kuiper
Department of Biology, NTNU, 7491 Trondheim, Norway

ABSTRACT

Transcriptional regulation of gene expression is an important mechanism in many biological processes. Aberrations in this mechanism have been implicated in cancer and other diseases. Effective investigation of gene expression mechanisms requires a system-wide integration and assessment of all available knowledge of the underlying molecular networks. This calls for a method that effectively manages and integrates the available data. We have built a semantic web based knowledge system that constitutes a significant step in this direction: the Gene Expression Knowledge Base (GeXKB). The GeXKB encompasses three application ontologies: the Gene Expression Ontology (GeXO), the Regulation of Gene Expression Ontology (ReXO), and the Regulation of Transcription Ontology (ReTO). These three ontologies, respectively, integrate gene expression information that is increasingly more specific, yet decreasing in coverage, from a variety of sources. The system is capable of answering complex biological questions with respect to gene expression and in this way facilitates the formulation or assessment of new hypothesis. Here we discuss the architecture of these ontologies and the data integration process and provide examples demonstrating the utility thereof. The knowledge base is freely available for download and can be queried through a SPARQL endpoint (http://www.semantic-systems-biology.org/apo/).

1 INTRODUCTION

Research in the Life Sciences is supported by a plethora of databases (see overview at www.pathguide.org). Moreover, the continuing advancements in functional genomics technologies make it possible to create an overwhelming amount of data in a single experiment. The many hypotheses that can be derived from such experiments must be assessed against a multitude of information and knowledge bases, often represented in a variety of formats. Scientists therefore become increasingly dependent on sophisticated computer technologies to integrate and manage all the available information. Furthermore, the drastic increase in the available information and a lack of adhering to accepted formal representations across all disparate knowledge bases allows only a fraction of the knowledge to be easily considered in the analysis of new data, or causes a user to query many databases individually, sometimes even without the support of ontology terms that would warrant a common semantics of queries in different databases. As discussed by Antezana et al. (2009), application ontologies can facilitate the query process itself as the ontology ensures a uniform semantics across all data.

1.1 Need for an integrated resource that captures gene expression knowledge

Transcriptional gene expression and its regulation depend on a large variety of cellular processes that control the timing and level of transcription of an individual gene, often in a cell- or condition specific manner. Regulation of the expression of protein coding genes is extensively studied. Gene expression falls into two main phases, *i.e.* transcription and translation. During the process of transcription, proteins called transcription factors bind to specific DNA sequence motifs (binding sites) of a gene, playing a key role in initiating or inhibiting the formation of an active RNA Polymerase II transcription complex. Active transcription produces pre-mRNAs which are subsequently processed (removal of introns, and polyadenylation of the transcript) upon which mature mRNAs are transported from the nucleus to the cytoplasm where the mRNA is translated into a protein. Regulatory processes of gene expression occur at different levels, enabling the cell to adapt to different conditions by controlling its structure and function. Furthermore, the process of gene expression may also be influenced at the epigenetic level, where nucleotide or protein modifications can cause heritable changes in expression of otherwise identical gene sequences. Abnormalities in the regulation of gene expression can cause diseases such as the occurrence of malignant cell proliferation.

The knowledge required to decipher the various processes involved in gene expression continues to grow. However,

[*] To whom correspondence should be addressed: mironov@nt.ntnu.no
[†] These authors contributed equally

for a systems-wide understanding of gene regulation, there is a need for efficiently capturing knowledge of this domain in its entirety and to further facilitate efficient querying of this data. For instance, the complex one-to-many relationships of a transcription factor like Myc includes thousands of target genes, representing a wide variety of functions and processes. An ontology-driven approach would best solve the issue of knowledge querying, representation and management. Previously, attempts have been made to model the gene regulation process; resulting in the Gene Regulation Ontology (GRO) (Beisswanger et al., 2008). GRO provides a conceptual model to represent common knowledge about the gene regulation domain. However, it was primarily built as a scaffold for knowledge intensive natural language processing (NLP) tasks and lacks the granularity in concepts much needed for advanced querying and hypothesis generation.

We have built a system that integrates existing ontologies relevant for the domain of gene expression to support the discovery of new scientific knowledge. We have named this knowledge system: the **Ge**ne E**x**pression **K**nowledge **B**ase (GeXKB). This system is conceived as part of the Semantic Systems Biology (SSB) (http://www.semantic-systems-biology.org) initiative and comprises at the current stage three application ontologies that capture the knowledge about gene expression, namely the **Ge**ne E**x**pression **O**ntology (GeXO), **Re**gulation of Gene E**x**pression **O**ntology (ReXO) and the **Re**gulation of **T**ranscription **O**ntology (ReTO).

2 GEXKB OBJECTIVES AND DESIGN PRINCIPLES

GeXKB is designed to provide the molecular biologist with a knowledge system that captures knowledge on a variety of aspects of the gene expression process. To this end it should be able to provide answers to questions like:

- *'Which are the proteins that act as chromatin remodeling proteins and as modulators of transcription factor activity?'*
- *'Which are the proteins that participate in two successive regulatory pathways?'.*
- *'Which are the transcription factors (Human) that are located in the cytoplasm?'.*

The following design principles were followed in the process of GeXKB development:

- 'is a' completeness
- 'all-some' semantics
- only classes used for modelling of the domain of discourse (see Table 1)
- maximal flexibility both for users and for future extensions

3 GEXKB ARCHITECTURE AND CONSTRUCTION

The core of the three ontologies is built of terms from a number of well established biomedical ontologies, first of all GO (Ashburner et al., 2000) and Molecular Interactions ontology (Kerrien et al., 2007), The core is used to integrate data from GOA (Barrell et al., 2009), IntAct database (Kerrien et al., 2007), KEGG (Kanehisa and Goto, 2000), UniProtKB (Magrane and Uniprot consortium, 2011) and NCBI Gene (Wheeler et al., 2005). In the subsequent sections we describe the architecture and the main features of the ontologies.

3.1 Data integration pipeline

The ontologies were built using an automated pipeline implemented with the use of the library ONTO-PERL (Antezana et al., 2008).

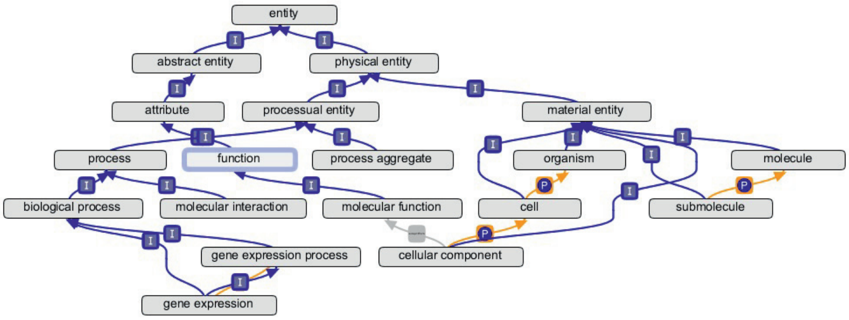


Figure 1: The figure illustrates the seed ontology of GeXO.

3.1.1 Building seed ontologies:

GeXO, ReXO and ReTO share a common Upper Level Ontology (ULO), which provides a general scaffold for data integration. It was developed on the basis of the Science Integrated Ontology (SIO) (http://code.google.com/p/semanticscience/wiki/SIO) with the addition of few terms from other ontologies. The origin of the terms is preserved in external references. The ULO is generated on the fly by the pipeline and does not exist as an individual artifact. The upper level term IDs are of the form 'SSB:nnnnnnn'.

The ULO is then merged with GO (domain specific fragments of Biological Process, complete Cellular Component, complete Molecular Function), MI ('interaction type' branch), and the Biorel ontology (Blondé et al. 2011). This yields three ontologies referred to as seed ontologies. To be more specific, in order to build the seed ontology for GeXO, the term 'gene expression' (GO:0010467) and all its descendants are imported. For ReXO and ReTO the corresponding GO terms are: 'regulation of gene expression' (GO:0010468) and 'regulation of transcription, DNA dependent' (GO:0006355). We refer to these three terms as sub-roots. Each of them is connected to the ULO as a subclass of 'biological process'. To ensure 'is a' completeness, each of the ontologies is complemented with an auxiliary term - ('gene expression process' (GeXO:0000001), 'process of regulation of gene expression' (ReXO:0000001), 'process of regulation of DNA-dependent transcription' (ReTO:0000001)), which becomes the parent of all the terms that did not have an 'is a' path to the sub-root. Apart from this, the three seed ontologies are structurally identical (Figure 1).

3.1.2 Building species specific intermediate ontologies:

The GeXKB ontologies support three model organisms: *Homo sapiens, Mus musculus* and *Rattus norvegicus*.

The corresponding three species-specific intermediate ontologies were developed in the following steps:

(1) For each species GOA annotations are used to extract all the associations involving domain specific Biological Process terms incorporated in the previous phase. The corresponding proteins are added as child terms to the upper level term 'protein' (SSB:0001211) and referred to as 'core proteins' hereafter.

(2) From the IntAct database all the interactions involving at least one of the core proteins are retrieved and incorporated into the knowledge base along with their pertinent information. This results in a further extension of the set of proteins in the KB.

3.1.3 Building the complete ontologies:

This is the final phase in the generation of the ontologies which proceed as follows:

(1) The species specific ontologies (from the previous step) are merged together.

(2) From the KEGG database all the pathways involving at least one of the core proteins are extracted and incorporated in the KB along with the pertinent information. The pathway terms become children of the term 'SSB:0011221' ('pathway', 'BioPAX:Pathway'). The corresponding KEGG orthology groups are incorporated as children of the term 'protein cluster' (SSB:0001122). This step results in a second extension of the set of proteins.

(3) Putative orthology relationships were computed with the use of the high-performance library TurboOrtho (Ekseth et al., 2010), a multi-threaded C++ implementation of the OrthoMCL algorithm (Li et al., 2003). The relations including core proteins are added to the KB, leading to the final extension of the set of proteins.

(4) The set of proteins in the GeXKB was finally augmented with:

- GOA annotations for Cellular Components and Molecular Functions,
- Additional information (e.g. protein modifications) from UniProtKB,
- The corresponding genes along with the pertinent information from NCBI.

The final result is the three ontologies in the OBO (Smith et al., 2007) format.

3.1.4 Enhancing the utility of the ontologies:

(1) Transitive closures were constructed with the use of the library ONTO-PERL for the following relation types: 'is a', 'part of', 'regulates'.

(2) The ontologies were exported in a number of formats: RDF, OWL, XML, and DOT.

(3) The RDF exports were used to populate a triple store, refer Table 2 (Virtuoso Open Link).

Ontology	No. of classes	No. of relations	No. of instances
GeXO	168417	15	0
ReXO	152962	15	0
ReTO	141095	15	0

Table 1: An overview of the ontologies in GeXKB

3.2 GeXKB and the Semantic Web

The Semantic Web (Berners-Lee and Hendler, 2001) is an extension of the WWW which aims at building a web of data accessible both by computers and human beings. This new technology is increasingly gaining momentum, in particular in the domain of Life Sciences (Antezana et al., 2009).

In order to make use of these new technologies, the RDF versions of the ontologies have been loaded into Open Link Virtuoso (http://virtuoso.openlinksw.com) and can be accessed via a SPARQL query page (http://www.semantic-systems-biology.org/apo/queryingcco/sparql). In contrast to other Semantic Web formalisms, such as OWL, RDF enables handling of large amounts of knowledge due to its simple and flexible syntax, making querying tractable. However, on the downside the low expressivity of RDF/RDFS imposes limitations on the inferencing over the knowledge base. To overcome this limitation, Blondé et al. (2011) have developed a novel approach for semi-automated reasoning on RDF stores with the use of the SPARUL update language (http://www.w3.org/TR/sparql11-update/). This allows for pre-computing the inferences supported by the store, thus making implicit knowledge explicit and available for querying. In order to provide maximum flexibility for querying, two graphs are available for each of the ontologies - with or without closures (e.g. GeXO-tc and GeXO, 'tc' standing for 'total closure').

The most convincing evidence of the success of the Semantic Web is the quick expansion of the Linked Data cloud (Heath and Bizer, 2011). In the course of the design of GeXKB a number of decisions were made to facilitate the migration of GeXKB eventually to the Linked Data cloud. For instance, we have re-used original IDs as much as possible. If the original IDs include a name-space (e.g. GO, MI) they were adopted without any modifications, otherwise the IDs were prepended with a name-space (for example UPKB for UniProtKB or NCBIgn for NCBI Gene), separated by a colon from the original ID (the colons are replaced with underscores in the RDF renderings). The re-use of the IDs benefits as well the users due to faster query execution and the familiarity of the IDs. Furthermore, in compliance with the Linked Data recommendations we minted the URIs in our own common name-space: http://www.semantic-systems-biology.org/ and have consistently used *rdfs:label* properties to aid human readability of the results.

RDF graphs	GeXO	GeXO-tc	ReXO	ReXO-tc	ReTO	ReTO-tc
No. of triples	~3.3 million	~23 million	~3 million	~19.9 million	~2.8 million	~19.1 million

Table 2: Shows the number of triples in the individual graphs of GeXKB

4 QUERYING GEXKB

In this section we demonstrate the utility of GeXKB with the help of a few example SPARQL queries. These queries are available as a part of a list of sample queries provided on the query page (http://www.semantic-systems-biology.org/apo/queryingcco/sparql). To query GeXKB, the base URI and the prefixes are set and the SELECT block specifies the variables to be part of the solution. The RDF triple pattern queried is defined in the WHERE block. The queries are as follows:

Q1: (see Table 3)

Biological question: *Which proteins can act as chromatin remodeling proteins and as modulators of transcription factor activity?*

SPARQL query:

```
BASE <http://www.semantic-systems-biology.org/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb: <SSB#>
PREFIX taxon: <SSB#NCBItx_9606>
PREFIX graph1: <ReXO>
PREFIX graph2: <ReTO-tc>

SELECT distinct ?protein_id ?protein_name
WHERE {
 GRAPH graph1: {
  ? protein_id ssb:is_a ssb:SSB_0001211 .
  ?b_process ssb:is_a ssb:GO_0040029 .
  ?b_process ssb:has_participant ? protein_id .
  ? protein_id ssb:has_source taxon: .
 }
 GRAPH graph2: {
  ssb:GO_0034401 ssb:has_participant ? protein_id .
  ? protein_id rdfs:label ?protein_name .
 }
}
LIMIT 4
```

Q2:
Biological question: *Which proteins participate in both the JAK/STAT signaling pathway and Apoptosis?*
SPARQL query:

```
BASE <http://www.semantic-systems-biology.org/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb: <SSB#>
PREFIX taxon: <SSB#NCBItx_9606>
PREFIX pathway1: <SSB#KEGG_ko04630>
PREFIX pathway2: <SSB#KEGG_ko04210>
PREFIX graph: <GeXO>

SELECT distinct ?protein
WHERE {
 GRAPH graph: {
  ?prot_id ssb:is_a ssb:SSB_0001211 .
  ?prot_id ssb:is_member_of ?cluster .
  pathway1: ssb:has_agent ?cluster .
  ?prot_id ssb:has_source taxon: .
 }
 GRAPH graph: {
  ?prot_id ssb:is_member_of ?cluster .
  pathway2: ssb:has_agent ?cluster .
  ?prot_id rdfs:label ?protein .
 }
}
```

Q3:
Biological question: *Which are the transcription factors (Human) that are located in the cytoplasm?*
SPARQL query:

```
BASE <http://www.semantic-systems-biology.org/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb: <SSB#>
PREFIX taxon: <SSB#NCBItx_9606>
PREFIX location: <SSB#GO_0005737>
PREFIX graph: <ReTO-tc>

SELECT distinct ?protein ?protein_name
WHERE {
 GRAPH graph: {
  ?protein ssb:is_a ssb:SSB_0001211 .
  ?protein rdfs:label  ?protein_name .
  ssb:GO_0006355 ssb:has_participant ?protein .
  ?protein ssb:has_function ?function .
  ?function ssb:is_a  ssb:GO_0003700 .
  location: ssb:contains ?protein .
  ?protein ssb:has_source taxon: .
 }
}
```

These queries offer just a glimpse of the repertoire of biological question that can be addressed to the knowledge system. In addition, users could also query the knowledge base in combination with other complementary semantic web resources to formulate advanced queries for hypothesis generation. This could be performed through the query federation features that are included in the latest version of SPARQL (ver. 1.1) and will be explored in the future.

Protein ID	**Protein Name**
http://www.semantic-systems-biology.org/SSB#UPKB_Q9NS37	ZHANG_HUMAN
http://www.semantic-systems-biology.org/SSB#UPKB_P14373	TRI27_HUMAN
http://www.semantic-systems-biology.org/SSB#UPKB_Q62158	TRI27_MOUSE
http://www.semantic-systems-biology.org/SSB#UPKB_P17947	SPI1_HUMAN

Table 3: The table shows the results for Q1

5 CONCLUSION

The drastic increase in the amount of data generated in the field of molecular biology and biomedicine requires efficient knowledge management practices. Ontologies certainly provide a robust method to integrate data and efficiently represent specific (sub) domain knowledge. With the creation of GeXKB, we have built a knowledge system that specifically supports researchers focusing on various aspects of gene expression. The three ontologies provide the user with the flexibility of choosing an ontology depending on the breadth and specificity of information needed. Further flexibility is afforded by a range of available formats for knowledge representation (OBO, RDF, OWL), data exchange (XML), and visualisation (DOT).

The presented examples demonstrate the utility of our knowledge base with respect to answering realistic domain specific questions, and this utility is expected to grow with its further development. The primary goal will be to augment the knowledge base with additional high quality, curated sources of information with documented transcription factor function and relations between transcription factors and their target genes.

ACKNOWLEDGEMENTS

This work is funded by the Norwegian University of Science and Technology (NTNU), Norway. AV was funded by Faculty of Natural Science and Technology and VM was funded by FUGE Mid-Norway.

REFERENCES

Antezana, E., Egaña, M., De Baets, B., Kuiper, M., and Mironov, V. (2008). ONTO-PERL: an API for supporting the development and analysis of bio-ontologies. Bioinformatics. Mar 15;24(6):885-7.

Antezana, E., Kuiper, M., and Mironov, V. (2009). Biological knowledge management: the emerging role of the Semantic Web technologies. Brief Bioinform., 10(4): 392-407.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski , K. et al., (2000). Gene ontology: tool for the unification of biology. *Nature Genetics,* 25**,** 25-9.

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., and Apweiler, R. (2009). The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Research* 37: D396-D403.

Beisswanger, E., Lee, V., Kim, J. J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., Hahn, U. (2008). Gene Regulation Ontology (GRO): design principles and use cases. Stud Health Technol Inform. 2008;136:9-14.

Berners-Lee, T. and Hendler, J. (2001). 'Publishing on the semantic web'. *Nature,* 410**,** 1023-4.

Blondé, W., Mironov, V., Venkatesan, A., Antezana, E., De Baets, B., and Kuiper M. (2011). Reasoning with bio-ontologies: using relational closure rules to enable practical querying. *Bioinformatics*, Jun 1;27(11):1562-8.

Ekseth, O., Lindi, B., Kuiper, M., and Mironov, V. TurboOrtho – a high performance alternative to OrthhoMCL. *European Conference on Computaional Biology*: *September 2010*; *Ghent.*

Heath, T., and Bizer, . (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res,* 28**:**27-30.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod N, Bader, G. D., Xenarios, I. et al. (2007). Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biol. 9;5:44.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M. et al. (2007). IntAct - open source resource for molecular interaction data. *Nucleic Acids Res*, 35:D561-565.

Li, L., Stoeckert, C. J. Jr., Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13:2178-2189.

Magrane M. and the UniProt consortium UniProt Knowledgebase: a hub of integrated protein data Database, 2011

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K. et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol, 25(11), 1251–1255.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D.M., DiCuccio, M. et al. (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2005, 33:D39-45.

Chapter 5

Paper IV

Network candidate discovery using the Gene eXpression Knowledge Base

Aravind Venkatesan[1†], Sushil Tripathi[2†], Alejandro Sanz de Galdeano[3], Ward Blonde[1], Astrid Lægreid[2], Vladimir Mironov[1], Martin Kuiper[1*]

[1]Department of Biology, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway.

[2]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), N-7489 Trondheim, Norway.

[3]Escuela Nacional de Sanidad, Instituto de Salud Carlos III, 28029 - Madrid, Spain.

[†]these authors contributed equally

*Corresponding author

Email addresses:

AV: aravind.venkatesan@ntnu.no

ST: sushil.tripathi@ntnu.no

ASG: asdgvarela@gmail.com

WB: w.blonde@vumc.nl

VM: vladimir.mironov@ntnu.no

AL: astrid.lagreid@ntnu.no

MK: martin.kuiper@ntnu.no

Abstract

Background

Network-based approaches for the analysis of large-scale genomics data have become well established, as biological networks provide a knowledge scaffold against which the patterns and dynamics of *'omics'* data can be interpreted. The background information required for the construction of such networks is often dispersed across a multitude of knowledge bases in a variety of formats. The integration of this information into one seamless knowledge resource is one of the main challenges in bioinformatics. The Semantic Web offers powerful technologies for the assembly of integrated knowledge bases that are computationally comprehensible, providing a potentially powerful resource for constructing biological networks and network-based analysis.

Results

We previously developed the Gene eXpression Knowledge Base (GeXKB), a semantic web technology based resource that contains integrated knowledge about gene expression regulation. We have enhanced and extended GeXKB and now demonstrate how this resource can be exploited for the identification of candidate regulator proteins that should be considered for integration into a regulatory network model. We present four use cases that were designed from a biological perspective in order to find candidate members relevant for our model network: the gastrin hormone signaling network. These candidates (regulator proteins and regulated genes) were subjected to a number of criteria to further substantiate their potential role in gastrin-mediated regulation of gene expression in our model system: AR42J cells. We have identified 33 potential regulator proteins and two regulated genes which may be considered for the extension of the gastrin response network.

Conclusions

The GeXKB offers biologists an integrated resource that allows complex biological questions pertaining to gene expression. Semantic Web technologies provide the means for integrating various heterogeneous knowledge sources in order to extract maximum information (if needed inferred computationally). This work illustrates how new potential candidates can be retrieved for the extension of a gene regulatory network.

Background

Cellular signaling cascades support the transmission of information from external signals (e.g. hormones) to distinct cellular responses, for instance changes in gene expression. Gene expression is controlled by a network of highly interconnected proteins known as transcription regulators [1, 2]. There is a large array of transcription regulators including sequence-specific DNA binding transcription factors (DbTFs), various transcription co-factors and chromatin modifiers [3, 4]. The information concerning these regulatory network components is scattered across a multitude of resources in a variety of formats, making it a challenge to consider all information when addressing questions to this fragmented knowledge resource.

More generally, the formulation and assessment of biological hypotheses against prior knowledge fundamentally relies on efficient knowledge integration that interlinks information and knowledge at various levels in standardized formats, after which the best supported hypotheses can be selected for testing in wet-lab experiments. The development of the technology for knowledge integration, metadata requirements and knowledge representation formats therefore has evolved to become a major research area [5, 6].

Ontologies have become a fundamental scaffold for the representation of biological knowledge. The Open Biomedical Ontologies (OBO) Foundry [7] provides a set of guidelines to structure the co-ordinated development of bio-ontologies. Bio-ontologies developed following the guidelines of the OBO Foundry are becoming widely used the life science community. The Gene Ontology (GO), a prominent example of this [8], provides a unified representation of properties of genes and their products. Furthermore, the Gene Ontology Annotation project [9] facilitates the unambiguous annotation of gene products with GO terms covering molecular function, cellular component and biological process aspects.

In parallel, the Semantic Web initiative [10] essentially aims at transforming the current Web into a global reasoning and semantics-driven knowledge base. The semantic web is founded on a collection of technologies such as the Resource Description Framework (RDF) [11], RDF Schema (RDFS) [12], Web Ontology Language (OWL) [13] and SPARQL Query Language (SPARQL) [14] which provides the foundation for data integration, and the means for querying and reasoning. RDF has become the foundation for data integration across computing platforms due the flexibility it offers in describing data. RDF models data in the form of so-called triple statements, comprising a subject, a predicate and an object. Triples can be joined in a large network of data that can be integrated from different sources, essentially in the form of a graph. At the core, RDF and the associated semantic web technologies like RDFS and OWL use the Uniform Resource Identifiers (URI) to identify real-world objects and concepts enabling interactions over the Web using the Hypertext Transfer Protocol (HTTP). The SPARQL querying language allows the retrieval of triples of interest (a sub-graph) from an arbitrary set of RDF graphs. It is considered as one of the key technologies of the semantic web as it enables users to query and integrate results from multiple RDF graphs that may reside at any location.

The semantic web promises to meet the challenges in knowledge representation and management by providing flexible frameworks for the modelling of knowledge of any given domain. We are currently witnessing a growing use of semantic web technologies for the management of broad repertoires of biological concepts and for providing a scaffold for the integration of concepts and data from disparate biological databases. As part of these efforts we previously built the Cell Cycle Ontology (CCO) [15] and the BioGateway knowledge base [16], both part of the Semantic Systems Biology platform [17]. Several other initiatives are demonstrating the potential of semantic web technologies. Semantic web resources such as Bio2RDF [18] and Linked Life Data [19] provide integrated knowledge that comprises information from NCBI Gene, UniProt, DrugBank, Pfam and the Protein Data Bank (PDB), to name just a few. These resources are generic in their scope by covering many aspects of the life science domain, but they lack information specific to gene regulation such as the relationships of DbTFs and the genes that they regulate ('target genes'). Furthermore, AmiGO [20] is the official web-based tool suite that allows users to analyse GO-annotated genes and proteins, such as browsing GO branches, searching for gene or gene product associations with specific GO terms, performing sequence similarity (BLAST) searches and viewing the associated GO terms for the returned list of genes or proteins. This tool however is restricted to the information housed in the GO database. The Orymold system [21] provides gene expression information integrated from TIGR rice genome database and Plant Proteome Annotation program for the model organism *Oryza sativa*. The system follows two tier architecture: the gene expression information is managed by a traditional relational database management system (RDBMS), OryDB; an ontology (Orymold ontology) is built on top of OryDB which serves as an integration scaffold and aids users make ontology based queries via a user interface. Data access in Orymold is made flexible by the ontology layer however it is built on a RDBMS platform which is a limiting factor towards efficient data integration. IntegromeDB [22] is a graph-based semantic knowledge base that integrates publicly available information focusing on transcriptional regulation. Users may opt to analyze the integrated data with the BiologicalNetworks application [23], but the possibilities to address complex queries of the type that SPARQL can handle are limited. Alternatively, WikiPathways [24] is an open collaborative platform for curated information on biological pathways. The information from WikiPathways has been converted to RDF providing access this knowledge via a SPARQL endpoint [25], but this is restricted to the information on biological pathways as the resource does not integrate information on protein-protein and DbTF-target gene interactions.

Broadly speaking, the various initiatives that provide semantic web solutions have greatly facilitated the process of integrating data from various sources. However, this does not mean that the semantic web has become deeply integrated in the repertoire of tools currently in use by experimental biologists. In that sense the semantic web represents a typical technology push, the technology can only make an impact when it becomes broadly embraced by the end users. To further establish the advantages of the semantic web and to encourage the experimental biologists to utilise these resources as part of their daily research activities requires involvement of the user community in the development of resources and the lowering of hurdles to adopt the new approach. For instance, many semantic web knowledge bases currently provide a SPARQL endpoint as a means to access the integrated data. Although in most cases sample queries are provided to guide the users, these are not very user-friendly. In some cases, the knowledge bases are equipped with faceted browser interfaces that allow users to explore the information by applying multiple filters and keywords. These are suitable for quick searches, such as retrieving a local neighborhood of a particular term (e.g. an ontological term, protein or gene identifiers). Although this helps the user in getting acquainted with the information offered by these resources, specific biological questions often need the formulation of much more complex queries. This requires at least a minimum knowledge of SPARQL, and as it is evident that this query language is somewhat intimidating for the biologists it results in an underuse of available semantic web resources. Providing interfaces that covert natural language questions to SPARQL queries is an active area of research with some advances been made in this regard outside the bioinformatics domain [26, 27, 28]. However, these approaches need to be carefully studied and adapted to suit the needs of the bioinformatics domain.

Therefore, in a close collaboration between semantic web specialists and experimental biologists we developed and exploited a set of resources designed as an analysis platform for the study of gene regulation events and built the Gene eXpression Knowledge Base (GeXKB) [29]. Here we describe the results of this joint work which focused on a number of concrete biological questions addressing the transcriptional regulatory network being regulated by gastrin-mediated signaling cascades.

Gastrin is a gastrointestinal peptide hormone which, similar to many other extracellular signals such as e.g. growth factors, plays a crucial role in both normal and pathological processes. After binding to its receptor CCK2R (cholecystokinin 2 receptor) gastrin triggers the activation of multiple intracellular signaling pathways and transcription regulatory networks culminating in the regulation of a vast number genes and cellular responses. We previously performed an extensive genome-wide gene expression time-series experiment on a gastrin-hormone induced AR42J cells [30]. This work allowed us to identify global changes in mRNA levels in response to gastrin, serving as an experimental reference for our study. In addition, we constructed a CellDesigner [31] map of gastrin-responsive intracellular signaling and transcription regulation networks based on an exhaustive search for experimental evidence reported in literature [32]. This network map was taken as a point of departure to identify new proteins that should be considered as putative network extensions. Sets of candidate proteins and genes were retrieved from the GeXKB resource based on a series of specific biological questions and converting those into computable queries that could be launched against GeXKB. The subsequent sections briefly describe the architecture and the improvements made to GeXKB, the construction of queries to generate the sets of candidate proteins and how these sets were assessed for significance with respect to the network model.

Methods

Experimental data

A genome-wide gene expression time-series experiment was performed on AR42J cells, available from the ArrayExpress database [33] (accession number: GSE32869). Analysis of this data set showed ~2000 genes with changes in expression in response to gastrin. For proteins to qualify as valid network candidate we reasoned that its corresponding gene expression should be responsive to gastrin.

Gastrin and CCK2R mediated signaling network

We constructed a map of gastrin-responsive intracellular signaling and transcription regulation proteins and genes based on an exhaustive search for experimental evidence reported in literature [32]. This map provided a comprehensive overview of all studies performed to date concerning gastrin inducible pathway members, and it served as a seed to query for new network candidates.

GeXKB Architecture

Capturing knowledge of the domain of gene expression requires the integration of information covering a broad range of biological processes, and molecular functions. Furthermore, it is important to capture the complex molecular interactions that represent the associations between DbTFs and their target genes and also the factors that modulate these interactions. To this end, GeXKB was designed to integrate high quality (curated) knowledge from a variety of sources. The knowledge base comprises three application ontologies: the **Ge**ne e**X**pression **O**ntology (GeXO); the **Re**gulation of Gene e**X**pression **O**ntology (ReXO); and the **Re**gulation of **T**ranscription **O**ntology (ReTO). GeXO, ReXO and ReTO are three nested ontologies (see Figure 1) with, in the order given, an increasingly narrower biological focus: users interested specifically in the regulation of nuclear transcription may find it more convenient and efficient to use ReTO instead of e.g. GeXO. GeXKB supports the three model organisms *Homo sapiens, Mus musculus* and *Rattus norvegicus*. The GeXKB ontologies share a common Upper Level Ontology (ULO), which serves to ‘glue’ together the various components of the application ontology. The ULO terms were developed on the basis of the Semanticscience Integrated Ontology (SIO) [34] and BioPAX [35]. In the current version of the GeXKB ontologies (v1.01), the ULO contains additional terms from other ontologies, including the Ontology for Biomedical Investigations (OBI) [36]; Chemical Entities of Biological Interest (ChEBI) [37]; and the Information Artifact Ontology (IAO) [38]. The ULO is merged with the GO through sub-domain-specific fragments of the Biological Process branch, and the complete Molecular Function and Cellular Component branches. More specifically, the GO gene expression sub-domain terms ‘gene expression’ (GO:0010467), 'regulation of gene expression' (GO:0010468) and 'regulation of transcription, DNA dependent' (GO:0006355) with all their respective descendants are imported to form GeXO, ReXO and ReTO, respectively. Additionally, the molecular interaction data is supported by the ‘interaction type’ branch of the Molecular Interaction (MI) ontology [39]. The Biorel ontology [40], an extension of the Relational Ontology [41] is added to provide additional vocabulary to logically link entities with relation attributes such as transitivity, reflexivity and subsumption.

GeXKB is fundamentally a protein-centric resource. Protein information from various sources is linked to the integrated ontologies to form the subsequent application ontologies. These sources include the UniProt Knowledgebase [42] (protein annotations including protein modification information); the Gene Ontology Annotations [9]; KEGG [43] (pathway data); and IntAct [44] (protein-protein interactions). The corresponding gene information is integrated from NCBI Entrez datasets [45]. Also, orthology relationships were predicted using orthAgogue [46], a high performance C++ implementation of OrthoMCL [47]. Furthermore, the current

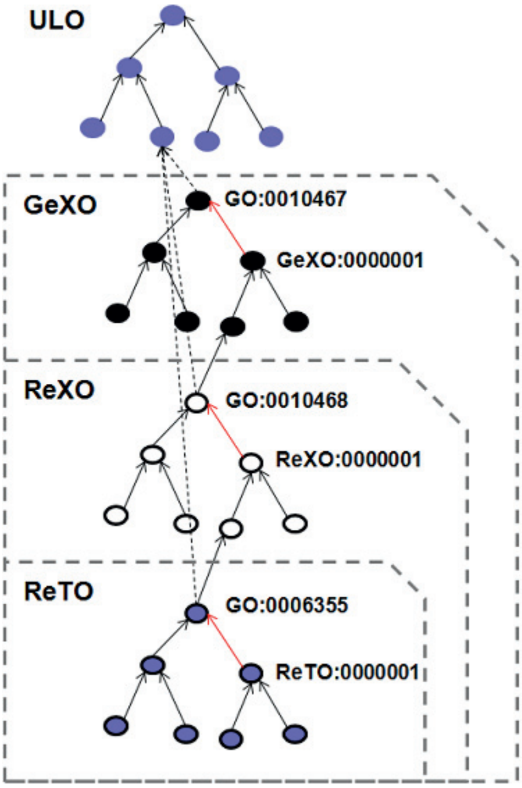


Figure 1: GeXKB ontologies.

The illustration shows the layout of the nested GeXKB ontologies (GeXO, ReXO and ReTO).The blue nodes represent the upper level ontology (ULO), the common root of the three ontologies. The black and red edges depict 'is_a' and 'part_of' relations, respectively. The three ontologies cover an increasingly wide domain from bottom to top. Each GO sub-domain term (e.g. GO:0010467; denoting 'gene expression') and its descendants are linked to the ULO as a subclass of 'Biological Process'.

version of GeXKB contains documented information about the functional interaction of DNA binding transcription factors with their target genes, added from a number of sources which includes a) PAZAR database [48], an open source framework that serves as an umbrella to bring together datasets pertaining to transcription factors and regulatory sequence annotations; b) Human Transcriptional Regulation Interactions (HTRI) database [49], an open-access database that serves as a repository for experimentally verified human transcription factor - target gene interactions; c) TFactS [50], a database that catalogs curated transcription factor - target gene interactions; and d) TFcheckpoint [51], a database that compiles curated information on human, rat and mouse DbTF candidates from many different resources.

Data integration pipeline

The GeXKB ontologies are generated by an automated data integration pipeline (Figure 2) that relies on the ability to programmatically manipulate ontologies through the ONTO-PERL API [52]. The pipeline is designed to accommodate the dynamic structure of biological information that is constantly being extended and updated. Hence the production of the knowledge base follows a particular cycle involving periodical download of data and integration of ontologies and data from scratch. As a first step, the ULO is assembled as a seed structure for linking both to the various GO sub-domain fragments and the MI and Biorel ontologies. This step generates three ontologies known as the seed ontologies. At the next step sets of proteins are retrieved from the Gene Ontology Annotation files by association with the Biological Process terms present in each of the seed ontologies. These sets of proteins (referred to as 'core' proteins) are used in the subsequent steps as a basis to select additional proteins from IntAct interactions, KEGG pathways and binary orthology relations as predicted by orthAgogue. Next, protein modifications, basic gene information and associations with Cellular Component and Molecular Function terms are added from UniProt, NCBI Entrez and the Gene Ontology Annotations. The pipeline finally outputs the three application ontologies in OBO [53], DOT [54] and XML [55] formats. A more detailed description of the pipeline has been provided earlier [29].

Accessing GeXKB

The semantic web extends on conventional Web technologies with the aim to facilitate automated machine interoperability by providing sophisticated frameworks for representation, management and retrieval of information. GeXKB utilizes the knowledge representation features offered by RDF and builds on previous efforts to use semantic web technology for the integration of knowledge [15,16,18,19,56]. The GeXKB ontologies are uploaded as RDF graphs to a Virtuoso data storage engine [57], which makes them accessible as a SPARQL query endpoint [58]. The query page contains sample queries that can be easily customized to help users explore the knowledge base.

Although RDF is efficient in integrating data, it has limited expressivity and it was not conceived to perform inferencing tasks. In GeXKB this limitation is partially overcome by the use of a semi-automated reasoning approach developed by Blondé *et al*. [40]. This approach allows the inference of new relations from the existing relation types in GeXKB based on five inference rules, namely reflexivity, transitivity, priority over the subsumption relation, superrelations and compositions. For details about the theoretical aspect of these inference rules see [40] and [59]. The inferencing process is implemented by using the SPARQL update language (SPARUL) [60] and a scaffold provided by the Biorel ontology to pre-compute the inferences, essentially to extract implicit relationships. This method enriches the RDF graph and offers increased power and flexibility in querying. Hence, each of the three nested ontologies is available as two graphs: either with or without the information obtained through inferencing. The graphs containing pre-computed inferences are suffixed with ‘tc’ (e.g. ReTO-tc, where 'tc' stands for total closures). Data sets from PAZAR, HTRI, TFactS and TFcheckpoint are loaded as separate RDF graphs.

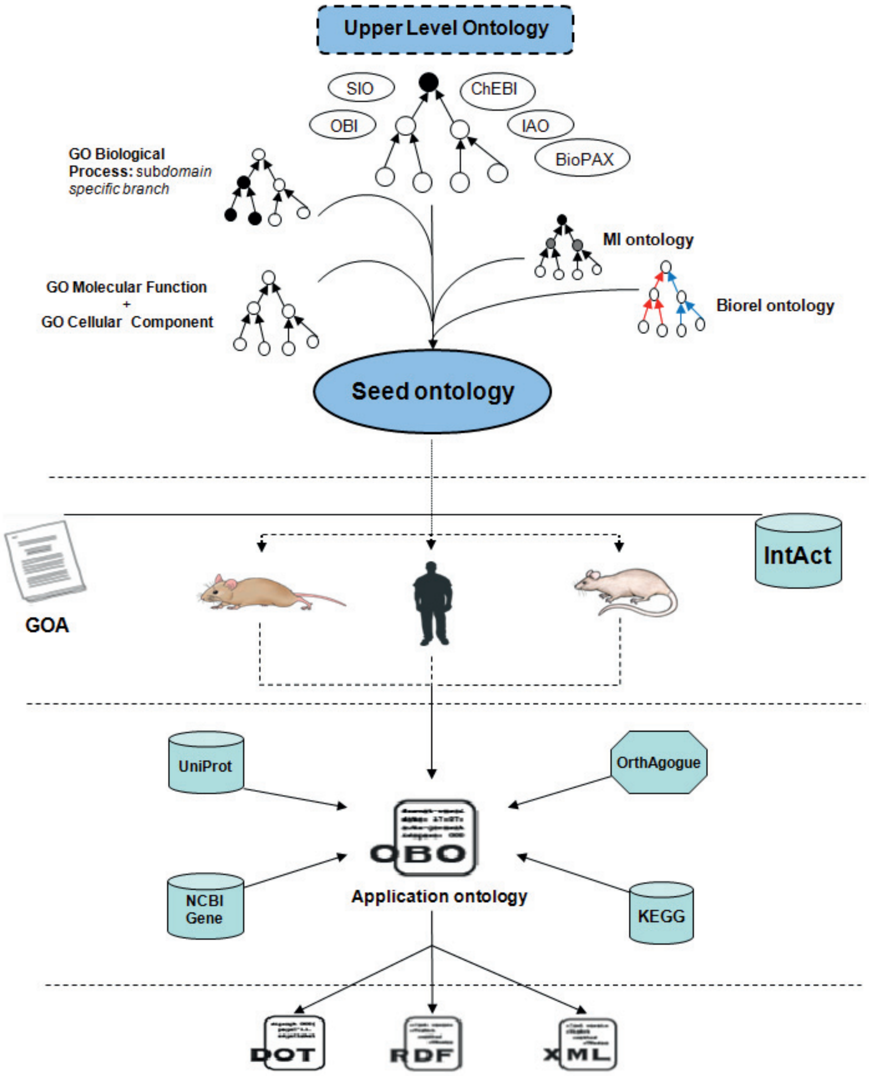


Figure 2: The data integration pipeline.

The first step of integration starts by generating an Upper Level Ontology, which is then linked with the different ontologies: GO (Biological Process, Molecular Function and Cellular Component fragments), the MI ontology and the Biorel ontology, forming a seed ontology. Mouse, human and rat-specific data are integrated from Gene Ontology Annotation files and IntAct. Next, these species-specific ontologies are merged and additional data is integrated including protein information

(UniProt), pathway annotations (KEGG), basic information for genes (NCBI) and orthology relations for proteins (orthAgogue). The final ontology is available in different formats.

A major effort of the semantic web community is focused towards making resources available as part of the Linked Data cloud [61]. We have taken initial steps in making the GeXKB resource Linked Data-compatible, therefore we re-use original IDs for all entities in GeXKB and we use a common namespace (http://semantic-systems-biology.org/SSB#) for all the URIs. This solution combines the benefits of faster query execution and familiarity of the IDs for the users. For instance, GeXKB can be queried using NCBI Gene IDs or UniProt accessions to retrieve information pertaining to the gene of interest. Additionally, the ID mapping file offered by UniProt has been uploaded in the RDF format to provide flexibility in using IDs produced by databases such as Ensembl [62].

Alternatively, the GeXKB ontologies can also be accessed through the NCBO BioPortal [63], where they can be visualized using the NCBO's FlexViz tool and queried via a SPARQL endpoint. However, it should be noted that BioPortal re-designs the URIs according to their internal policy.

Use cases

Our use cases were designed to identify new putative network components related to gene expression regulation that may play a role in the response to gastrin. We formulated biological questions that, once converted to SPARQL queries, should allow us to retrieve proteins and genes related to the current literature-based gastrin signaling network map. We reasoned that given the knowledge sources integrated into GeXKB, these queries should yield both well established and new gastrin response network participants. In total we formulated 6 questions, identified as Q1 through Q6 below, all converted to customized SPARQL queries on the *Homo sapiens* data in GeXKB.

The results returned for uses cases I through III were investigated for their relevance to the gastrin response network by categorizing them into two disjoint groups: a) proteins that are already documented as memebers in the gastrin response network, and b) potential novel components of the gastrin response network. The proteins from group *b* were further assessed for two criteria: 1) whether they show evidence of gastrin-induced regulation at the mRNA level, based on transcriptome observations previously obtained in the AR42J cell line model system (essentially a 14h time series gastrin response data set, see Methods); and 2) whether there is any literature reference implicating them to respond to stimuli in general (other than gastrin). Proteins qualifying for both criteria were deemed to constitute the most promising set of new putative network members obtained from GeXKB. Further, for use case IV the results returned for Q6 were assessed based on whether the genes (regulated by the corresponding DbTFs) are expressed in the AR42J cell line and their expression in response to gastrin stimulation. Below we provide a detailed description of the use cases and queries submitted to GeXKB.

Use Case I: Protein candidates involved in regulation of transcription factor CREB1

The cAMP response element binding protein 1 (CREB1) is a specific DNA binding transcription factor. It is known that many different signal transduction and gene regulation proteins can modulate CREB1 activity. These regulators include activators, repressors, chromatin modifiers and signaling proteins, each of which may belong to one or more of the network protein classes such as DbTFs, co-factors and kinases. We were interested in gathering all known regulators of CREB1 that could be relevant in the context of gastrin signaling.

To investigate the regulators of CREB1 (UniProt accession: P16220; commonly referred to as "CREB"), three queries were formulated (Q1-Q3, supplementary material). Query Q1 retrieves all proteins that are involved in the activation of CREB1. To achieve this, the query combined different terms that suggest the activation of CREB1. First of all, we used the ReTO and ReTO-tc graphs as default graphs for the queries as they are suitable to query nuclear transcriptional processes. Further, the GO terms *positive regulation of CREB transcription factor activity* (GO:0032793) and *cAMP response element binding protein binding* (GO:0008140) were included in the query. These terms suggest direct association with the process of regulating CREB1. Additionally, the term *direct interaction* (MI:0407) was included in the query to retrieve proteins that interact directly with CREB1. Then, to widen the breadth of the query, a broader GO term, *positive regulation of sequence-specific DNA binding transcription factor activity* (GO:0051091) was included. However, in this case only proteins that have a *physical association* (MI:0914) with the CREB1 protein were considered, thus reducing the number of false positives (see Figure 3). Similarly, Q2 retrieves proteins associated with biological process terms *negative regulation of CREB transcription factor activity* (GO:0032792) and *negative regulation of sequence-specific DNA binding transcription factor activity* (GO:0043433).

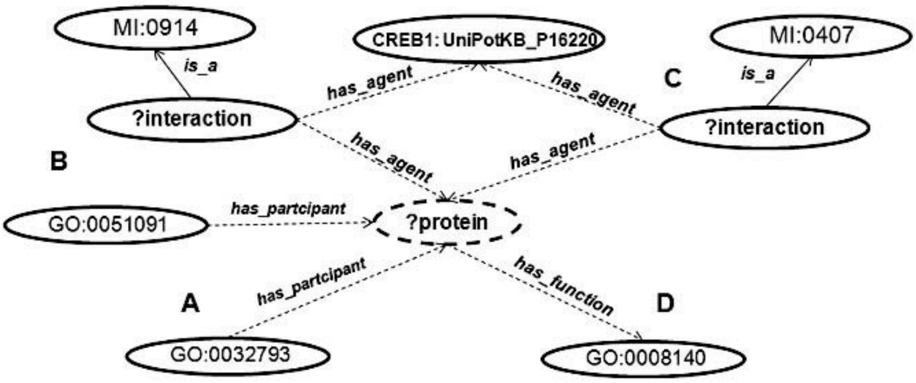


Figure 3: Conceptual model of Q1.

The cartoon displays the different concepts, ontology terms and relationships that together form a graph that was used as a SPARQL query to find matching patterns in GeXKB. The query specifies proteins that A) exhibit positive regulation of CREB transcription factor activity (GO:0032793); B) exhibit positive regulation of sequence-specific DNA binding transcription factor activity (GO:0051091) and are linked to the CREB1 protein through an association (MI:0914); C) are linked to the CREB1 protein through a direct interaction (MI:0407); and D) have function cAMP response element binding protein binding (GO:0008140).

The query Q3 specifies chromatin modifiers that are involved in the regulation of CREB1. It retrieves the union of proteins associated with molecular function terms *histone acetyltransferase* (GO:0004402) and *histone deacetylase* (GO:0004407) activity that are involved in the biological process *regulation of sequence-specific DNA binding transcription factor activity* (GO:0051090), and are interacting with the CREB1 protein. Other than providing putative network components these queries also serve to demonstrate the utility of targeting relations obtained through the inferencing process. By using the ReTO-tc graph, we were able to include implicit knowledge statements in the query output, meaning ontology term relationships not directly annotated to proteins, but shown to belong and directly linked to them through the inferencing process (see Methods 2.2: Accessing GeXKB).

Use Case II: Repressors of NFκB1 and RELA that undergo proteasomal degradation

NFκB1 and RELA are members of the NFκB transcription factor family. Members of this family are involved in regulating apoptosis, proliferation, and immune responses [64]. Gastrin-dependent regulation of this transcription factor is reported to be mediated through PKC and Rho GTPase signaling cascades [65, 66] (Figure 4). The activity of NFκB is under the control of a family of inhibitors, known as '*inhibitors of kB (IkB)*' that sequester NFκB in the cytoplasm and thereby keep these transcription factors in their inactive state. Proteasomal degradation of IkB factors results in restoration of the active state of the NFκB transcription factor. In order to gain detailed mechanistic insight in NFκB regulation in the context of the gastrin response, we were interested in retrieving the proteins that contribute to NFκB down-regulation, and also are functionality related to proteasomal degradation. This was achieved by formulating Q4, which was constructed similar to the previous queries by using a combination of terms. First, the GO term *negative regulation of NFκB transcription factor activity* (GO:0032088) was chosen as the central term, as this would retrieve all proteins annotated as negative regulators of NKκB and RELA. Next, GeXKB was explored to identify terms that suggested an involvement with proteasomal degradation. Several terms were identified: *ubiquitin ligase complex* (cellular component: GO:0000151), *ubiquitin binding* (molecular function: GO:0043130), *ubiquitination reaction* (interaction type: MI:0220), and *ubiquitin mediated proteolysis* (KEGG pathway: ko04120). The SPARQL *union* construct was used to formulate a combination of the central term and the additional term set.

Use Case III: List of components that function as repressors for TFC7L2 and activators for NFκB1 or CREB1

In the gastrin response signaling cascade, DbTFs are implicated in different cellular processes. TCF7L2 plays a central role in gastrin mediated cellular migration [67], whereas NFκB1 and CREB1 are central in regulation of gastrin-dependent immune responses and proliferation, respectively [68, 69]. Investigation of proteins that function as repressors for one transcription factor and activators for another can be of potential significance in cellular decision making.

No terms were found specifically suggesting negative regulation of TCF7L2. Therefore, Q5 was formulated by using generic terms that indicated a dual role of proteins. As a result, Q5 retrieves all proteins that interact with the TCF7L2 protein (UniProt accession: Q9NQB0) and are furthermore annotated with the terms *negative regulation of sequence-specific DNA binding transcription factor activity* (GO:0043433), and *positive regulation of sequence-specific DNA binding transcription factor activity* (GO:0051091).

Use Case IV: Identification of genes that are shared targets of DbTF regulators and the DbTFs in use case I-III

Most signal-induced cellular responses involve regulation of gene expression. DbTFs are central in regulating gene transcription rates which in turn play a key role in determining gene expression levels. Often, several DbTFs act together in the regulation of transcription of a specific gene. To enhance our understanding of mechanisms involved in gastrin mediated cellular responses we were interested in retrieving the shared target genes of the selected DbTFs (CREB1, NFKB1 and TCF7L2) and their regulators that function as DbTFs. TFcheckpoint data contains literature curated classifications of true DbTFs. Thus, queries Q1, Q2, Q4 and Q5 were extended to identify the DbTFs among the regulators using the TFcheckpoint graph. Furthermore, Q6 was formulated to retrieve target genes shared between the regulators and their corresponding DbTF from TFactS, PAZAR and HTRI graphs.

Results

The six SPARQL queries and the results of use cases I through III are available in the supplementary material. All queries combined returned 148 proteins and 20 target genes. Because we used both standard RDF graphs and graphs containing triples obtained from pre-computing relations through total closures (the tc graphs, see Methods) we were able to differentiate the results. Queries Q1, Q3, Q4 and Q5 were launched against RDF graphs containing inferred triples, meaning the contained results from tc graphs. Q1 returned 37 proteins, 24 of them obtained by inferencing; Q4 returned 32 proteins with 17 proteins resulting from inferencing. Moreover, the results produced by Q3 and Q5 were solely based on the tc graphs, and yielded 21 and six proteins, respectively. Table 1 shows the break-down of the number of proteins returned.

The proteins were classified based on the criteria described above (see section Use cases), grouping them in known members of the CCKR network (a), new candidates with evidence of regulation at the mRNA level (b1) and proteins which are described in literature as responding to stimuli other than gastrin (b2). Considering the new putative network proteins among the 105 proteins identified in Use case I, 60 proteins

Table 1: SPARQL query results

The table shows the break-down of results returned from the five SPARQL queries that were part of Use case I - III. **Asserted components:** the number of proteins retrieved by direct statements; **Inferred components:** proteins retrieved by inferred statements; **Union**: the number of proteins retrieved by using a combination of asserted and inferred statements in the queries; **Intersection:** the number of proteins that are common between asserted and inferred statements; **Total:** the total number of proteins retrieved by the five queries. Note: n/a – not applicable.

	Use Case I			**Use Case II**	**Use Case III**
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Asserted components	13	52	-	15	-
Inferred components	24	-	21	17	6
Union	37	n/a	n/a	32	n/a
Intersection	3	n/a	n/a	0	n/a
Total	37	52	21	32	6

qualify as b_1 and 16 proteins as b_2 (Table 2a, Supplementary material). Similarly, Use case II yielded 32 proteins, 20 of which belonging to b_1 and 12 to b_2 (Table 2b). Use case III resulted in six proteins; all of them are member of both group b_1 and b_2 (Table 2c). Furthermore, Use case IV yielded 18 potential regulators of CREB1, three of NFKB1 and two of TCF7L2; all of them are DbTFs based on the TFcheckpoint data (Table 2d). These regulators proteins were subsequently used in Q6 from Use case IV to identify target genes that they share with CREB1, NFKB1 or TCF7L2. This query yielded 20 target genes (19 unique target genes), and were further assessed based on 1) their expression in AR42J cells and 2) their response to gastrin-induced stimulation. Finally two target genes that were considered as valid hypotheses (Table 3).

Discussion

Network based analysis of biological data forms one of the cornerstones of systems biology. Finding new candidate network components is an area of active research [70, 71, 72]. Our objective was to demonstrate the use of semantic knowledge bases for such network expansion work, in order to illustrate the potential value of the semantic web for the biologists. We extended the semantic knowledge base GeXKB in order to make it more suitable for regulatory network analysis and construction. Starting from a literature-based gastrin signaling network that we built previously we chose three of its documented DNA binding transcription factors (CREB1, NFKB1 and TCF7L2) for the design of a set of biological questions that were formulated as SPARQL queries. This allowed us to retrieve 148 candidate regulators of these three DbTFs, and 20 shared target genes that are likely to be regulated by both the candidate regulators and the DbTFs.

Use Case I was designed to identify new activators of CREB1. The only known activator of CREB1 reported in the context of a gastrin-mediated response is Ribosomal S6 Kinase 1/2 (RSK1/2, see Fig. 4), a member of the 90 kDa ribosomal S6 kinase (RSK) protein family [73]. The results obtained from GeXKB suggest several other members of the RSK family to be involved in the activation of CREB1: Ribosomal protein S6 kinase alpha-4 (RPS6KA4) and Ribosomal protein S6 kinase alpha-5 (RPS6KA5), as indicated in Table 2a and Figure 3. Our literature search revealed that activation of CREB1 was indeed shown to be regulated by RPS6KA4 and RPS6KA5 [74, 75], however, only RPS6KA4 is expressed in AR42J cells and is thus an interesting candidate for experimental investigation in our gastrin response model system. Similarly, the network candidates PRKD1 (PKD1) and PRKD2 (PKD2) were reported to play a role in CREB1 activation in other cellular responses [76, 77], making them interesting candidates for AR42J experiments since they are expressed in this cell line (Table 2a). Furthermore, among the repressors, TCF7L2, SIRT1 and SIK1 (Table 2a, and Figure 4) are well documented as negative regulators of the CREB1 transcriptional complex in other experimental systems [78, 79, 80]. Proteins such as CREB-binding protein (CREBBP, also termed CBP), which have multiple functions to accommodate different contexts and environments [81, 82], also appear in the query result (see Table 2a, and Figure 4). This reflects the complexity where various factors interplay and contribute to CREB1 regulation in response to a stimulus. Taken together, our analysis of GeXKB for information relevant for the CCKR network showed that gastrin-mediated regulation of CREB1 activity involves several other proteins in addition to RSK1/2, which is the only CREB1-modulator reported so far in gastrin-responses. Rather, the cellular response outcomes upon gastrin stimuli and mediated by CREB1 are likely to be dependent on the interplay between different activators such as RPS6KA4 and PRKD1/2 and repressors such as TCF7L2, SIRT1 and SIK1, resulting in fine tuning of CREB1-mediated gene regulatory events triggered by gastrin.

For *Use Case II*, literature screening showed that several proteins, including NFKBIA, CYLD, TAX1BP1, ITCH, SIRT1 and IRAK, have been reported to undergo proteasomal degradation and are implicated in contributing to NFκB down-regulation (see Table 2b and references herein, and Figure 4). However, in the gastrin response signaling cascade, so far only NFKBIA has been experimentally shown to be associated with negative regulation of NFκB (reference in Table 2b, Figure 4). The GeXKB query result suggests additional proteins e.g. CYLD, TAX1BP1, ITCH, SIRT1 and IRAK that are documented as NFκB repressors and undergoing proteasomal degradation (see Table 2b, and Figure 4) and that can therefore be interesting to pursue in future experimental work.

Furthermore, in *Use Case III*, interestingly, the genes encoding these six proteins are all expressed in AR42J cells. Five of these proteins (see Table 2c, and Figure 4) have literature evidence indicating that they function both as activators and repressors, depending on the context. Of these six proteins, only beta-catenin (CTNNB1) has previously been shown to modulate TCF7L2 in gastrin mediated intracellular signaling. In *Use Case IV*, the result suggests that regulators CREM, FOXP3, TCF7L2, SMAD3 and PAPR1 are DbTFs and share 20 target genes that are also regulated by CREB1 and NFkB1 DbTFs (see Table 3). The genes encoding the regulators CREM, TCF7L2 and PAPR1 are found to be expressed in AR42J cells. Therefore, potential targets of any of the AR42J expressed regulators would be of greater significance, especially if these target genes would show evidence of regulation in the gastrin-induced AR42J cell system. GeXKB provided five genes (*JUN, NFkB1*, *IER3, ALOX5AP* and *BRCA2*) (see Table 3), however, only *JUN* and *BRCA2* are identified as genes that are targets of regulators (CREM and PAPR1, Figure 4), and show changed expression in AR42J cells after gastrin perturbation.

The results presented in this paper demonstrate that GeXKB can facilitate the identification of potential novel regulatory network candidates. Our analysis suggests that GeXKB also offers good perspectives as a resource for relevant information

Table 2: Regulators of DbTFs

List of proteins returned for use case I (a); use case II (b); use case III (c). Key for columns (left to right) of tables a-c: **Regulator(s):** A subset of proteins (complete list in supplementary material) that are retrieved after querying GeXKB for regulators of the targeted transcription factors. **Function:** Role of the retrieved regulators. **GeXKB result categories:** b_1 novel regulators that are expressed in AR42J cells; and b_2 novel regulators which are implicated based on observations in responses to other stimuli than gastrin. **Evidence:** Literature references for the given functionality for the transcription factor regulators in question. The PubMed IDs (PMIDs) are chosen from PubMed-based searches guided by GeXKB results.

a)

Regulator(s)	**Function**	**GeXKB result categories**		**Evidence**
		b_1	**b_2**	
CAMK1D	CREB activator		Yes	PMID:16324104
CREM	CREB activator	Yes	Yes	PMID:1370576, PMID:7961842
CRTC1	CREB activator		Yes	PMID:17565599
CRTC2	CREB activator		Yes	PMID:17565599
CRTC3	CREB activator		Yes	PMID:17565599
MAPK3	CREB activator	Yes		PMID:8688081
PRKAA1	CREB activator	Yes		PMID:19442239, PMID:18063805, PMID:17565599
PRKAA2	CREB activator	Yes		PMID:19442239, PMID:18063805
PRKD1	CREB activator	Yes		PMID:20497126, PMID:17389598
PRKD2	CREB activator	Yes		PMID:20497126, PMID:17389598
RPS6KA1	CREB activator	Yes		PMID:8688081, PMID:17565599
RPS6KA4	CREB activator	Yes	Yes	PMID:16125054
RPS6KA5	CREB activator		Yes	PMID:16125054
CEBPG	CREB repressor	Yes		Similarity bZIP
DDIT3	CREB repressor	Yes		Similarity bZIP
HDAC2	CREB repressor	Yes		Functional similarity (PMID:10669737)
HDAC4	CREB repressor	Yes		Functional similarity (PMID:10669737)
SIRT1	CREB repressor	Yes	Yes	PMID:23292070
SIK1	CREB repressor	Yes	Yes	PMID:15511237
TCF7L2	CREB repressor	Yes		PMID:23028378
CREBBP	CREB activator and chromatin modifier	Yes	Yes	PMID:11094091, PMID:9413984
EP300	CREB chromatin modifier	Yes	Yes	PMID:17565599

b)

Regulator(s)	Function	GeXKB result categories		Evidence
		b_1	**b_2**	
NFKBIA	NFκB1 and RELA repressor	Yes		PMID:12740336
COMMD1	NFκB1 and RELA repressor		Yes	PMID:15799966, PMID:16573520, PMID:20068069
COMMD7	NFκB1 and RELA repressor		Yes	PMID:15799966, PMID:16573520, PMID:20068069
TAX1BP1	NFκB1 and RELA repressor	Yes	Yes	PMID:22435550
TNFAIP3	NFκB1 and RELA repressor		Yes	PMID:19494296, PMID:19608751, PMID:16684768, PMID:19380639
ITCH	NFκB1 and RELA repressor	Yes	Yes	PMID:22435550, PMID:21119682
CYLD	NFκB1 and RELA repressor	Yes	Yes	PMID:22435550, PMID:21119682, PMID:19373246
SIRT1	NFκB1 and RELA repressor	Yes	Yes	PMID:19373246
CTNNB1	NFκB1 and RELA repressor	Yes		PMID:12398896, PMID:14991743
IRAK1	NFκB1 and RELA repressor	Yes	Yes	PMID:20300215, PMID:17457343
IRAK2	NFκB1 and RELA repressor	Yes	Yes	PMID:20300215, PMID:17457343
IRAK3	NFκB1 and RELA repressor	Yes	Yes	PMID:20300215, PMID:17457343
ZNF675	NFκB1 and RELA repressor		Yes	PMID:11751921

c)

Regulator(s)	Function	GeXKB result categories		Evidence
		b_1	**b_2**	
PARP1	TCF repressor, NFκB activator	Yes	Yes	PMID:17504138, PMID:19060926
RUNX3	TCF repressor, NFκB activator	Yes	Yes	PMID:18772112, PMID:21523770
CTNB1	TCF repressor	Yes		PMID:20122174
XRCC5	TCF repressor, NFκB activator	Yes	Yes	PMID:17283121, PMID:17031478
XRCC6	TCF repressor, NFκB activator	Yes	Yes	PMID:17283121, PMID:17031478
DAXX	TCF repressor	Yes	Yes	PMID:16569639

Table 3: DbTF – target gene categorisation

The table lists shared target genes of the novel DbTFs and CCKR core DbTFs, retrieved through Use Case I-III. Key for columns (left to right): **Novel DbTFs:** Proteins that regulate the core CCKR-DbTFs (CREB1, NFkB1 and TCF7L2) transcriptionally; **Function:** Role of the regulators; **CCKR-DbTF:** core CCKR-DbTF that is regulated by the candidates from column one; **TGs:** Target genes retrieved from GeXKB that are found to be common between the novel DbTFs and the CCKR core DbTF(s); **AR42J expressed:** target genes that are expressed in AR42J cells; **Gastrin responsive:** target genes that show change in gene expression after gastrin treatment..

.

Novel DbTF	Function	CCKR DbTF	TGs	AR42J expressed	Gastrin responsive
CREM	activator	CREB1	JUN	Yes	Yes
FOXP3	repressor	CREB1	IFNG	No	
	repressor	CREB1	IL10	No	
	repressor	CREB1	BCL2	No	
	repressor	CREB1	MALAT1	No	
TCF7L2	repressor	CREB1	MYOD1	No	
FOXP3	repressor	NFkB1	PIGR	No	
	repressor	NFkB1	CXCL5	No	
	repressor	NFkB1	VCAM1	No	
	repressor	NFkB1	VWF	No	
	repressor	NFkB1	IFNG	No	
	repressor	NFkB1	IL8	No	
	repressor	NFkB1	BCL2A1	No	
	repressor	NFkB1	NFKB1	Yes	Yes
	repressor	NFkB1	IER3	Yes	Yes
	repressor	NFkB1	CD40LG	No	
	repressor	NFkB1	SELE	No	
	repressor	NFkB1	ALOX5AP	Yes	Yes
SMAD3	repressor	NFkB1	MMP9	No	
PARP1	activator	NFkB1	BRCA2	Yes	Yes

concerning transcription factor regulation. Obviously the involvement of the candidate regulators and target genes in gastrin mediated regulation of transcription factors requires further experimental validation. The observations made through small scale experiments such as RNAi-mediated knock-down of novel regulators or large-scale studies on knock-out model organisms can greatly enhance our current understanding of gastrin mediated transcription regulation and subsequent cellular

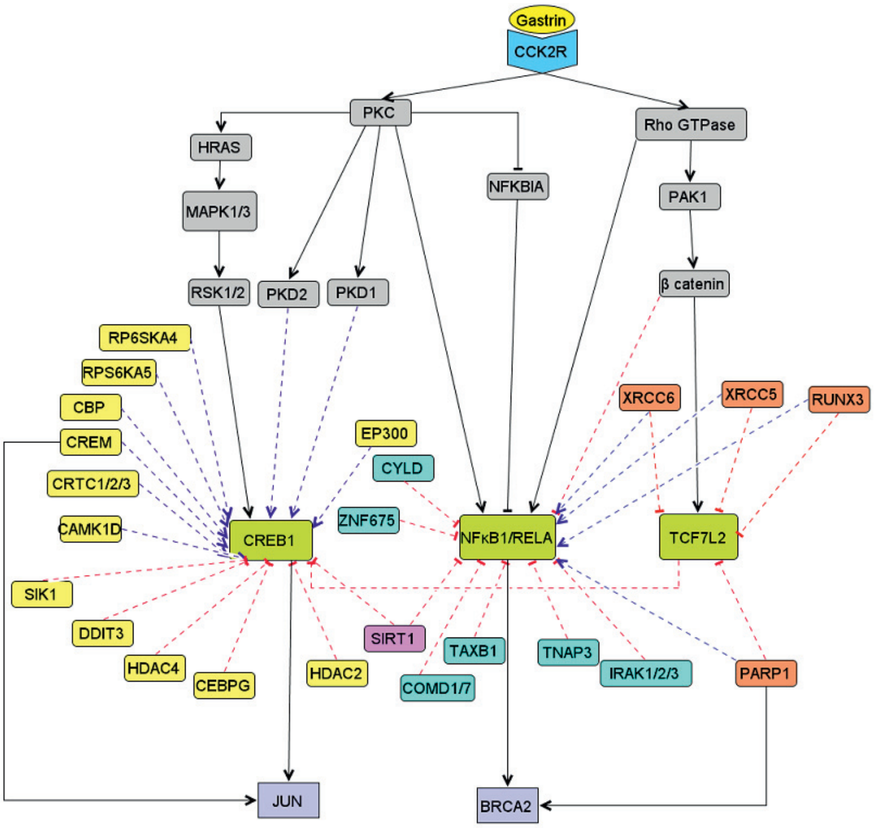


Figure 4: Core CCKR network and novel candidate regulators

The core of the Gastrin mediated signal transduction network (CCKR) and novel candidate regulators resulting from our queries are shown. The CCKR DbTFs that were targeted in our queries are colored *light green*. The network components in *grey* and the *solid lines* connecting them are part of the core CCKR network and documented as regulators of the CCKR DbTFs and respond to gastrin. The *dotted lines* represent new relations identified by the queries which could be verified against literature: *blue pointed* arrows denote ‘activation or positive influence’ and *red bar-headed* arrows depict ‘repression or negative influence’. CREB1 candidate regulators identified through Q1, Q2 and Q3 are colored *yellow*. Candidate regulators of NFκB1 identified through Q4 are colored *turquoise*, and candidate regulators of TCF7L2 identified through Q5 are colored *orange*. The *pink-colored* candidate is identified by both Q2 and Q4. The target genes shared by the CCKR DbTFs (CREB1 and NFκB1) and the DbTF candidates identified through Q6 are colored *light blue* and their connections are shown as *solid* arrows.

outcomes. Alternatively, information from gene expression databases such as ArrayExpress could already provide some evidence to elucidate the role of the candidate regulators. We searched for gene knockout experiments for the candidates in ArrayExpress and found evidence supporting two regulators. Proteins CRTC1 and COMMD1 were among the novel regulators suggested by GeXKB to play a role in gastrin mediated transcriptional regulation. Gene knock-out experiments indicate CRTC1 and COMMD1 as a potential regulator of CREB1 (ArrayExpress accession: E-GEOD-12209) and NFkB1 (ArrayExpress accession: E-MEXP-832) respectively.

Conclusions

Our work demonstrates the level of knowledge discovery that can be achieved when knowledge from a broad range of GO annotations and experimental evidence is semantically integrated. Interlinking various data sets using RDF provides the much needed homogeneity and extensibility for advanced data analysis. Additionally, we have shown the implications of using computational inferencing in building the knowledgebase, as this approach allows the retrieval of information that would otherwise have remained implicit and hidden from querying. Our efforts have involved a close collaboration between semantic web specialists and biological domain experts, resulting in novel ways for generating hypotheses and an initial assessment of these hypotheses against the current understanding of a regulatory network.

The utility of GeXKB is expected to grow with its further development. The goal for future releases will be to expand the knowledge base with additional high quality datasets which will include relations between DbTFs and other interactors from curated texts, partially based on our current work on checking the full repertoire of transcription factors of human, mouse and rat, and their respective target genes.

Authors' contributions

AV and ST designed the experiments and wrote the manuscript; AV carried out the querying exercise and contributed in the data integration pipeline. ST analyzed the results for biological relevance. ASG and WB contributed to the integration of data from databases HTRIdb, TFcheckpoint and TFactS. VM helped in the design of the GeXKB project and its implementation, and reviewed the manuscript. AL guided the development of the use cases, assisted in the interpretation of the results and reviewed the manuscript. MK conceived the GeXKB project, helped with the designing the use cases, and reviewed and revised the manuscript. All the authors approved the final manuscript.

Acknowledgements

Technical support was provided by the High-Performance Computing team at the Norwegian University of Science and Technology.

References

1. Weake VM, Workman JL: **Inducible gene expression: diverse regulatory mechanisms**. *Nature reviews. Genetics* 2010, **11**: 426–37.

2. Perissi V, Jepsen K, Glass CK, Rosenfeld MG: **Deconstructing repression: evolving models of co-repressor action**. *Nature reviews. Genetics* 2010, **11**: 109–23.

3. Thomas MC, Chiang CM: **The general transcription machinery and general cofactors**. *Critical reviews in biochemistry and molecular biology* 2006, **41**: 105–78.

4. Mitchell PJ, Tjian R: **Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins**. *Science* 1989, **245**: 371–8.

5. Davidson SB, Overton C, Buneman P: **Challenges in Integrating Biological Data Sources**. *Journal of Computational Biology* 1995, **2**: 557-572.

6. Goble C, Stevens R: **State of the nation in data integration for bioinformatics**. *Journal of Biomedical Informatics* 2008, **41**: 687-693.

7. Smith B, Ashburner M, Rosse C, *et al.*: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration**. *Nat Biotechnol* 2007, **25**: 1251–1255.

8. Ashburner M, Ball CA, Blake JA, *et al.*: **Gene ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**: 25-9.

9. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R: **The GOA database in 2009--an integrated Gene Ontology Annotation resource**. *Nucleic Acids Research* 2009*,* **37**: D396-D403.

10. Berners-Lee T, Hendler J: **Publishing on the semantic web**. *Nature* 2001*,* **410**: 1023-4.

11. **Resource Description Framework** [http://www.w3.org/RDF/]

12. **RDF Schema** [http://www.w3.org/TR/2004/REC-rdf-schema-20040210/]

13. **Web Ontology Language** [http://www.w3.org/TR/owl2-profiles/]

14. **SPARQL Query Language** [http://www.w3.org/TR/rdf-sparql-query/]

15. Antezana E, Egaña M, Blondé W, Illarramendi A, Bilbao I, De Baets B, Stevens R, Mironov V, Kuiper M: **The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process**. *Genome Biology* 2009*,* **10**: R58.

16. Antezana E, Blondé W, Egaña M, Rutherford A, Stevens R, De Baets B, Mironov V, Kuiper M: **BioGateway: a semantic systems biology tool for the life sciences**. *BMC Bioinformatics* 2009*,* **10**: S11.

17. **Semantic Systems Biology** [http://www.semantic-systems-biology.org]

18. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: **Bio2RDF: towards a mashup to build bioinformatics knowledge systems**. *Journal of Biomedical Informatics* 2008*,* **41**: 706-16.

19. Momtchev V, Peychev D, Primov T, Georgiev G: **Expanding the Pathway and Interaction Knowledge in Linked Life Data**. *In: International Semantic Web Challenge* 2009.

20. Carbon S, Ireland A, Mungall CJ, *et al.*: **AmiGO: online access to ontology and annotation data**. *Bioinformatics* 2009*,* **25**(2): 288-289.

21. Mercadé J, Espinosa A, Adsuara J E, Adrados R, Segura J, Maes T: **Orymold: ontology based gene expression data integration and analysis tool applied to rice**. *BMC bioinformatics* 2009, **10**(1), 158.

22. Baitaluk M, Kozhenkov S, Dubinina Y, Ponomarenko J: **IntegromeDB: an integrated system and biological search engine**. *BMC Genomics* 2012*,* **13**(1): 35.

23. **Biological Networks application** [http://www.biologicalnetworks.org/]

24. Pico AR, Kelder T, van Iersel MP, *et al.*: **WikiPathways: pathway editing for the people**. *PLoS biology* 2008*,* **6**(7): e184.

25. **WikiPathways SPARQL endpoint** [http://sparql.wikipathways.org/]

26. Bernstein A, Kaufmann E, Kaiser C: **Querying the semantic web with ginseng: A guided input natural language search engine**. *15th Workshop on Information Technologies and Systems:* 2005; *Las Vegas* : 112-126.

27. Lehmann J, Bühmann L: **AutoSPARQL: Let users query your knowledge base**. *The Semantic Web: Research and Applications* : 2011; Springer Berlin Heidelberg: 63-79.

28. Ngonga N, Bühmann L, Unger C, Lehmann, J, Gerber D: **Sorry, i don't speak SPARQL: translating SPARQL queries into natural language**. *Proceedings of the 22nd international conference on World Wide Web* 2013; International World Wide Web Conferences Steering Committee.

29. Venkatesan A, Mironov V, Kuiper M: **Towards an integrated knowledge system for capturing gene expression events**. *Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO)*: 2012; Graz, Austria.

30. Selvik LK, Fjeldbo CS, Flatberg A, Steigedal TS, Misund K, Anderssen E, Doseth B, Langaas M, Tripathi S, Beisvag V, Lægreid A, Thommesen L, Bruland T: **The duration of gastrin treatment affects global gene**

31. Funahashi A, Morohashi M, Kitano H: **CellDesigner: a process diagram editor for gene-regulatory and biochemical networks**. *Biosilico* 2003, **1**(5): 159-162.

32. Tripathi S: **Laying the foundations for Gastrin Systems Biology: Conceptual models and knowledge resources to enhance research on gastrin mediated intracellular signaling and gene regulation.** *PhD thesis*. Norwegian University of Science and Technology, Department of Cancer Research and Molecular Medicine; 2013.

33. Rustici G, Kolesnikov N, Brandizi M, *et al.*: **ArrayExpress update—trends in database growth and links to data analysis tools**. *Nucleic acids research* 2013, **41**(D1): D987-D990.

34. **Semanticscience Integrated Ontology** [https://code.google.com/p/semanticscience/wiki/SIO]

35. Demir E, Cary MP, Paley S, *et al.*: **The BioPAX community standard for pathway data sharing**. *Nature Biotechnology* 2010, **28**: 935–942

36. Brinkman RR, Courtot M, Derom D, *et al.*: **Modeling biomedical experimental processes with OBI**. *J Biomed Semantics* 2010, 1 (Suppl 1): S7.

37. Degtyarenko K, de Matos P, Ennis M, *et al.*: **ChEBI: a database and ontology for chemical entities of biological interest**. *Nucleic Acids Res.* 2008*,* **36:** D344–D350.

38. **Information Artifact Ontology** [https://code.google.com/p/information-artifact-ontology/]

39. Kerrien S, Orchard S, Montecchi-Palazzi L, *et al.*: **Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions**. *BMC Biol.* 2007, **9**(5): 44.

40. Blondé W, Mironov V, Venkatesan A, *et al.*: **Reasoning with bio-ontologies: using relational closure rules to enable practical querying**. *Bioinformatics* 2011, **27**: 1562-8.

41. Smith B, Ceusters W, Klagges B, *et al.*: **Relations in biomedical ontologies**. *Genome biology* 2005, **6**(5): R46.

42. Magrane M: **UniProt Knowledgebase: a hub of integrated protein data**. *Database: the journal of biological databases and curation* 2011.

43. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 2000*,* **28**: 27-30.

44. Kerrien S, Alam-Faruque Y, Aranda B, *et al.*: **IntAct - open source resource for molecular interaction data**. *Nucleic Acids Res* 2007, **35**: D561-565.

45. Wheeler DL, Barrett T, Benson DA, *et al.*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2005*,* **33**: D39-45.

46. Ekseth OK, Kuiper M, Mironov V: **orthAgogue: an agile tool for the rapid prediction of orthology relations**. *Bioinformatics* 2013, btt582.

47. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome Res* 2003, **13**: 2178-2189.

48. Casamar PE, Arenillas D, Lim J, *et al.*: **The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences**. *Nucleic Acids Res.* 2009, **37** (Database issue): D54-60.

49. Bovolenta LA, Acencio ML, Lemke N: **HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions**. *BMC Genomics* 2012, **13**(1): 405.

50. Essaghir A, Toffalini F, Knoops L, *et al.*: **Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data**. *Nucleic acids research* 2010, **38**(11): e120-e120.

51. Chawla K, Tripathi S, Thommesen L, Lægreid A, Kuiper M: **TFcheckpoint: a curated compendium of specific DNA-binding RNA polymerase II transcription factors**. *Bioinformatics* 2013, **29**(19): 2519-2520.

52. Antezana E, Egaña M, De Baets B, Kuiper M, Mironov V: **ONTO-PERL: an API for supporting the development and analysis of bio-ontologies**. *Bioinformatics* 2008, **24**: 885-7.

53. **OBO format** [http://www.geneontology.org/GO.format.obo-1_2.shtml]

54. **DOT format** [http://www.graphviz.org/doc/info/lang.html]

55. **XML format** [http://www.w3.org/TR/xml/]

56. Jupp S, Klein J, Schanstra J, Stevens R: **Developing a kidney and urinary pathway knowledge base**. *J Biomed Semantics* 2011*,* **17**(2): S7.

57. **Openlink Virtuoso** [http://virtuoso.openlinksw.com]

58. **GeXKB SPARQL endpoint** [http://www.semantic-systems-biology.org/apo/queryingcco/sparql]

59. Blondé W, Antezana E, Mironov V, *et al.*: **Using the relation ontology Metarel for modelling Linked Data as multi-digraphs**. *Semantic Web Journal* 2013.

60. **SPARQL update language** [http://www.w3.org/TR/sparql11-update/]

61. Heath T, Bizer C: **Linked Data: Evolving the Web into a Global Data Space (1st edition)**. *Synthesis Lectures on the Semantic Web: Theory and Technology* 2011, **1**(1): 1-136.

62. Hubbard T, Barker D, Birney E, *et al.*: **The Ensembl genome database project**. *Nucleic acids research* 2002, **30**(1): 38-41.

63. Noy NF, Shah NH, Whetzel PL, *et al.*: **BioPortal: ontologies and integrated data resources at the click of a mouse**. *Nucleic Acids Res* 2009, **37**(Web Server issue): W170-173.

64. Dolcet X, Llobet D, Pallares J, Matias-Guiu X: **NF-kB in development and progression of human cancer**. *Virchows Archiv : an international journal of pathology* 2005, **446**: 475–82.

65. Hiraoka S, Miyazaki Y, Kitamura S, *et al.*: **Gastrin induces CXC chemokine expression in gastric epithelial cells through activation of NF-kappaB**. *American journal of physiology. Gastrointestinal and liver physiology* 2001, **281**: G735–42.

66. Varro A, Noble PJ, Pritchard DM, *et al.*: **Helicobacter pylori induces plasminogen activator inhibitor 2 in gastric epithelial cells through nuclear factor-kappaB and RhoA: implications for invasion and apoptosis**. *Cancer research* 2004, **64**: 1695–702.

67. He H, Shulkes A, Baldwin GS: **PAK1 interacts with beta-catenin and is required for the regulation of the beta-catenin signalling pathway by gastrins**. *Biochimica et biophysica acta* 2008, **1783**: 1943–54.

68. Pradeep A, Sharma C, Sathyanarayana P, *et al.*: **Gastrin-mediated activation of cyclin D1 transcription involves beta-catenin and CREB pathways in gastric cancer cells**. *Oncogene* 2004, **23**: 3689–99.

69. Subramaniam D, Ramalingam S, May R, *et al.*: **Gastrin-mediated interleukin-8 and cyclooxygenase-2 gene expression: differential transcriptional and posttranscriptional mechanisms**. *Gastroenterology* 2008, **134**: 1070–82.

70. Tipney HJ, Leach SM, Feng W, Spritz R, Williams T, Hunter L: **Leveraging existing biological knowledge in the identification of candidate genes for facial dysmorphology**. *BMC bioinformatics* 2009, **10** (Suppl 2): S12.

71. Wu G, Stein L: **A network module-based method for identifying cancer prognostic signatures**. *Genome biology* 2012, **13**: R112.

72. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes**. *American journal of human genetics* 2006, **78**: 1011–25.

73. Hauge C, Frödin M: **RSK and MSK in MAP kinase signalling**. *Journal of cell science* 2006, **119**: 3021–3.

74. Delghandi MP, Johannessen M, Moens U: **The cAMP signalling pathway activates CREB through PKA, p38 and MSK1 in NIH 3T3 cells**. *Cellular signaling* 2005, **17**: 1343–51.

75. Wu GY, Deisseroth K, Tsien RW: **Activity-dependent CREB phosphorylation: convergence of a fast, sensitive calmodulin kinase pathway and a slow, less sensitive mitogen-activated protein kinase pathway**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**: 2808–13.

76. Johannessen M, Delghandi MP, Rykx A, Dragset M, Vandenheede JR, Van Lint J, Moens U: **Protein kinase D induces transcription through direct phosphorylation of the cAMP-response element-binding protein**. *The Journal of biological chemistry* 2007, **282**: 14777–87.

77. Evans IM, Bagherzadeh A, Charles M, Raynham T, Ireson C, Boakes A, Kelland L, Zachary IC: **Characterization of the biological effects of a novel protein kinase D inhibitor in endothelial cells**. *The Biochemical journal* 2010, **429**: 565–72.

78. Oh KJ, Park J, Kim SS, Oh H, Choi CS, Koo SH: **TCF7L2 modulates glucose homeostasis by regulating CREB- and FoxO1-dependent transcriptional pathway in the liver**. *PLoS genetics* 2012, **8**: e1002986.

79. Monteserin GJ, Al-Massadi O, Seoane LM, *et al.*: **Sirt1 inhibits the transcription factor CREB to regulate pituitary growth hormone synthesis**. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 2013, **1**: 11.

80. Katoh Y, Takemori H, Min L, Muraoka M, Doi J, Horike N, Okamoto M: **Salt-inducible kinase-1 represses cAMP response element-binding protein activity both in the nucleus and in the cytoplasm**. *European journal of biochemistry / FEBS* 2004, **271**: 4307–19.

81. Shaywitz AJ, Dove SL, Kornhauser JM, Hochschild A, Greenberg ME: **Magnitude of the CREB-dependent transcriptional response is determined by the strength of the interaction between the kinase-inducible domain of CREB and the KIX domain of CREB-binding protein**. *Molecular and cellular biology* 2000, **20**(24): 9409-9422.

82. Radhakrishnan I, Pérez-Alvarado GC, Parker D, Dyson HJ, Montminy MR, Wright PE: **Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions**. *Cell* 1997, **91**: 741–52.

Supplementary material

Q1:

Biological Question: List of proteins involved in activation of CREB1 Transcription factor

Parameters:

- GO_0032793 - positive regulation of CREB transcription factor activity,
- GO_0051091 - positive regulation of sequence-specific DNA binding transcription factor activity + MI_0914 – association,
- MI_0407 - Direct interactors,
- GO_0008140 - cAMP response element binding protein binding

SPARQL query:

```
BASE   <http://www.semantic-systems-biology.org/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>

PREFIX term: <SSB#UniProtKB_P16220> # CREB1

SELECT distinct ?gene ?name ?description ?dbtf ?protein
WHERE {
 GRAPH <ReTO> {
  ?protein ssb:has_source ssb:NCBITaxon_9606 .
  ?protein ssb:Definition ?d .
  ?d ssb:def ?description .
  ?g ssb:codes_for ?protein .
  ?g rdfs:label ?gene .
  ?protein rdfs:label ?name .
 }
 {
  GRAPH <ReTO> {
   ssb:GO_0032793 ssb:has_participant ?protein .
  }
 }UNION {
  GRAPH <ReTO-tc> {
   ?biological_process ssb:is_a ssb:GO_0051091 .
   ?biological_process ssb:has_participant ?protein .
   ?interaction ssb:is_a ssb:MI_0914 .
   ?interaction ssb:has_agent ?protein .
   ?interaction ssb:has_agent term: .
  }
 } UNION {
  GRAPH <ReTO-tc> {
   ?interaction ssb:is_a ssb:MI_0407 .
   ?interaction ssb:has_agent ?protein .
   ?interaction ssb:has_agent term: .
  }
 } UNION {
  GRAPH <ReTO> {
   ?protein ssb:has_function ssb:GO_0008140 .
  }
 }
   FILTER (?protein != term:)
 OPTIONAL {
  GRAPH <tfcheckpoint> {
   ?protein ssb:is_dbtf ?dbtf.
  }
 }
}

ORDER BY ?gene
```

Q2:

Biological Question: Name transcriptional repressors of CREB1 Transcription factor

Parameters:

- GO_0043433 - negative regulation of sequence-specific DNA binding transcription factor activity,
- GO_0032792 - negative regulation of CREB transcription factor activity

SPARQL query:

```
BASE   <http://www.semantic-systems-biology.org/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>

SELECT distinct ?gene ?name ?description ?dbtf ?repressor
WHERE {
  GRAPH <ReTO> {
   ?repressor ssb:has_source ssb:NCBITaxon_9606 .
   ?repressor rdfs:label ?name .
   ?repressor ssb:Definition ?d .
   ?d ssb:def ?description .
   ?g ssb:codes_for ?repressor .
   ?g rdfs:label ?gene .
  }
  GRAPH <ReTO> {
   {
    ssb:GO_0043433 ssb:has_participant ?repressor .
   } UNION {
    ssb:GO_0032792 ssb:has_participant ?repressor .
   }
  }
  OPTIONAL {
  GRAPH <tfcheckpoint> {
   ?repressor ssb:is_dbtf ?dbtf.
  }
 }
}

ORDER BY ?gene
```

Q3:

Biological Question: List chromatin modifiers which are part of CREB transcription factor complex

Parameters:

- GO_0004402 - histone acetyltransferase activity,
- GO_0004407 - histone deacetylase activity,
- GO_0051090 - regulation of sequence-specific DNA binding transcription factor activity
- GO_0005667 - transcription factor complex + MI_0914 - association

SPARQL Query:

```
BASE   <http://www.semantic-systems-biology.org/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>

PREFIX term: <SSB#UniProtKB_P16220> # CREB1-P16220

SELECT distinct ?gene ?chromatin_modifier_name ?description ?chromatin_modifier
WHERE {
 GRAPH <ReTO> {
  ?chromatin_modifier ssb:has_source ssb:NCBITaxon_9606 .
  ?chromatin_modifier ssb:Definition ?Def .
  ?Def  ssb:def   ?description .
  ?chromatin_modifier rdfs:label ?chromatin_modifier_name .
  ?g ssb:codes_for ?chromatin_modifier .
  ?g rdfs:label ?gene .
 }
 GRAPH <ReTO-tc> {
  {
    ?chromatin_modifier ssb:has_function ?function .
    ?function ssb:is_a ssb:GO_0004402 .
  } UNION {
    ?chromatin_modifier ssb:has_function ?function .
    ?function ssb:is_a ssb:GO_0004407 .
  }
  ssb:GO_0051090 ssb:has_participant ?chromatin_modifier .
 }
 GRAPH <ReTO-tc> {
  ?complex ssb:is_a ssb:GO_0005667 .
  ?complex ssb:contains ?chromatin_modifier .
  ?complex ssb:contains term: .
  OPTIONAL {
   ?interaction ssb:is_a ssb:MI_0914 .
   ?interaction ssb:has_agent term: .
   ?interaction ssb:has_agent ?chromatin_modifier .
  }
 }
}

ORDER BY ?gene
```

Q4:

Biological Question: List all transcriptional repressors of Transcription factors NFkB1 and RELA which undergoes proteosomal degradation.

Parameters:

- GO_0032088 - negative regulation of NF-kappaB transcription factor activity
- KEGG_ko04120 - Ubiquitin mediated proteolysis,
- GO_0000151 - ubiquitin ligase complex,
- GO_0043130 - ubiquitin binding,
- MI_0220 - ubiquitination reaction

SPARQL Query:

```
BASE   <http://www.semantic-systems-biology.org/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>

SELECT distinct ?gene ?name ?description ?dbtf ?protein
WHERE {
  GRAPH <ReTO> {
  ?protein ssb:has_source ssb:NCBITaxon_9606 .
  ?protein rdfs:label ?name .
  ?protein ssb:Definition ?d .
  ?d ssb:def ?description .
  ?g ssb:codes_for ?protein .
  ?g rdfs:label ?gene .
 }
 {
  GRAPH <ReTO-tc> {
   ?cellular_component ssb:is_a ssb:GO_0000151 .
   ?cellular_component ssb:contains ?protein .
  }
  GRAPH <ReTO> {
   ssb:GO_0032088 ssb:has_participant ?protein .
  }
 }
 UNION {
  GRAPH <ReTO> {
   ssb:GO_0032088 ssb:has_participant ?protein .
  }
  GRAPH <ReTO> {
   ssb:KEGG_ko04120 ssb:has_agent ?protein_cluster .
   ?q_prot ssb:is_member_of ?protein_cluster .
  }
  GRAPH <ReTO> {
   ?interaction ssb:is_a ssb:MI_0915 .
   ?interaction ssb:has_agent ?protein .
   ?interaction ssb:has_agent ?q_prot .
  }
 }
 UNION {
  GRAPH <ReTO-tc> {
   ?function ssb:is_a ssb:GO_0043130 .
   ?protein ssb:has_function ?function .
   ssb:GO_0032088 ssb:has_participant ?protein .
  }
 }
 UNION {
  GRAPH <ReTO-tc> {
   ?interaction ssb:is_a ssb:MI_0220 .
   ?interaction ssb:has_agent ?protein .
   ssb:GO_0032088 ssb:has_participant ?protein .
  }
 }
  OPTIONAL {
  GRAPH <tfcheckpoint> {
   ?protein ssb:is_dbtf ?dbtf.
  }
 }
}

ORDER BY ?gene
```

Q5:

Biological Question: List all transcriptional repressors of TCF7L2 which are activators of NFkB1 or CREB1.

Parameters:

- GO_0043433 - negative regulation of sequence-specific DNA binding transcription factor activit
- GO_0051091 - positive regulation of sequence-specific DNA binding transcription factor activity,
- MI_0914 - association

SPARQL Query:

```
BASE <http://www.semantic-systems-biology.org/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>

PREFIX term: <SSB#UniProtKB_Q9NQB0> # TCF7L2

SELECT distinct ?gene ?name ?description ?dbtf ?protein
WHERE {
 GRAPH <ReTO> {
  ?protein ssb:has_source ssb:NCBITaxon_9606 .
  ?protein rdfs:label ?name .
  ?protein ssb:Definition ?d .
  ?d ssb:def ?description .
  ?g ssb:codes_for ?protein .
  ?g rdfs:label ?gene .
 }
 GRAPH <ReTO-tc> {
  ?interaction ssb:is_a ssb:MI_0914 .
  ?interaction ssb:has_agent term: .
  ?interaction ssb:has_agent ?protein .
  FILTER (?protein != term:)
 }
 GRAPH <ReTO-tc> {
  ?biological_process1 ssb:is_a ssb:GO_0043433 .
  ?biological_process1 ssb:has_participant ?protein .
  ?biological_process2 ssb:is_a ssb:GO_0051091 .
  ?biological_process2 ssb:has_participant ?protein .
 }
  OPTIONAL {
  GRAPH <tfcheckpoint> {
   ?protein ssb:is_dbtf ?dbtf.
  }
 }
}

ORDER BY ?gene
```

Q6:

Biological Question: Identification of shared target genes between regulators and their DbTFs

Parameters:

- Regulators retrieved from Q1, Q2, Q4 and Q5 that are DbTF,
- DbTF of interest (CREB1, NFKB1 and TCF7L2)

SPARQL Query:

```
BASE   <http://www.semantic-systems-biology.org/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>

PREFIX reg_term:<SSB#UniProtKB_Q9BZS1>  # FOXP3, DbTF terms to be changed
accordingly

PREFIX creb:<SSB#UniProtKB_P16220> # CREB1 protein term
PREFIX nfkb:<SSB#UniProtKB_P19838> # NFKB1 protein term
PREFIX tcf7l2:<SSB#UniProtKB_Q9NQB0> # TCF7L2 protein term

SELECT distinct ?name ?tg
WHERE {
 {
  GRAPH <htridb> {
   reg_term: ssb:acts_on ?tg.
   creb: ssb:acts_on ?tg.
#   nfkb: ssb:acts_on ?tg.
#   tcf7l2: ssb:acts_on ?tg.
   ?tg rdfs:label ?name.
  }
 }
 UNION {
  GRAPH <tfacts> {
   reg_term: ssb:acts_on ?tg.
   creb: ssb:acts_on ?tg.
#   nfkb: ssb:acts_on ?tg.
#   tcf7l2: ssb:acts_on ?tg.
   ?tg rdfs:label ?name.
  }
 }
 UNION {
  GRAPH <UP-IDMAP> {
   reg_term: ssb:ensembl_trs ?term_mrna.
   creb: ssb:ensembl_trs ?creb_mrna.
#   nfkb: ssb:ensembl_trs ?nfkb_mrna.
#   tcf7l2: ssb:ensembl_trs ?tcf7l2_mrna.
  }
  GRAPH <PAZAR> {
   ?term_mrna ssb:acts_on ?tg.
   ?creb_mrna ssb:acts_on ?tg.
#   ?nfkb_mrna ssb:acts_on ?tg.
#   ?tcf7l2_mrna ssb:acts_on ?tg.
   OPTIONAL {
     GRAPH ?g {
      ?tg rdfs:label ?name.
     }
   }
  }
 }
}
```

Uniprot-ID	Description	Uniprot-Accession	Q1	Q2	Q3	Q4	Q5	CCK2R model protein (*a*)	AR42J expressed (b_1)	Response to other stimuli (b_2)	Evidence
CEBPG_HUMAN	CCAAT&#47;enhancer-binding protein gamma.	P53567		CREB1 repressor					Yes		Similarity bZIP
CNBP1_HUMAN	Beta-catenin-interacting protein 1.	Q9NSA3		CREB1 repressor					Yes		
Q5T4V2_HUMAN	CTNNBIP1 protein.	Q5T4V2		CREB1 repressor							
DDIT3_HUMAN	DNA damage-inducible transcript 3 protein.	P35638		CREB1 repressor					Yes		Similarity bZIP
EGLN1_HUMAN	Egl nine homolog 1.	Q9GZT9		CREB1 repressor					Yes		
FLNA_HUMAN	Filamin-A.	P21333		CREB1 repressor					Yes		
FOXA2_HUMAN	Hepatocyte nuclear factor 3-beta.	Q9Y261		CREB1 repressor					Yes		
FOXP3_HUMAN	Forkhead box protein P3.	Q9BZS1		CREB1 repressor		NFkB1 repressor					
FOXS1_HUMAN	Forkhead box protein S1.	O43638		CREB1 repressor							
FZD6_HUMAN	Frizzled-6 precursor.	O60353		CREB1 repressor							
GLIS2_HUMAN	Zinc finger protein GLIS2.	Q9BZE0		CREB1 repressor					Yes		
HDAC2_HUMAN	Histone deacetylase 2.	Q92769		CREB1 repressor					Yes		Functional similarity (PMID:10669737)
HDAC4_HUMAN	Histone deacetylase 4.	P56524		CREB1 repressor					Yes		Functional similarity (PMID:10669737)
HMOX1_HUMAN	Heme oxygenase 1.	P09601		CREB1 repressor					Yes		
HNF4A_HUMAN	Hepatocyte nuclear factor 4-alpha.	P41235		CREB1 repressor							
HAIR_HUMAN	Protein hairless.	O43593		CREB1 repressor							
ID1_HUMAN	DNA-binding protein inhibitor ID-1.	P41134		CREB1 repressor					Yes		

ID2_HUMAN	DNA-binding protein inhibitor ID-2.	Q02363	CREB1 repressor			Yes		
ID3_HUMAN	DNA-binding protein inhibitor ID-3.	Q02535	CREB1 repressor			Yes		
IRAK2_HUMAN	Interleukin-1 receptor-associated kinase-like 2.	O43187	CREB1 repressor	NFkB1 repressor		Yes	Yes	PMID:20300215, 17457343
KDM1A_HUMAN	Lysine-specific histone demethylase 1A.	O60341	CREB1 repressor					
M3K10_HUMAN	Mitogen-activated protein kinase kinase kinase 10.	Q02779	CREB1 repressor			Yes		
MEN1_HUMAN	Menin.	O00255	CREB1 repressor		Yes	Yes		
MSX2_HUMAN	Homeobox protein MSX-2.	P35548	CREB1 repressor					
NR0B1_HUMAN	Nuclear receptor subfamily 0 group B member 1.	P51843	CREB1 repressor					
NR0B2_HUMAN	Nuclear receptor subfamily 0 group B member 2.	Q15466	CREB1 repressor			Yes		
PEX14_HUMAN	Peroxisomal membrane protein PEX14.	O75381	CREB1 repressor			Yes		
PIM1_HUMAN	Serine&#47;threonine-protein kinase pim-1.	P11309	CREB1 repressor			Yes		
PA2GX_HUMAN	Group 10 secretory phospholipase A2 precursor.	O15496	CREB1 repressor					
PRIO_HU	Major prion	P04156	CREB1			Yes		

MAN	protein precursor.			repressor					
PROX1_HUMAN	Prospero homeobox protein 1.	Q92786		CREB1 repressor					
PTHR_HUMAN	Parathyroid hormone-related protein precursor.	P12272		CREB1 repressor			Yes		
Q53XY9_HUMAN	Parathyroid hormone-like hormone.	Q53XY9		CREB1 repressor					
RB_HUMAN	Retinoblastoma-associated protein.	P06400		CREB1 repressor			Yes		
RNF12_HUMAN	E3 ubiquitin-protein ligase RLIM.	Q9NVW2		CREB1 repressor			Yes		
SFRP4_HUMAN	Secreted frizzled-related protein 4 precursor.	Q6FHJ7		CREB1 repressor					
SFRP5_HUMAN	Secreted frizzled-related protein 5 precursor.	Q5T4F7		CREB1 repressor					
SIGIR_HUMAN	Single Ig IL-1-related receptor.	Q6IA17		CREB1 repressor					
SIK1_HUMAN	Serine&#47;threonine-protein kinase SIK1.	P57059	CREB1 activator	CREB1 repressor			Yes	Yes	PMID:15511237
SIRT1_HUMAN	NAD-dependent deacetylase sirtuin-1.	Q96EB6		CREB1 repressor		NFkB1 repressor	Yes	Yes	PMID:23292070, 19373246
SMAD7_HUMAN	Mothers against decapentaplegic homolog 7.	O15105		CREB1 repressor			Yes		
SP100_HUMAN	Nuclear autoantigen Sp-100.	P23497		CREB1 repressor					

SUMO1_HUMAN	Small ubiquitin-related modifier 1 precursor.	P63165		CREB1 repressor			Yes	
TAF3_HUMAN	Transcription initiation factor TFIID subunit 3.	Q5VWG9		CREB1 repressor				
TF7L2_HUMAN	Transcription factor 7-like 2.	Q9NQB0		CREB1 repressor		Yes	Yes	PMID:23028378
TNR4_HUMAN	Tumor necrosis factor receptor superfamily member 4 precursor.	P43489		CREB1 repressor				
TNFL4_HUMAN	Tumor necrosis factor ligand superfamily member 4.	P23510		CREB1 repressor				
TRIB1_HUMAN	Tribbles homolog 1.	Q96RU8		CREB1 repressor				
TWST1_HUMAN	Twist-related protein 1.	Q15672		CREB1 repressor				
WFS1_HUMAN	Wolframin.	O76024		CREB1 repressor			Yes	
WWP2_HUMAN	NEDD4-like E3 ubiquitin-protein ligase WWP2.	O00308		CREB1 repressor			Yes	
XCL1_HUMAN	Lymphotactin precursor.	P47992		CREB1 repressor				
ARRB1_HUMAN	Beta-arrestin-1.	P49407			NFkB1 repressor		Yes	
ARRB2_HUMAN	Beta-arrestin-2.	P32121			NFkB1 repressor			
ATF6A_HUMAN	Cyclic AMP-dependent transcription factor ATF-6 alpha.	P18850	CREB1 activator					
ATR_HUMAN	Serine&#47;threonine-protein kinase ATR.	Q13535	CREB1 activator					

KCC1D_HUMAN	Calcium&#47; calmodulin-dependent protein kinase type 1D.	Q8IU85	CREB1 activator				Yes	PMID:16324104
CD2A1_HUMAN	Cyclin-dependent kinase inhibitor 2A, isoforms 1&#47;2&#47; 3.	P42771			NFkB1 repressor			
CHD3_HUMAN	Chromodomain-helicase-DNA-binding protein 3.	Q12873	CREB1 activator			Yes		
CHD8_HUMAN	Chromodomain-helicase-DNA-binding protein 8.	Q9HCK8	CREB1 activator			Yes		
CLOCK_HUMAN	Circadian locomoter output cycles protein kaput.	O15516		CREB1_chromatin modifier				
COMD1_HUMAN	COMM domain-containing protein 1.	Q8N668			NFkB1 repressor		Yes	PMID:15799966, 16573520, 20068069
COMD7_HUMAN	COMM domain-containing protein 7.	Q86VX2			NFkB1 repressor		Yes	PMID:15799966, 16573520, 20068069
CBP_HUMAN	CREB-binding protein.	Q92793	CREB1 activator	CREB1_chromatin modifier		Yes	Yes	PMID:11094091, 9413984
CREM_HUMAN	cAMP-responsive element modulator.	Q03060	CREB1 activator			Yes	Yes	PMID:1370576, 7961842
CRTC1_HUMAN	CREB-regulated transcription coactivator 1.	Q6UUV9	CREB1 activator				Yes	PMID:17565599
CRTC2_HUMAN	CREB-regulated transcription coactivator 2.	Q53ET0	CREB1 activator				Yes	PMID:17565599
CRTC3_H	CREB-	Q6UUV7	CREB1				Yes	PMID:17565599

UMAN	regulated transcription coactivator 3.		activator							
CTNB1_HUMAN	Catenin beta-1.	P35222			NFkB1 repressor	TCF repressor_NFkB1/CREB1 activator	Yes	Yes		PMID:12398896, 14991743, 20122174
CYLD_HUMAN	Ubiquitin carboxyl-terminal hydrolase CYLD.	Q9NQC7			NFkB1 repressor			Yes	Yes	PMID:22435550, 21119682, 19373246
DAXX_HUMAN	Death domain-associated protein 6.	Q9UER7				TCF repressor_NFkB1/CREB1 activator		Yes	Yes	PMID:16569639
DNJC2_HUMAN	DnaJ homolog subfamily C member 2.	Q99543			NFkB1 repressor			Yes		
EDF1_HUMAN	Endothelial differentiation-related factor 1.	O60869		CREB1_chromatin modifier				Yes		
EP300_HUMAN	Histone acetyltransferase p300.	Q09472		CREB1_chromatin modifier				Yes	Yes	PMID:17565599
EPHA5_HUMAN	Ephrin type-A receptor 5 precursor.	P54756	CREB1 activator							
GFI1_HUMAN	Zinc finger protein Gfi-1.	Q99684			NFkB1 repressor			Yes		
TF2AA_HUMAN	Transcription initiation factor IIA subunit 1.	P52655	CREB1 activator							
HDAC6_HUMAN	Histone deacetylase 6.	Q9UBN7		CREB1_chromatin modifier	NFkB1 repressor					
HDAC9_HUMAN	Histone deacetylase 9.	Q9UKV0		CREB1_chromatin modifier						
HIPK3_HUMAN	Homeodomain-interacting protein kinase 3.	Q9H422	CREB1 activator							
IRAK1_HUMAN	Interleukin-1 receptor-associated kinase 1.	P51617			NFkB1 repressor			Yes	Yes	PMID:20300215, 17457343
IRAK3_H	Interleukin-1	Q9Y616			NFkB1			Yes	Yes	PMID:20300215, 17457343

UMAN	receptor-associated kinase 3.			repressor					
ITCH_HUMAN	E3 ubiquitin-protein ligase Itchy homolog.	Q96J02		NFkB1 repressor			Yes	Yes	PMID:22435550, 21119682
KAT2A_HUMAN	Histone acetyltransferase KAT2A.	Q92830		CREB1_chromatin modifier			Yes		
KAT2B_HUMAN	Histone acetyltransferase KAT2B.	Q92831		CREB1_chromatin modifier					
KAT7_HUMAN	Histone acetyltransferase KAT7.	O95251		CREB1_chromatin modifier			Yes		
MK03_HUMAN	Mitogen-activated protein kinase 3.	P27361	CREB1 activator			Yes	Yes		PMID:8688081
MED15_HUMAN	Mediator of RNA polymerase II transcription subunit 15.	Q96RN5	CREB1 activator				Yes		
MEIS1_HUMAN	Homeobox protein Meis1.	O00470	CREB1 activator						
MTA2_HUMAN	Metastasis-associated protein MTA2.	O94776	CREB1 activator	CREB1_chromatin modifier			Yes		
IKBA_HUMAN	NF-kappa-B inhibitor alpha.	P25963		NFkB1 repressor		Yes	Yes		PMID:12740336
PARP1_HUMAN	Poly [ADP-ribose] polymerase 1.	P09874			TCF repressor_NFkB1/CREB1 activator		Yes	Yes	PMID:17504138, 19060926
PBX1_HUMAN	Pre-B-cell leukemia transcription factor 1.	P40424	CREB1 activator						
PIAS1_HUMAN	E3 SUMO-protein ligase PIAS1.	O75925	CREB1 activator				Yes		
PIAS4_HUMAN	E3 SUMO-protein ligase PIAS4.	Q8N2W9		NFkB1 repressor			Yes		
POGZ_HU	Pogo	Q7Z3K3	CREB1				Yes		

MAN	transposable element with ZNF domain.		activator					
AAPK1_HUMAN	5'-AMP-activated protein kinase catalytic subunit alpha-1.	Q13131	CREB1 activator			Yes		PMID:19442239, PMID:18063805, PMID:17565599
AAPK2_HUMAN	5'-AMP-activated protein kinase catalytic subunit alpha-2.	P54646	CREB1 activator			Yes		PMID:19442239, PMID:18063805
KPCD1_HUMAN	Serine&#47;threonine-protein kinase D1.	Q15139	CREB1 activator		Yes	Yes		PMID:20497126, 17389598
KPCD2_HUMAN	Serine&#47;threonine-protein kinase D2.	Q9BZL6	CREB1 activator		Yes	Yes		PMID:20497126, 17389598
RBCC1_HUMAN	RB1-inducible coiled-coil protein 1.	Q8TDY2	CREB1 activator			Yes		
HOIL1_HUMAN	RanBP-type and C3HC4-type zinc finger-containing protein 1.	Q9BYM8		NFkB1 repressor		Yes		
BRE1A_HUMAN	E3 ubiquitin-protein ligase BRE1A.	Q5VTR2		NFkB1 repressor		Yes		
KS6A1_HUMAN	Ribosomal protein S6 kinase alpha-1.	Q15418	CREB1 activator		Yes	Yes		PMID:8688081, 17565599
KS6A4_HUMAN	Ribosomal protein S6 kinase alpha-4.	O75676	CREB1 activator			Yes	Yes	PMID:16125054
KS6A5_HUMAN	Ribosomal protein S6 kinase alpha-5.	O75582	CREB1 activator				Yes	PMID:16125054

RUNX3_HUMAN	Runt-related transcription factor 3.	Q13761				TCF repressor_NFkB1/CREB1 activator	Yes	Yes	PMID:18772112, 21523770
SMAD3_HUMAN	Mothers against decapentaplegic homolog 3.	P84022			NFkB1 repressor				
SRCAP_HUMAN	Helicase SRCAP.	Q6ZRS2		CREB1_chromatin modifier			Yes		
SRBP2_HUMAN	Sterol regulatory element-binding protein 2.	Q12772	CREB1 activator				Yes		
SSBP3_HUMAN	Single-stranded DNA-binding protein 3.	Q9BWW4	CREB1 activator						
SUPT3_HUMAN	Transcription initiation protein SPT3 homolog.	O75486		CREB1_chromatin modifier					
TADA1_HUMAN	Transcriptional adapter 1.	Q96BN2		CREB1_chromatin modifier					
TADA3_HUMAN	Transcriptional adapter 3.	O75528		CREB1_chromatin modifier			Yes		
TAF1_HUMAN	Transcription initiation factor TFIID subunit 1.	P21675		CREB1_chromatin modifier			Yes		
TAF10_HUMAN	Transcription initiation factor TFIID subunit 10.	Q12962		CREB1_chromatin modifier			Yes		
TAF1L_HUMAN	Transcription initiation factor TFIID subunit 1-like.	Q8IZX4		CREB1_chromatin modifier					
TAF5_HUMAN	Transcription initiation factor TFIID subunit 5.	Q15542		CREB1_chromatin modifier					
TAF5L_HUMAN	TAF5-like RNA polymerase II p300&#47;CBP-associated factor-	O75529		CREB1_chromatin modifier			Yes		

	associated factor 65 kDa subunit 5L.						
TAF9_HUMAN	Transcription initiation factor TFIID subunit 9.	Q16594		CREB1_chromatin modifier	Yes		
TAXB1_HUMAN	Tax1-binding protein 1.	Q86VP1		NFkB1 repressor	Yes	Yes	PMID:22435550
UB2V1_HUMAN	Ubiquitin-conjugating enzyme E2 variant 1.	Q13404		NFkB1 repressor			
TNAP3_HUMAN	Tumor necrosis factor alpha-induced protein 3.	P21580		NFkB1 repressor		Yes	PMID:19494296, 19608751, 16684768, 19380639
TOP2A_HUMAN	DNA topoisomerase 2-alpha.	P11388		NFkB1 repressor			
RO52_HUMAN	E3 ubiquitin-protein ligase TRIM21.	P19474		NFkB1 repressor	Yes		
TSSK4_HUMAN	Testis-specific serine&#47;threonine-protein kinase 4.	Q6SA08	CREB1 activator				
UBC9_HUMAN	SUMO-conjugating enzyme UBC9.	P63279	CREB1 activator				
UB2L3_HUMAN	Ubiquitin-conjugating enzyme E2 L3.	P68036		NFkB1 repressor	Yes		
UB2V1_HUMAN	Ubiquitin-conjugating enzyme E2 variant 1.	Q13404		NFkB1 repressor	Yes		
UIMC1_HUMAN	BRCA1-A complex subunit RAP80.	Q96RL1		NFkB1 repressor	Yes		
UBP16_HUMAN	Ubiquitin carboxyl-terminal hydrolase 16.	Q9Y5T5		NFkB1 repressor	Yes		
UBP22_H	Ubiquitin	Q9UPT9		CREB1_chromatin modifier			

UMAN	carboxyl-terminal hydrolase 22.							
VEGFA_HUMAN	Vascular endothelial growth factor A precursor.	P15692	CREB1 activator			Yes		
VPS36_HUMAN	Vacuolar protein-sorting-associated protein 36.	Q86VN1		NFkB1 repressor		Yes		
XRCC5_HUMAN	X-ray repair cross-complementing protein 5.	P13010			TCF repressor_NFkB1/CREB1 activator	Yes	Yes	PMID:17283121, 17031478
XRCC6_HUMAN	X-ray repair cross-complementing protein 6.	P12956			TCF repressor_NFkB1/CREB1 activator	Yes	Yes	PMID:17283121, 17031478
ZHX1_HUMAN	Zinc fingers and homeoboxes protein 1.	Q9UKY1	CREB1 activator			Yes		
ZMYM2_HUMAN	Zinc finger MYM-type protein 2.	Q9UBW7	CREB1 activator			Yes		
ZN451_HUMAN	Zinc finger protein 451.	Q9Y4E5	CREB1 activator					
ZN675_HUMAN	Zinc finger protein 675.	Q8TD23		NFkB1 repressor			Yes	PMID:11751921

Chapter 6

Paper V

Vercruysse *et al. BMC Bioinformatics* 2012, **13**:116
http://www.biomedcentral.com/1471-2105/13/116

BMC Bioinformatics

SOFTWARE **Open Access**

OLSVis: an animated, interactive visual browser for bio-ontologies

Steven Vercruysse*, Aravind Venkatesan and Martin Kuiper

Abstract

Background: More than one million terms from biomedical ontologies and controlled vocabularies are available through the Ontology Lookup Service (OLS). Although OLS provides ample possibility for querying and browsing terms, the visualization of parts of the ontology graphs is rather limited and inflexible.

Results: We created the OLSVis web application, a visualiser for browsing all ontologies available in the OLS database. OLSVis shows customisable subgraphs of the OLS ontologies. Subgraphs are animated via a real-time force-based layout algorithm which is fully interactive: each time the user makes a change, *e.g.* browsing to a new term, hiding, adding, or dragging terms, the algorithm performs smooth and only essential reorganisations of the graph. This assures an optimal viewing experience, because subsequent screen layouts are not grossly altered, and users can easily navigate through the graph. URL: http://ols.wordvis.com

Conclusions: The OLSVis web application provides a user-friendly tool to visualise ontologies from the OLS repository. It broadens the possibilities to investigate and select ontology subgraphs through a smooth visualisation method.

Keywords: Bio-ontologies, Visualisation, Browsing, Web application

Background

Ontologies constitute an increasingly important knowledge resource. In the biomedical domain the engineering of ontologies is predominantly organised by the Open Biomedical Ontology (OBO) Foundry [1]. Ontologies arrange terms hierarchically, connected by relationships in directed acyclic graphs. OBO ontologies represent formalised biological knowledge and are broadly used in the analysis and interpretation of experimental results, *e.g.* by linking Gene Ontology (GO) terms [2] to gene sets [3,4]. Ontologies provide also an important resource to find accurate terms for use in scientific reports.

Many tools are available for browsing ontologies (see [5,6]). Several of them are integrated in systems dedicated to analyse specific data sets (*e.g.* calculating overrepresented GO categories in a gene list: GOrilla [7], agriGO [8], and GOTermFinder [3]). Other tools are designed for more general-purpose ontology exploration, such as QuickGO [9], AmiGO [10], or NCBO's FlexViz [11]. Some of these ontology viewers are text-based, *i.e.* they use a folder/subfolder-interface to explore hierarchies (*e.g.* AmiGO [10], MGI GO Browser [12]). However, many ontologies feature multiple-inheritance: they have terms that are linked to more than one parent. This multiple-inheritance is more clearly visualised in a two-dimensional display, with nodes and connectors in between. For instance, the Ontology Lookup Service (OLS) offers static images that clarify better how terms are positioned and related to adjacent terms in the hierarchy, and it provides this unified interface for the browsing of 79 bio-ontologies [13]. Also, the NCBO BioPortal features the graph browser FlexViz, which draws subgraphs from 293 ontologies and allows clicking on terms to bring up its local environment (*e.g.* child or parent terms) [11]. FlexViz is one of the most powerful viewers currently available. But despite the added flexibility and user-interaction support, this graphical browser may feel rigid and sometimes confusing, because it only shifts between static, pre-calculated, and often sub-optimal configurations. The addition of new terms may therefore result in large graph reorganisations that are often hard to follow.

* Correspondence: vercruys@nt.ntnu.no
Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

One can easily experience why we consider the NCBO Bioportal's FlexViz not an optimal viewer even in simple use scenarios, by trying for example the following exercise in FlexViz: open the ontology 'Gene Ontology', search for 'mitochondrion', and then expand some terms upward towards the root, *e.g.* 'intracellular membrane-bound organelle' and then 'intracellular organelle'. When doing so, one is confronted with terms moving all over and far out of the viewport, with the viewport shifting over large distances. This is caused by many terms being placed next to each other on a too wide hierarchy level. Much of the overview is lost, and an attempt to regain some of it back by zooming out will leave the node labels too small to read. Using other layout algorithms than the default one ('tree layout') seems also less than satisfactory.

Although FlexViz constitutes an interesting first step towards a fully flexible and user-friendly browsing experience, it leaves room to explore alternative approaches to ontology visualisation. We therefore investigated if the use of a fundamentally different layout method would give a better user-experience for the general-purpose browsing of ontologies. We chose to implement a *real-time, force-based* layout algorithm, which can organise nodes and connections globally and dynamically. First, it uses a 'minimum energy' principle, ensuring that nodes and connection-structures are distributed optimally relative to each other in the available screen space. Second, it immediately responds when (and as long as) the user interacts with the graph, updating the nodes' positions continuously until a new optimal configuration is reached.

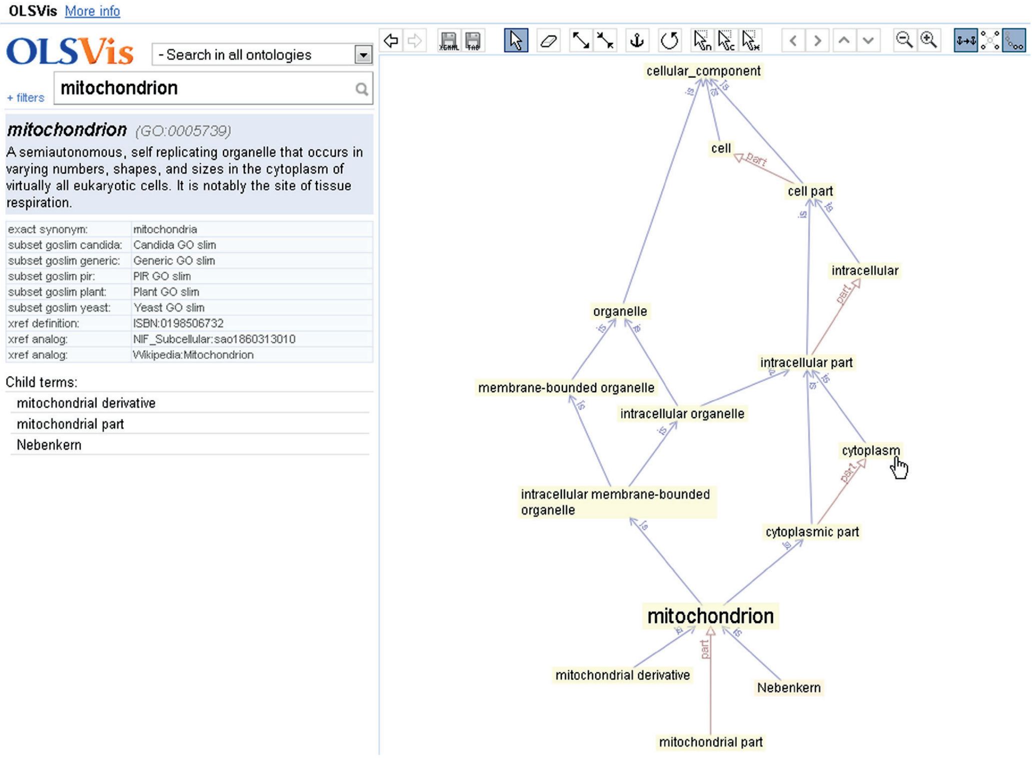
Results and discussion

Web-based ontology visualisers are largely used for browsing and to analyse the placement of a given term in an ontology. They help to get a grasp of the local environment of a term of interest or to view terms that form the connection to the root term (path to the root). As bio-ontologies are getting increasingly complex, browsing through them requires a visualiser that offers more intuitive functionalities such as the autosuggestion of terms, an 'undo' function, filtering for relationships and additional functions that facilitate smooth and user-friendly browsing. The visualisers that are currently available only have some of these characteristics, and often show limitations with respect to browsing speed, scalability issues, context-based display of a term's environment, or overall user interaction support. This prompted us to create the web application *OLSVis*: a fast, interactive visualiser to explore OLS ontologies based on minimal and smooth relayouts. OLSVis exploits the speed and ease-of-use of the WordVis application [14,15]. Inspired by the Ontology Lookup Service, we applied the concept of a term's local environment (child terms and path to the root [13]) as the basic viewing unit for the visualiser. We illustrate the advantages of OLSVis through three use cases, exemplifying both the added functionalities and the enhanced user-experience that OLSVis brings to ontology visualisation. Use Case I demonstrates a general overview of the features of OLSVis, highlighting its interactive environment using the Gene Ontology. Use Case II illustrates an approach to view common ancestor terms shared between two Gene Ontology terms; and Use Case III demonstrates the visualisation of the local neighbourhood of a protein.

Use case I: Browsing ontologies in OLSVis

The first use case illustrates how OLSVis can make ontology browsing more intuitive: A user is interested in the placement of the term 'mitochondrion' in the Gene Ontology hierarchy. She can proceed in two ways: a) select the ontology of choice and then search for the chosen term, or b) do a direct search for the GO term 'mitochondrion'. Autosuggestion enables her to perform a quick selection of the term from the autosuggest list. Autosuggestion also highlights the occurrence of the chosen term in other ontologies. The user selects the GO entry, in this case 'mitochondrion' (GO:0005739) and OLSVis shows the official GO term centered in the visualiser, along with its child terms and all paths of ancestor terms up to a GO root term (Figure 1). The display of the local environment of the term is dynamic and the visualiser allows the use of various features to further refine the display (see the toolbar). For instance, the 'Eraser' tool can be used to hide unnecessary terms from the display panel. In some cases the relation names are abbreviated for a clearer view and displayed in full by mouse-hovering. Parts of the graph can be made less/more compact by increasing/decreasing the length of connectors. Also, similar to modern map-applications, OLSVis supports moving the graph by dragging its background and zooming by mouse scrolling. Furthermore, a 'filters' panel is provided to assist the user in narrowing or broadening the search space. In Figure 1 both 'is_a' and 'part_of' relations are shown.

Other improvements that OLSVis provides concern the animation and presentation of terms after specific user actions. For instance, clicking on 'cytoplasm' will shift the display into cytoplasm's local environment (the children and all ancestors of the term 'cytoplasm'). The algorithm switches between local environments by gently pushing out terms and inserting new terms, which allows a user to easily keep track of the changing display. A button on the toolbar may be used to prevent automatic removing of nodes. Its dynamic layout algorithm and the additional graph interaction tools all contribute to the user-friendliness of OLSVis. Furthermore, OLSVis allows



OLSVis More info

- Search in all ontologies

+ filters

mitochondrion

mitochondrion *(GO:0005739)*

A semiautonomous, self replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration.

exact synonym:	mitochondria
subset goslim candida:	Candida GO slim
subset goslim generic:	Generic GO slim
subset goslim pir:	PIR GO slim
subset goslim plant:	Plant GO slim
subset goslim yeast:	Yeast GO slim
xref definition:	ISBN:0198506732
xref analog:	NIF_Subcellular:sao1860313010
xref analog:	Wikipedia:Mitochondrion

Child terms:

- mitochondrial derivative
- mitochondrial part
- Nebenkern

Figure 1 Screenshot of the OLSVis web application. OLSVis has been used to search the Gene Ontology term ‘mitochondrion’ (see search box). The graph panel on the right shows the term’s child, parent and ancestor terms linked by various relationship types. Clicking any term launches a new search focused on that term. The left panel shows ontology details for the displayed term.

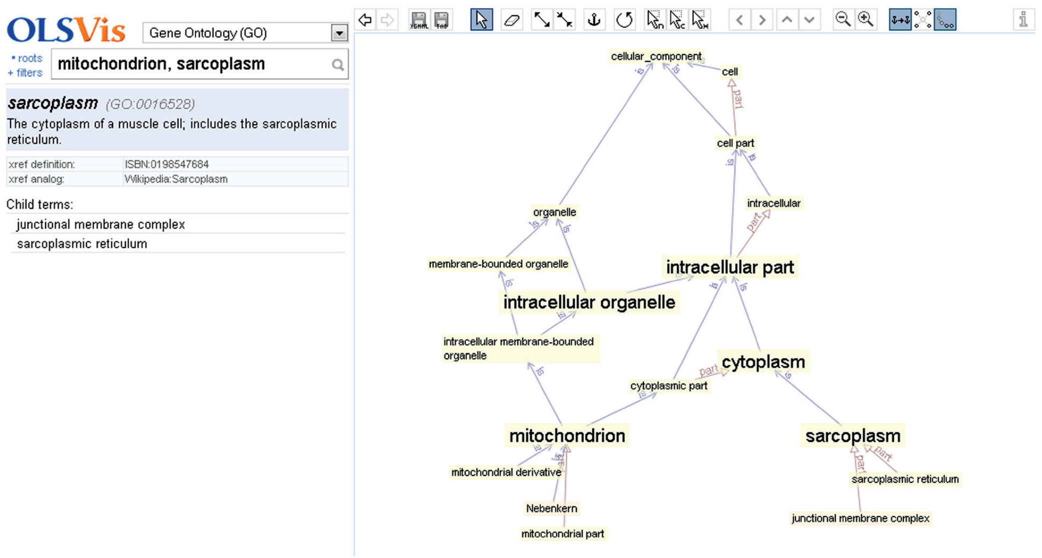
the user to save the local environment in XGMML format that may be imported in network building tools such as Cytoscape [16,17]. Alternately, the user can obtain the list of nodes and relationships in the current view in a tab-delimited file.

Use case II: Identifying shared ancestor terms between two ontology terms

Suppose a user wants to identify the common ancestry between two different terms, in order to assess their relatedness. Use case II shows an example based on the cellular components ‘mitochondria’ and ‘sarcoplasm’. Here the user first selects ‘Gene Ontology’ from the ontology list and then enters two terms in the search box, separated by a comma. OLSVis reads both text strings as separate terms, matches them to their respective terms in the selected ontology, GO, and then displays a merged view of their local environments. Figure 2 shows the terms that hereby are displayed, linking ‘mitochondria’ and ‘sarcoplasm’ and showing their shared connections. Additionally, for customised visualisation, shared terms could be repositioned and fixed by using the ‘Anchor node’ functionality. Non-anchored terms will slide to new optimal positions. This example demonstrates the potential of OLSVis in displaying environments for multiple terms which is currently not available in any other visualiser.

Use case III: Visualising the local neighbourhood of a protein

Biologists are often interested in understanding the various attributes of a particular protein such as protein modifications, biological functions, or protein interactions. Use case III illustrates how OLSVis can be used for visualising the local neighbourhood of a protein. In this example the protein is cdc23 (*H. sapiens*). The user enters the string ‘cdc23’ and the autosuggestion list shows a number of matches from the Cell Cycle Ontology (CCO) [18]. Selection of the term ‘cdc23_HUMAN (CCO:B0002212)’ displays the local neighbourhood of



OLSVis

Gene Ontology (GO)

• roots
+ filters

mitochondrion, sarcoplasm

sarcoplasm (GO:0016528)

The cytoplasm of a muscle cell; includes the sarcoplasmic reticulum.

xref definition:	ISBN:0198547684
xref analog:	Wikipedia:Sarcoplasm

Child terms:

- junctional membrane complex
- sarcoplasmic reticulum

Figure 2 OLSVis screenshot of use case II. The canvas shows the combined local environments of two search terms, their paths to the root and thereby the relatedness between the terms. The search box in the left panel shows the two terms. Also a number of terms were 'anchored' by the user.

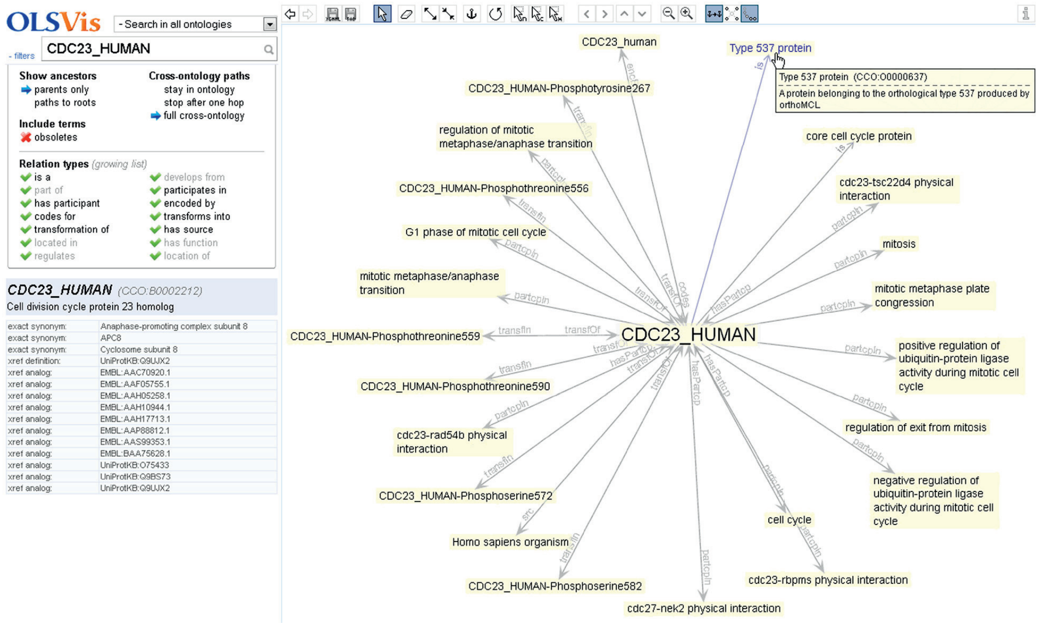
this term whilst providing a warning message that alerts the user as to the large number of terms associated with the chosen protein. When browsing large ontologies (*e.g.* CCO), a user usually has to deal with performance issues as the visualiser may actually fail to load the subgraph due to its size. Instead, OLSVis loads up to 500 terms smoothly and if more it gives a notification to the user. The user is suggested to use the filter panel to narrow the search space for improved performance and viewing. For example, clicking on 'parents only' will update the current view with a simplified graph (Figure 3). Alternatively, a number of relation types could be filtered away. Here we note that CCO includes bidirectional relationships, so leaving some out can clarify the intended parent–child hierarchy. The user may then choose to save the current display in formats provided by OLSVis. For instance, biologists to a large extent still work on spreadsheets where they periodically associate a particular protein of interest with an ontological term. In such cases, saving the current view in a tab-delimited format makes it easier for them to use the terms associated with a protein in their annotation work.

Implementation

For the client-side of the software, we used the modern web technologies of JavaScript and the new HTML5 standard. In contrast to traditional Flash-objects or Java-applets, which are isolated objects in the web page, JavaScript and HTML5 make it possible to create animations that are fully connectable with other elements on the web page, and that require no extra browser plugins. HTML5 defines the < canvas > HTML-element, basically a rectangular empty space on the web page, onto which JavaScript code (which is included in the web page) draws basic shapes like lines, circles, text, etc. Note that the older SVG (Scalable Vector Graphics) technology requires computationally expensive (slow) DOM-updates; therefore only canvas is appropriate for smooth animation of large graphs. Because a sufficiently powerful JavaScript library for animated graph browsing did not exist yet, we wrote one from scratch: GraphVis. We first applied GraphVis in the webapp WordVis [14,15], which visualises WordNet, a lexical database of English [19].

Layout engine

We applied our GraphVis layout module to the exploration of ontologies in OLSVis, and upgraded it among others with hierarchical layout for parent/ancestor terms, see Figure 1. When the user searches for an ontology term, OLSVis will by default show it together with its child terms and parent terms up to the ontology root(s), see Figure 1. After initial placement of ontology terms, OLSVis uses a real-time force-based layout algorithm that gently moves the terms towards more optimal positions. The algorithm is explained in [14] and [15]. It



OLSVis

- Search in all ontologies

CDC23_HUMAN

- filters

Show ancestors
➡ parents only
paths to roots

Include terms
✗ obsoletes

Cross-ontology paths
stay in ontology
stop after one hop
➡ full cross-ontology

Relation types *(growing list)*
✓ is a
✓ part of
✓ has participant
✓ codes for
✓ transformation of
✓ located in
✓ regulates
✓ develops from
✓ participates in
✓ encoded by
✓ transforms into
✓ has source
✓ has function
✓ location of

CDC23_HUMAN *(CCO:B0002212)*
Cell division cycle protein 23 homolog

exact synonym:	Anaphase-promoting complex subunit 8
exact synonym:	APC8
exact synonym:	Cyclosome subunit 8
xref definition:	UniProtKB:Q9UJX2
xref analog:	EMBL:AAC70920.1
xref analog:	EMBL:AAF05755.1
xref analog:	EMBL:AAH05258.1
xref analog:	EMBL:AAH10944.1
xref analog:	EMBL:AAH17713.1
xref analog:	EMBL:AAP88812.1
xref analog:	EMBL:AAS99353.1
xref analog:	EMBL:BAA75628.1
xref analog:	UniProtKB:O75433
xref analog:	UniProtKB:Q9BS73
xref analog:	UniProtKB:Q9UJX2

Type 537 protein (CCO:00000637)
A protein belonging to the orthological type 537 produced by orthoMCL

CDC23_HUMAN
CDC23_human
Type 537 protein
CDC23_HUMAN-Phosphotyrosine267
regulation of mitotic metaphase/anaphase transition
core cell cycle protein
cdc23-tsc22d4 physical interaction
CDC23_HUMAN-Phosphothreonine556
mitosis
G1 phase of mitotic cell cycle
mitotic metaphase/anaphase transition
mitotic metaphase plate congression
CDC23_HUMAN-Phosphothreonine559
positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle
CDC23_HUMAN-Phosphothreonine590
regulation of exit from mitosis
cdc23-rad54b physical interaction
negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle
CDC23_HUMAN-Phosphoserine572
cell cycle
Homo sapiens organism
cdc23-rbpms physical interaction
CDC23_HUMAN-Phosphoserine582
cdc27-nek2 physical interaction
is
partOf
partcpIn
transfIn
transfOf
codes
hasPartcp
src

Figure 3 Screenshot illustrating use case III. The canvas shows the local neighbourhood of a CCO term. The left panel shows the term name in the search box and the pop-out panel with filter settings. The filter setting enables the display of only parent terms.

models nodes as repelling, electrically charged rectangles. This distributes them over the screen, prevents them from occupying the same space (if possible), and prevents term labels from overlapping. Connections are modelled as mechanical springs, which hold nodes together and which may be given a specific preferred-length in order to create a certain global structure in the graph. Connections may also have a preferred orientation (*e.g.* down-to-up for ‘is a’ links). This layout is fully interactive: each time the user makes a change (such as focusing on a new term, hiding, adding or dragging terms, changing connection lengths), it smoothly yet minimally reorganises the graph. This assures an optimal viewing experience that minimises each operation’s effect on graph reorganisation, and maximises the user’s ability to keep track of changes and comprehend the new layout.

Data source

OLSVis visualises the contents of the OLS database [13,20,21], which holds around 80 bio-ontologies and over 1 million concepts. OLSVis uses a local copy of OLS‘ publicly available database, in order to provide a smooth visualisation with fast response times. Only via a local copy placed on OLSVis’ server can the node environments be calculated sufficiently fast. The use of the OLS web-service to retrieve data proved to be painfully slow, because each mouse click required several web-service queries, which typically resulted in total query times of several tens of seconds. EBI updates the OLS database weekly by polling its ontology providers through the CVS and SVN version control systems. OLSVis detects OLS’ updates automatically and then updates its local copy. In addition, a number of table-indexes and pre-calculated fields are added to enable the speed of OLSVis. On the server-side of OLSVis, PHP scripts translate client-side requests into custom queries on the local MySQL database. Note that the web-application’s front-end is designed independent from the database back-end. Given software that would be able to calculate node-environments (filterable paths-to-roots) in reasonably short times, the visualiser would be usable also for other semantic resources.

Term searching

While the user types one (or several) terms or identifiers (*e.g.* ‘mito’ or ‘PO:0009001’) in the search box, a selection of best known matches is shown in a pop-out list. This includes preferred terms as well as their synonyms. For each autosuggested term, the ontology’s (short)name and identifier is shown, and mouse-hovering shows its ontology’s full name. Autosuggestions can be confined to a single specific ontology by selecting one from the dropdown list. Pressing 'Enter' in the search box will display the term that is selected in the autosuggestion list. If the user has *no term selected*, OLSVis will take the *first* term (also if the autosuggestion list has not appeared yet).

Basic visualisation

The chosen term is then *expanded*: it is placed in the centre of the graph panel, amidst its *local environment* of child terms and parent terms, and connected with further ancestors up to the ontology root(s). This configuration is inspired by OLS' static images [13]. Child terms are ordered in a half circle under the expanded term; ancestors are put in hierarchical levels above it. Relations are shown as labelled arrows; their lengths are adjusted for good hierarchical positioning. After initial placement, the visualiser slides terms to more optimal positions via real-time animation; hereby the graph 'feels' and behaves as if terms are repelling electric charges (or repelling magnets) that are connected over mechanic/elastic springs. This creates a layout that minimises term overlap. In addition, the connecting arrows undergo a small north–south orienting force to enhance a hierarchical alignment of terms. The visible graph is fully customisable: see the toolbar in Figure 1 or the online description for mouse/keyboard shortcuts. It has undo/redo history, and terms can be dragged and pushed around. Clicking on any displayed term will re-centre on that term and expand its local neighbourhood. Hereby the graph is subtly reorganised via real-time animation, and is transformed into the new term's local environment (by addition and removal of terms). This enables easy and intuitive browsing through ontologies. The automatic removal of already visible terms can be turned off via the rightmost button on the toolbar. Hovering over any term makes its definition pop up. For a relation arrow, its non-abbreviated name pops up. Leaf terms (=without child terms) get a slightly orange background. The three most common relations (is_a, part_of, develops_from) get a coloured arrow. In the left panel, data for the *last* expanded term is shown: its identifier (hovering shows ontology's full name) and definition; its synonyms, annotations and cross-references (as in the OLS database); and its child terms (each clickable to expand), to make them easier visible when there are many. When zooming in, OLSVis increases distances faster than font sizes; this is more useful and is an extra method (next to electrostatic repulsion) to counteract overlapping terms.

Customised visualisation

A click on 'filters' (left of the search box) brings up a panel to set filters that prune the expanded node's environment. For instance, any relation type can be excluded; this means that they are omitted when building *e.g.* the path-to-root. Initially the three most common relation types are listed in the panel; this list grows each time the visualiser encounters new types. Relation types that are currently in the visualiser are highlighted. The filter that hides obsolete terms also hides them in autosuggestion lists. Earlier expanded terms and their environment are by default automatically removed when clicking on a new term, but can be kept in the visualiser by turning off the rightmost toolbar button. Several toolbar tools enable further customisation of the graph. Connections can be made longer or shorter (also via Alt + scrolling up/down). Terms can be anchored to a fixed position, and anything can be removed manually via the Eraser tool.

More features

A 'roots' link appears next to the search box after selecting a specific ontology. Clicking it shows and expands this ontology's root term(s) (if defined in the OLS data), *i. e.* showing them and their child terms. This enables easy top-down ontology exploration. Multiple terms and identifiers can be searched, separated by commas. Therefore in-term commas must be preceded by a backslash, and genuine backslashes doubled. First hits from autosuggestion are then expanded. When a term's local environment contains too many terms (this happens with application ontologies such as the Cell Cycle Ontology [18]), OLSVis will only show the first 500 terms and will suggest using filters. OLSVis supports URL-shortcuts:

(1) A term or identifier can be expanded directly via URLs like: *ols.wordvis.com/q = GO:0005739*, or *. . ./q = mito*. The part after */q* = will be put in the search box and the first term that would have been autosuggested will be expanded.
(2) A specific ontology can be preselected via a URL like: *ols.wordvis.com/ont = GO*. The part after */ont* = is the ontology's short name from the selection list. This is a shortcut for users that are mainly interested in a specific ontology.
(3) 'q' and 'ont' can be combined like: *. . ./ont = GO&q = mitochondrion,sarcoplasm* , which also illustrates a multi-term query.
(4) Some ontologies use non-standard prefixes in term-identifiers (GO has 'GO:', but ZFA may use 'ZFS:', and NEWT has none), so identifiers may be disambiguated by adding their ontology's short name as prefix, e.g. *. . ./q = NEWT:1234*, or *. . ./q = ZFA: ZFS:0000019.*

Terms in the graph can be right-clicked for more options. The visible graph can be exported to an XGMML file (eXtensible Graph Markup and Modeling Language) and can subsequently be imported in Cytoscape [16,17] for further analysis. There, node labels will show the term

Table 1 Comparison of OLSVis and other two-dimensional ontology visualisers

	OLSVis	OLS	FlexViz	OntoViz	IsaViz	GOSurfer	GOMiner	OntoTrack	OBO-Edit	QuickGO	AmiGO
Full Dynamic layout method	✓		✓(semi) [1]	✓(semi) [1]				✓(semi) [1]	✓(semi) [1]		
Layout user-interaction	Interactive; + continuous optimisation[2]		one-by-one dragging	one-by-one dragging	highlighting a selected branch			one-by-one dragging	one-by-one dragging		
Hierarchy layout	ancestors: layered; children: circular	ancestors: layered; OR: parents + children	several[3]	ancestors + children: layered	graph-view or radar-view	tree view	ancestors + children: layered	ancestors + children: layered	ancestors + children: layered	ancestors	subfolders
Term searching	✓	✓	✓	✓	✓		✓	✓		✓	✓
Multiple term search	✓										
Auto-suggest	✓	✓	✓							✓	
Filters	✓		✓					✓			
Undo	✓	✓[4]	✓	✓	✓			✓		✓[4]	✓[4]
Context dependent display[5]	✓										
Click on term expands it	✓	page reload	✓	✓				✓	✓		✓
Path to root	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Simple tree view		✓	✓	✓		✓	✓	✓	✓		✓
Handling 500 terms	✓						✓				
Term source	OLS (79 ontologies)	OBO (79 ontologies)	NCBO (293 ontologies)	OWL ontologies	RDF graph visualiser	GO (1)	GO (1)	OWL ontologies	OBO ontologies	GO (1)	GO (1)
Web tool (Technology)	✓ (Javascript)	✓ (images)	✓ (Flash)	–(Protégé)	–(Java)	–(.exe)	–(Java)	–(Java)	–(Java)	✓ (HTML)	✓ (HTML)

[1]: Only shifting between static configurations. [2]: Also during the several editing functions (which are displayed on OLSVis' toolbar). [3]: Each with their own shortcomings, see 'Background'. [4]: Browser's back button. [5]: The visualiser can optionally clean up nodes that are not in the latest expanded node's "environment".

Comparison with: OLS [13], FlexViz [11], OntoViz [22], IsaViz [23], GOSurfer [24], GOMiner [25], OntoTrack [26], OBO-Edit [27], QuickGO [9], AmiGO [10]).

names, and 'ontID' attributes store the ontology identifiers. In addition, nodes and relations can be exported to a tab-delimited text file.

Comparison with other visualisation tools

The utility and performance of OLSVis was assessed in comparison to other tools commonly used for ontology visualisation, including some biological data analysis tools that have visualisation components integrated in them, as listed in Table 1. The evaluation addressed a number of criteria, including tool functionality (*e.g.* support of multiple term searching); scalability (*e.g.* handling of large numbers of terms); and some aspects that capture user-friendliness and intuitiveness of browsing (*e.g.* context-dependent browsing). The table shows that some of OLSVis' features are not provided by any other visualiser, and that the other tools only support a subset of what OLSVis offers. Clearly, OLSVis offers the most interactive visualisation environment. FlexViz ranks well too, as it also provides a relatively high level of user-interaction; however, OLSVis makes more efficient and intuitive use of the available screen space.

Conclusions

OLSVis was created to improve the exploration of bio-ontologies. Other visualisers like FlexViz, may feel rigid and sometimes confusing, because the addition of new terms may result in largely rearranged term displays. OLSVis demonstrates that the user experience for ontology exploration can be substantially improved by using real-time animation of force-based graph relayout, and by providing improved user interaction on the graph's structure. This new webapp provides the scientific community with a versatile and more user-friendly tool to explore ontologies and to find related and more precise ontology terms.

Availability and requirements

- Project name: OLSVis
- Project home page: http://ols.wordvis.com
- Operating system: Platform independent
- Programming language: JavaScript, PHP, (MySQL)
- Other requirements: Modern browser: recent version of Firefox, Chrome, Opera, Safari or Internet Explorer. (IE 8 not recommended; please upgrade to IE 9, which supports 'canvas' and thus is much faster). No browser plugin needed.
- License: The web-application is freely accessible for use.
- Any restrictions to use by non-academics: No specific restrictions.

Abbreviations

CCO: Cell Cycle Ontology; CVS: Concurrent Versions System; EBI: European Bioinformatics Institute; GO: Gene Ontology; NCBO: National Center for Biomedical Ontology; OBO: Open Biomedical Ontology; OLS: Ontology Lookup Service; SVN: Apache Subversion; XGMML: eXtensible Graph Markup and Modeling Language.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

SV was supported by Funksjonell Genomforskning (FUGE) Midt-Norge and the Norwegian University of Science and Technology [grant number 81179100]. AV was supported by Faculty of Natural Sciences and Technology, NTNU, Norway. We thank EBI for making the OLS database available and keeping it up-to-date.

Authors' contributions

SV designed and programmed the software and drafted the manuscript. AV helped testing the software and draft the manuscript. MK helped conceive the web application and draft the manuscript. All authors read and approved the final manuscript.

Received: 9 December 2011 Accepted: 30 May 2012
Published: 30 May 2012

References

1. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Consortium OBI, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007, **25**(11):1251–1255.
2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. *Nat Genet* 2000, **25**(1):25–29.
3. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 2004, **20**(18):3710–3715.
4. Maere S, Heymans K, Kuiper M: BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 2005, **21**(16):3448–3449.
5. Katifori A, Halatsis C, Lepouras G, Vassilakis C, Giannopoulou EG: Ontology Visualization Methods - A Survey. *ACM Comput Surv* 2007, **39**(4):Article10.
6. Lanzenberger M, Sampson J, Rester M: Ontology Visualization: Tools and Techniques for Visual Representation of Semi-Structured Meta-Data. *Journal of Universal Computer Science* 2010, **16**(7):1036–1054.
7. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinforma* 2009, **10**:48.
8. Du Z, Zhou X, Ling Y, Zhang Z, Su Z: agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res* 2010, **38**(Web Server issue):W64–W70.
9. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R: QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* 2009, **25**(22):3045–3046.
10. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, AmiGO Hub, Web Presence Working Group: AmiGO: online access to ontology and annotation data. *Bioinformatics* 2009, **25**(2):288–289.
11. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA: BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009, **37** (Web Server issue):W170–W173.
12. Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE, Mouse Genome Database Group: The Mouse Genome Database genotypes::phenotypes. *Nucleic Acids Res* 2009, **37**(Database issue):D712–D719.
13. Côté R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H: The Ontology Lookup Service: bigger and better. *Nucleic Acids Res* 2010, **38** (Web Server issue):W155–W160.
14. WordVis [http://wordvis.com].
15. Vercruysse S, Kuiper M: WordVis: JavaScript and Animation to Visualize the WordNet Relational Dictionary. *In Proceedings of the Third International*

Conference on Intelligent Human Computer Interaction: 29–31 August 2011, Prague. In press.

16. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**(11):2498–2504.
17. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD: **Integration of biological networks and gene expression data using Cytoscape.** *Nat Protoc* 2007, **2** (10):2366–2382.
18. Antezana E, Egaña M, Blondé W, Illarramendi A, Bilbao I, De Baets B, Stevens R, Mironov V, Kuiper M: **The Cell Cycle Ontology: An application ontology for the representation and integrated analysis of the cell cycle process.** *Genome Biol* 2009, **10:**R58.
19. Fellbaum C: **WordNet and wordnets**. In *Encyclopedia of Language and Linguistics*. 2nd edition. Edited by Brown K, *et al*. Oxford: Elsevier; 2005:665–670.
20. Côté RG, Jones P, Apweiler R, Hermjakob H: **The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries.** *BMC Bioinformatics* 2006, **7:**97.
21. Côté RG, Jones P, Martens L, Apweiler R, Hermjakob H: **The Ontology Lookup Service: more data and better tools for controlled vocabulary queries.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W372–W376.
22. Sintek M: **OntoViz Tab: Visualizing Prot ég é Ontologies**; 2003. [http://protegewiki.stanford.edu/wiki/OntoViz]
23. **IsaViz** [http://www.w3.org/2001/11/IsaViz]
24. Zhong S, Storch KF, Lipan O, Kao MC, Weitz CJ, Wong WH: **GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space.** *Appl Bioinformatics* 2004, **3**(4):261–264.
25. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN: **GoMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data.** *Genome Biol* 2003, **4**(4):R28.
26. Liebig T: **OntoTrack: Fast browsing and easy editing of large ontologies**, *Proceedings of The Second International Workshop on Evaluation of Ontology-based Tools*; 2003.
27. Day-Richter J, Harris MA, Haendel M: **Gene Ontology OBO-Edit Working Group, Lewis S: OBO-Edit–an ontology editor for biologists.** *Bioinformatics* 2007, **23**(16):2198–2200.

doi:10.1186/1471-2105-13-116
Cite this article as: Vercruysse *et al.*: **OLSVis: an animated, interactive visual browser for bio-ontologies.** *BMC Bioinformatics* 2012 **13**:116.

Chapter 7

Discussion

The Semantic Web technology is playing a key role in creating a content-oriented research environment within the life science domain by making structured description of data over the web. To that end, bio-ontologies have provided the required knowledge scaffold for achieving seamless data-integration. Initiatives such as the OBO Foundry and the OBI consortium are committed to improve biological knowledge discovery by making knowledge and data semantically interoperable. The growing acceptance of the semantic web in the life science domain as a means to manage biological knowledge is noteworthy. Particularly, the development of the RDF technology has helped to enable this transition as large amounts of data can be integrated due to the flexibility it offers in modelling. RDF is currently the most widely adopted Semantic Web technology.

This being said, to further establish the semantic web as a robust technology in the life science domain depends greatly on how this caters to the end-users' (biologists') needs. Further development and application of semantic web technologies should go in parallel with the adoption of the integrated resources by the users. The work presented in this thesis covers a number of steps in the direction of the end-users, including efforts to provide them with customised knowledge bases, possibilities to analyse, engineer and visualise ontologies, and some elaborate use cases in collaboration with these end-users that demonstrate the potential of bringing the semantic web closer to the biologists.

- **Paper I:** Computational approaches to analyse biological data have become an essential part in life science research. Biologists often have to utilise many different tools to conduct complex analysis. However, the skilled use of bioinformatics tools frequently poses a steep learning curve for the life science researchers. Additionally, reproducing results becomes a crucial problem while dealing with different tools. Galaxy is aimed to overcome these barriers by catering to the needs of the biologists with limited computational skills. The Galaxy platform offers flexibility in integrating various tools to build reproducible workflows. The ONTO-ToolKit plug-in extends the functionality of the ONTO-PERL software suite, allowing users to handle and manipulate OBO ontologies within the Galaxy environment. As ontologies provide a means to add semantic annotations to the data contained in various biological databases. ONTO-ToolKit allows users to perform ontology-based analysis to improve the depth of their overall analysis.

- **Paper II:** As automated computational reasoning is still an active research area and not commonly available to end-users, it is important to provide alternative solutions that accommodate the exponential growth in biological data and provides scalable inferencing methods to exploit data housed in knowledge bases and the semantic web. The thesis describes a novel method to perform semi-automated reasoning on RDF stores with the use of the Metarel ontology and the SPARQL update language, SPARUL. This implementation allows the inference of new relations from existing relations in RDF stores, thus aiding in the extraction of implicit knowledge through querying. This inferencing approach is used in knowledge bases under the Semantic Systems Biology platform, including the Gene eXpression Knowledge Base (GeXKB).

- **Paper(s) III and IV:** A key aspect in establishing the role of semantic web resources as a knowledge discovery platform is to engage with the domain experts. The participation of end-users promotes development of resources that could be used to address real world use cases. The active collaboration of the Semantic Systems Biology group with experts at the Norwegian University of Science and Technology (NTNU) who are investigating gene regulatory networks has resulted in the development of GeXKB. GeXKB exploits the power of seamless data integration offered by the semantic web technologies. GeXKB demonstrates the ability to build large networks of varied datasets, afforded by homogenising data using RDF. Furthermore, the use cases demonstrate the potential of GeXKB in the identification of candidate regulators of transcription factors and their target genes in a given biological system. To this end, SPARQL technology plays a crucial role in interrogating the data represented in RDF. SPARQL in conjunction with pre-computed inferencing allows addressing questions that were unapproachable at the time the data were produced, strategically filling knowledge gaps.

- **Paper V:** As the complexity of bio-ontologies grows, visualisation plays an important role in grasping the knowledge represented in these ontologies. OLSVis was developed to provide a scalable solution to make ontology browsing intuitive. OLSVis offers a user-friendly, interactive visualisation environment for e.g. the browsing of large ontology branches (including an analysis of the path to the root), viewing the local neighbourhood of terms of interest or alternatively viewing the relationship between two ontological terms.

Enhancing semantic data integration:

As the semantic web technology gains acceptance in the life science domain, it is observed that the full potential of semantically encoded knowledge for querying, and hypotheses generation has not yet been completely realised. All the semantic web resources constructed so far are essentially ‘data warehouses’ with all the classical shortcomings such as a large up-front time investment required for data integration and querying, technical challenges with respect to the infrastructure, maintenance issues and data redundancy. Projects such as SWObjects [1] have used RDF to explore the possibility of utilising query federation for fast track data integration avoiding the aforementioned technical shortcomings, offering an interesting alternative to traditional ‘data warehousing’ approach. However, query federation is still in its nascent stage and has been hampered by RDF’s major strength: its flexibility. RDF allows for a variety of modeling practices creating differences among the knowledge bases in the way they use RDF [2], which may result in the production of non-interoperable resources, such as the use of different types of predicates to represent the same datasets across different semantic web resources. Therefore, in the process of developing real world use cases, a semantic web specialist would be required to spend considerable time to understand the layout of the knowledge bases; then formulate suitable queries (based on the biological question) and produce consolidated results for the biologists. Although these resources follow a consistent pattern in the formulation of their URIs, the user needs to be acquainted with the varying usage of synonymous predicates to query different semantic web resources housing similar data (with possible differences in the scope). For instance, in GeXKB, biological process (BP), molecular function (MF) and interaction (INT) terms are linked to their corresponding proteins using the predicates *has_participant* (*http://www.semantic-systems-biology.org/SSB#has_participant*), *has_function* (*http://www.semantic-systems-biology.org/SSB#has_function*) and *has_agent* (*http://www.semantic-systems-biology.org/SSB#has_agent*), respectively. In contrast, Bio2RDF makes use of the predicates *process* (*http://bio2rdf.org/goa_vocabulary:process*) for BP, *function* (*http://bio2rdf.org/goa_vocabulary:function*) for MF and *interactor_b* (*http://bio2rdf.org/irefindex_vocabulary:interactor_b*) for INT terms.

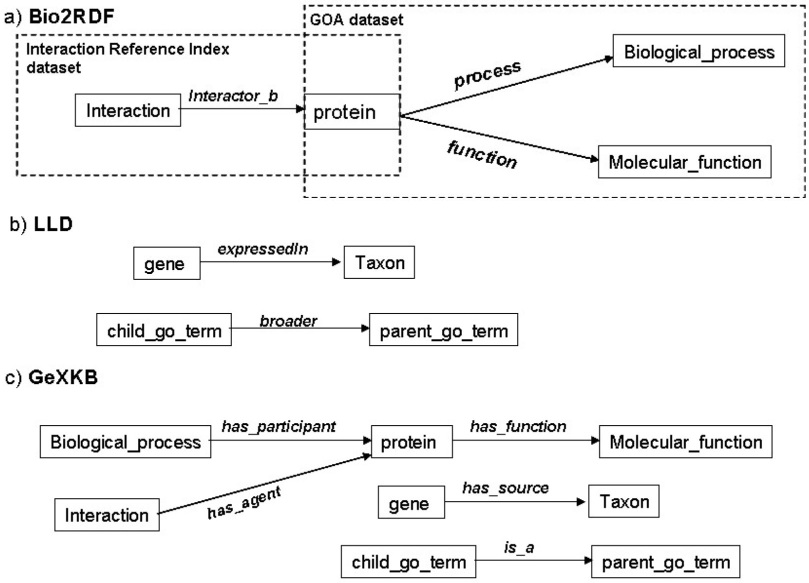


Figure 1: The illustration compares the relations used to represent similar data sets across three different semantic web resources. ***a)*** Represents the Gene Ontology Annotations data set and IntAct data set modeled in Bio2RDF; ***b)*** represents relation between NCBI Gene term to its corresponding taxon term and parent-child Gene Ontology relations used in Linked Life Data and finally ***c)*** shows an alternative data representation of similar data sets (Bio2RDF and LLD) modelled in GeXKB.

Similarly, GeXKB uses *has_source* (*http://www.semantic-systems-biology.org/SSB#has_source*) to link gene/protein terms to their corresponding taxon and child-parent GO term relationship, represented by *is_a* (*http://www.semantic-systems-biology.org/SSB#is_a*), whereas Linked Life Data (LLD) uses *expressedIn* (*http://linkedlifedata.com/resource/entrezgene/expressinedIn*) and *broader* (*http://www.w3.org/2004/02/skos/ core#broader*), respectively (Figure 1). Therefore, a clear distinction between the representations of these datasets is observed.

To enhance the process of knowledge discovery, data-specific queries across different semantic web resources have to be executable. Hence, there is a need to develop approaches that mitigate the variety of otherwise synonymous predicates. One way to realise this is by formulating SPARUL-based rule transformations that effectively align the knowledge bases and ease the process of querying. For example, SPARUL INSERT constructs can be used to map the chosen predicates from Bio2RDF and LLD into the GeXO graph of GeXKB as follows:

a) Predicate mapping between Bio2RDF (Gene Ontology Annotation graph) and GeXKB (GeXO graph):

```
BASE <http://www.semantic-systems-biology.org/>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>
PREFIX goa:<http://bio2rdf.org/goa_vocabulary:>
INSERT INTO GRAPH <GeXO> {
  ?subject goa:process ?object.
}
WHERE {
  GRAPH <GeXO> {
    ?object ssb:has_participant ?subject.
  }
}
```

b) Predicate mapping between LLD (NCBI Entrez graph) and GeXKB (GeXO graph):

```
BASE <http://www.semantic-systems-biology.org/>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>
PREFIX entrezgene:<http://linkedlifedata.com/resource/entrezgene/>
INSERT INTO GRAPH <GeXO> {
  ?subject entrezgene:expressedIn ?object.
}
WHERE {
  GRAPH <GeXO> {
   ?subject ssb:has_participant ?object.
  ?subject rdf:type ssb:NCBIGene.
 }
}
```

The transformation rules must take into account the difference in the way the RDF data has been modelled. The aforementioned examples provide a clear picture on how SPARUL transformations would work. The first case (*a*) concerns with proteins' participation in a BP, deals with inverse relations. Apart from the replacement of the relations, the subjects and the objects must be inverted to map it to the triples in GeXKB. This example demonstrates alignment of relations where one of the knowledge base uses all-some relations (GeXKB). In the example *b*, a distinction in the modelling practice is observed in which GeXKB uses a generic *has_source* predicate to link gene and proteins to their corresponding taxon, whereas LLD uses the *expressedIn* relation for the NCBI Gene data sets. Hence, the relation should be mapped to GeXKB genes only, by restricting on the meta-class ssb:NCBIGene. Furthermore, once the transformation rules are in place generic queries could be launched to GeXKB using predicates from Bio2RDF or LLD, for example:

```
BASE <http://www.semantic-systems-biology.org/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>
PREFIX goa:<http://bio2rdf.org/goa_vocabulary:>
PREFIX graph: <GeXO>
SELECT ?subject ?object
WHERE {
GRAPH graph: {
?subject goa:process ?object.
}
}
```

This query would return proteins and their corresponding BPs, although originally the knowledge was modelled as an all-some relation type (BP *has_participant* protein); the incorporation of SPARUL rules bridges the gap, in this case between Bio2RDF and GeXKB.

Thus, these transformation rules could provide an ad-hoc solution to promote data integration between disparate semantic web resources. In principle, the SPARUL transformation rules could be automated by a) modelling a set of relations representing biomedical knowledge; b) developing applications that would fetch predicates from various semantic web resources and use the model as template to make rule transformations. This will perceptibly affect only the sections the resources have in common and thus effectively aligning resources and enhancing the process of query federation.

Reference

[1] Prud'hommeaux, E., Deus, H. and Marshall, M. S. Tutorial: Query Federation with SWObjects. *In:* SWAT4LS, 2010. Available from *Nature Preceedings.*

[2] Cheung, K. H., Frost, H. R., Marshall, M. S., Prud'hommeaux, E., Samwald, M.. Zhao, J. and Paschke A. (2009) A journey to Semantic Web query federation in the life sciences. *BMC Bioinformatics,* **10**(10)**,** S10.

Chapter 8

Conclusions

The work presented in this thesis demonstrates a series of advances made in bringing Semantic Systems Biology to biological domain experts. These advances cover a variety of aspects of Semantic Systems Biology: developing tools that allow biologists to exploit bio-ontology based analysis as part of biological workflows; the development of intuitive visualisation approaches; the building of customised knowledge bases; the integration of additional data sets on users request; a multidisciplinary approach in the conversion of biological questions in SPARQL queries and the analysis of query results.

Biologists are beginning to adopt Systems Biology approaches to study biological systems in their entirety, as opposed to working on studying functional aspects only of single components in isolation from a system. The semantic web technology complements this approach by providing a sound framework for biological knowledge integration and management. Knowledge bases semantically enriched by bio-ontologies and founded on the RDF technology allow biologists to ask biological questions pertaining to their domain of interest. This is achieved by following and developing new and efficient data integration methods that enable explicit description of data and their relationships. Furthermore, querying and inferencing are key components that potentially reduce the knowledge discovery cycle for the biologists. This has been demonstrated with the development of the Gene eXpression Knowledge Base. This work indicates the level of knowledge discovery that can be achieved with active involvement of the user group in designing knowledge resources and formulation of queries. The collaboration has resulted in the identification of novel components involved in transcription factor regulation in a biological system and provides a proof of concept to the Semantic Systems Biology paradigm in the extension of biological networks.

The full power of Semantic Systems Biology paradigm has even more potential than what is described here, however, this cannot as yet be exploited. The choice taken here, in essence the development of the 'data warehouse' type knowledge base presents a steep learning curve to master the integration of data from various sources, and poses a considerable liability with respect to maintenance. As large amounts of RDF data are being made available for querying, by various primary data providers, future semantic web systems need to effectively build on this trend and support the federated querying of *data warehouse* type knowledge bases together with the external RDF stores.

However, the execution of federated queries is far from being straightforward as it is severely hindered by the variation in modelling practices adopted by independent RDF stores to represent biological data. A number of initiatives have been undertaken to harmonise the use of RDF. The Linked Open Data project [1] provides a list of recommendations for consistent RDF representation. The Vocabulary of Interlinked Datasets (VoID) [2] is an emerging standard to facilitate the linking of various datasets by providing a common vocabulary to describe data in RDF Schema. The Banff Manifesto [3] provides some best practices for the design and implementation of RDF documents in the life science domain. The Concept Web Alliance (CWA) [4] developed an initial proposal of the nano-publication model that enables the aggregation of fine-grained scientific information across the web in RDF. Furthermore, members of the BioRDF task force have produced substantial literature on guidelines of how to produce RDF[5, 6, 7]. Nonetheless, there is a need for consolidating the aforementioned efforts for developing RDF standard practices to improve machine readability. These issues should be discussed and addressed at the community level, and standardisation efforts similar to the MIAME initiative must be broadly accepted in order to unleash the full potential of the semantic web for hypothesis generation, *in-silico* validation, and the answering of complex biological questions.

Finally, the current knowledge bases need a more sophisticated and user-friendly front-end interface. While experts can easily retrieve knowledge from semantic web resources it still is not very friendly or intuitive to the non-expert users. The general practice is to provide standard 'backbone' queries that can easily be modified and customised to guide the lay user to address specific questions to the knowledge housed in these resources. In order to generate hypotheses from complex questions the user still needs to have a moderate knowledge of SPARQL. Therefore, further work is required to for instance build a natural language query interface that intuitively converts natural language questions to SPARQL queries. Such an application should take into account the evolving architecture of the underlying resource in terms of the RDF model. Although some advances have been made in this regard [8, 9, 10] outside the bioinformatics domain and in principle offer great potential, these approaches must be carefully studied and adapted to suit the needs of the life science domain.

Reference

[1] LOD project: Linking Open Data (LOD), W3C: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData#Project_Description

[2] Describing Linked Datasets with the VoID Vocabulary, W3C:, http://www.w3.org/2001/sw/interest/void/

[3] HCLS-DI Banff Manifesto 2007, Banff: http://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff_Manifesto

[4] Groth, P., Gibson, A. and Velterop, J. (2010) The anatomy of a nanopublication. *Information Services and Use,* **30,** 51-56.

[5] Deus, H., et al. (2010) Provenance of Microarray Experiments for a Better Understanding of Experiment Results. *In:* Proceeding of the second International Workshop on the role of Semantic Web in Provenance Management (SWPM2010, ISWC), Shanghai, China.

[6] Deus, H., et al. (2012) Translating standards into practice – one semantic web API for gene expression. *Journal of Biomedical Informatics,* **45**(4), 782-794.

[7] Marshall, M. S., et al. (2012) Emerging practices for mapping and linking life sciences data using RDF — a case series. *WebSemantics: Science, Services and Agents on the World Wide Web*, **4**, 2-13.

[8] Bernstein, A., Kaufmann, E. and Kaiser C. (2005) Querying the semantic web with ginseng: A guided input natural language search engine. *15th Workshop on Information Technologies and Systems*, *Las Vegas*, 112-126.

[9] Lehmann, J. and Bühmann, L. (2011) AutoSPARQL: Let users query your knowledge base. *The Semantic Web: Research and Applications*, *Springer Berlin Heidelberg*, 63-79.

[10] Ngonga, N., et al. (2013) Sorry, i don't speak SPARQL: translating SPARQL queries into natural language. *Proceedings of the 22nd international conference on World Wide Web*, *International World Wide Web Conferences Steering Committee*.

Doctoral theses in the Department of Biology

Doctoral theses in Biology
Norwegian University of Science and Technology
Department of Biology

Year	Name	Degree	Title
1974	Tor-Henning Iversen	Dr. philos Botany	The roles of statholiths, auxin transport, and auxin metabolism in root gravitropism
1978	Tore Slagsvold	Dr. philos Zoology	Breeding events of birds in relation to spring temperature and environmental phenology
1978	Egil Sakshaug	Dr.philos Botany	"The influence of environmental factors on the chemical composition of cultivated and natural populations of marine phytoplankton"
1980	Arnfinn Langeland	Dr. philos Zoology	Interaction between fish and zooplankton populations and their effects on the material utilization in a freshwater lake
1980	Helge Reinertsen	Dr. philos Botany	The effect of lake fertilization on the dynamics and stability of a limnetic ecosystem with special reference to the phytoplankton
1982	Gunn Mari Olsen	Dr. scient Botany	Gravitropism in roots of *Pisum sativum* and *Arabidopsis thaliana*
1982	Dag Dolmen	Dr. philos Zoology	Life aspects of two sympartic species of newts (*Triturus, Amphibia*) in Norway, with special emphasis on their ecological niche segregation
1984	Eivin Røskaft	Dr. philos Zoology	Sociobiological studies of the rook *Corvus frugilegus*
1984	Anne Margrethe Cameron	Dr. scient Botany	Effects of alcohol inhalation on levels of circulating testosterone, follicle stimulating hormone and luteinzing hormone in male mature rats
1984	Asbjørn Magne Nilsen	Dr. scient Botany	Alveolar macrophages from expectorates – Biological monitoring of workers exosed to occupational air pollution. An evaluation of the AM-test
1985	Jarle Mork	Dr. philos Zoology	Biochemical genetic studies in fish
1985	John Solem	Dr. philos Zoology	Taxonomy, distribution and ecology of caddisflies (*Trichoptera*) in the Dovrefjell mountains
1985	Randi E. Reinertsen	Dr. philos Zoology	Energy strategies in the cold: Metabolic and thermoregulatory adaptations in small northern birds
1986	Bernt-Erik Sæther	Dr. philos Zoology	Ecological and evolutionary basis for variation in reproductive traits of some vertebrates: A comparative approach
1986	Torleif Holthe	Dr. philos Zoology	Evolution, systematics, nomenclature, and zoogeography in the polychaete orders *Oweniimorpha* and *Terebellomorpha*, with special reference to the Arctic and Scandinavian fauna
1987	Helene Lampe	Dr. scient	The function of bird song in mate attraction and territorial defence, and the importance of song

		Zoology	repertoires
1987	Olav Hogstad	Dr. philos Zoology	Winter survival strategies of the Willow tit *Parus montanus*
1987	Jarle Inge Holten	Dr. philos Botany	Autecological investigations along a coust-inland transect at Nord-Møre, Central Norway
1987	Rita Kumar	Dr. scient Botany	Somaclonal variation in plants regenerated from cell cultures of *Nicotiana sanderae* and *Chrysanthemum morifolium*
1987	Bjørn Åge Tømmerås	Dr. scient. Zoolog	Olfaction in bark beetle communities: Interspecific interactions in regulation of colonization density, predator - prey relationship and host attraction
1988	Hans Christian Pedersen	Dr. philos Zoology	Reproductive behaviour in willow ptarmigan with special emphasis on territoriality and parental care
1988	Tor G. Heggberget	Dr. philos Zoology	Reproduction in Atlantic Salmon (*Salmo salar*): Aspects of spawning, incubation, early life history and population structure
1988	Marianne V. Nielsen	Dr. scient Zoology	The effects of selected environmental factors on carbon allocation/growth of larval and juvenile mussels (*Mytilus edulis*)
1988	Ole Kristian Berg	Dr. scient Zoology	The formation of landlocked Atlantic salmon (*Salmo salar* L.)
1989	John W. Jensen	Dr. philos Zoology	Crustacean plankton and fish during the first decade of the manmade Nesjø reservoir, with special emphasis on the effects of gill nets and salmonid growth
1989	Helga J. Vivås	Dr. scient Zoology	Theoretical models of activity pattern and optimal foraging: Predictions for the Moose *Alces alces*
1989	Reidar Andersen	Dr. scient Zoology	Interactions between a generalist herbivore, the moose *Alces alces*, and its winter food resources: a study of behavioural variation
1989	Kurt Ingar Draget	Dr. scient Botany	Alginate gel media for plant tissue culture
1990	Bengt Finstad	Dr. scient Zoology	Osmotic and ionic regulation in Atlantic salmon, rainbow trout and Arctic charr: Effect of temperature, salinity and season
1990	Hege Johannesen	Dr. scient Zoology	Respiration and temperature regulation in birds with special emphasis on the oxygen extraction by the lung
1990	Åse Krøkje	Dr. scient Botany	The mutagenic load from air pollution at two work-places with PAH-exposure measured with Ames Salmonella/microsome test
1990	Arne Johan Jensen	Dr. philos Zoology	Effects of water temperature on early life history, juvenile growth and prespawning migrations of Atlantic salmion (*Salmo salar*) and brown trout (*Salmo trutta*): A summary of studies in Norwegian streams
1990	Tor Jørgen Almaas	Dr. scient Zoology	Pheromone reception in moths: Response characteristics of olfactory receptor neurons to intra- and interspecific chemical cues
1990	Magne Husby	Dr.	Breeding strategies in birds: Experiments with the

		scient Zoology	Magpie *Pica pica*
1991	Tor Kvam	Dr. scient Zoology	Population biology of the European lynx (*Lynx lynx*) in Norway
1991	Jan Henning L'Abêe Lund	Dr. philos Zoology	Reproductive biology in freshwater fish, brown trout *Salmo trutta* and roach *Rutilus rutilus* in particular
1991	Asbjørn Moen	Dr. philos Botany	The plant cover of the boreal uplands of Central Norway. I. Vegetation ecology of Sølendet nature reserve; haymaking fens and birch woodlands
1991	Else Marie Løbersli	Dr. scient Botany	Soil acidification and metal uptake in plants
1991	Trond Nordtug	Dr. scient Zoology	Reflctometric studies of photomechanical adaptation in superposition eyes of arthropods
1991	Thyra Solem	Dr. scient Botany	Age, origin and development of blanket mires in Central Norway
1991	Odd Terje Sandlund	Dr. philos Zoology	The dynamics of habitat use in the salmonid genera *Coregonus* and *Salvelinus*: Ontogenic niche shifts and polymorphism
1991	Nina Jonsson	Dr. philos	Aspects of migration and spawning in salmonids
1991	Atle Bones	Dr. scient Botany	Compartmentation and molecular properties of thioglucoside glucohydrolase (myrosinase)
1992	Torgrim Breiehagen	Dr. scient Zoology	Mating behaviour and evolutionary aspects of the breeding system of two bird species: the Temminck's stint and the Pied flycatcher
1992	Anne Kjersti Bakken	Dr. scient Botany	The influence of photoperiod on nitrate assimilation and nitrogen status in timothy (*Phleum pratense* L.)
1992	Tycho Anker-Nilssen	Dr. scient Zoology	Food supply as a determinant of reproduction and population development in Norwegian Puffins *Fratercula arctica*
1992	Bjørn Munro Jenssen	Dr. philos Zoology	Thermoregulation in aquatic birds in air and water: With special emphasis on the effects of crude oil, chemically treated oil and cleaning on the thermal balance of ducks
1992	Arne Vollan Aarset	Dr. philos Zoology	The ecophysiology of under-ice fauna: Osmotic regulation, low temperature tolerance and metabolism in polar crustaceans.
1993	Geir Slupphaug	Dr. scient Botany	Regulation and expression of uracil-DNA glycosylase and O^6-methylguanine-DNA methyltransferase in mammalian cells
1993	Tor Fredrik Næsje	Dr. scient Zoology	Habitat shifts in coregonids.
1993	Yngvar Asbjørn Olsen	Dr. scient Zoology	Cortisol dynamics in Atlantic salmon, *Salmo salar* L.: Basal and stressor-induced variations in plasma levels ans some secondary effects.
1993	Bård Pedersen	Dr. scient Botany	Theoretical studies of life history evolution in modular and clonal organisms

1993	Ole Petter Thangstad	Dr. scient Botany	Molecular studies of myrosinase in Brassicaceae
1993	Thrine L. M. Heggberget	Dr. scient Zoology	Reproductive strategy and feeding ecology of the Eurasian otter *Lutra lutra*.
1993	Kjetil Bevanger	Dr. scient. Zoology	Avian interactions with utility structures, a biological approach.
1993	Kåre Haugan	Dr. scient Bothany	Mutations in the replication control gene trfA of the broad host-range plasmid RK2
1994	Peder Fiske	Dr. scient. Zoology	Sexual selection in the lekking great snipe (*Gallinago media*): Male mating success and female behaviour at the lek
1994	Kjell Inge Reitan	Dr. scient Botany	Nutritional effects of algae in first-feeding of marine fish larvae
1994	Nils Røv	Dr. scient Zoology	Breeding distribution, population status and regulation of breeding numbers in the northeast-Atlantic Great Cormorant *Phalacrocorax carbo carbo*
1994	Annette-Susanne Hoepfner	Dr. scient Botany	Tissue culture techniques in propagation and breeding of Red Raspberry (*Rubus idaeus* L.)
1994	Inga Elise Bruteig	Dr. scient Bothany	Distribution, ecology and biomonitoring studies of epiphytic lichens on conifers
1994	Geir Johnsen	Dr. scient Botany	Light harvesting and utilization in marine phytoplankton: Species-specific and photoadaptive responses
1994	Morten Bakken	Dr. scient Zoology	Infanticidal behaviour and reproductive performance in relation to competition capacity among farmed silver fox vixens, *Vulpes vulpes*
1994	Arne Moksnes	Dr. philos Zoology	Host adaptations towards brood parasitism by the Cockoo
1994	Solveig Bakken	Dr. scient Bothany	Growth and nitrogen status in the moss *Dicranum majus* Sm. as influenced by nitrogen supply
1994	Torbjørn Forseth	Dr. scient Zoology	Bioenergetics in ecological and life history studies of fishes.
1995	Olav Vadstein	Dr. philos Botany	The role of heterotrophic planktonic bacteria in the cycling of phosphorus in lakes: Phosphorus requirement, competitive ability and food web interactions
1995	Hanne Christensen	Dr. scient Zoology	Determinants of Otter *Lutra lutra* distribution in Norway: Effects of harvest, polychlorinated biphenyls (PCBs), human population density and competition with mink *Mustela vision*
1995	Svein Håkon Lorentsen	Dr. scient Zoology	Reproductive effort in the Antarctic Petrel *Thalassoica antarctica*; the effect of parental body size and condition
1995	Chris Jørgen Jensen	Dr.	The surface electromyographic (EMG) amplitude as

		scient Zoology	an estimate of upper trapezius muscle activity
1995	Martha Kold Bakkevig	Dr. scient Zoology	The impact of clothing textiles and construction in a clothing system on thermoregulatory responses, sweat accumulation and heat transport
1995	Vidar Moen	Dr. scient Zoology	Distribution patterns and adaptations to light in newly introduced populations of *Mysis relicta* and constraints on Cladoceran and Char populations
1995	Hans Haavardsholm Blom	Dr. philos Bothany	A revision of the *Schistidium apocarpum* complex in Norway and Sweden
1996	Jorun Skjærmo	Dr. scient Botany	Microbial ecology of early stages of cultivated marine fish; inpact fish-bacterial interactions on growth and survival of larvae
1996	Ola Ugedal	Dr. scient Zoology	Radiocesium turnover in freshwater fishes
1996	Ingibjørg Einarsdottir	Dr. scient Zoology	Production of Atlantic salmon (*Salmo salar*) and Arctic charr (*Salvelinus alpinus*): A study of some physiological and immunological responses to rearing routines
1996	Christina M. S. Pereira	Dr. scient Zoology	Glucose metabolism in salmonids: Dietary effects and hormonal regulation
1996	Jan Fredrik Børseth	Dr. scient Zoology	The sodium energy gradients in muscle cells of *Mytilus edulis* and the effects of organic xenobiotics
1996	Gunnar Henriksen	Dr. scient Zoology	Status of Grey seal *Halichoerus grypus* and Harbour seal *Phoca vitulina* in the Barents sea region
1997	Gunvor Øie	Dr. scient Bothany	Eevalution of rotifer *Brachionus plicatilis* quality in early first feeding of turbot *Scophtalmus maximus* L. larvae
1997	Håkon Holien	Dr. scient Botany	Studies of lichens in spurce forest of Central Norway. Diversity, old growth species and the relationship to site and stand parameters
1997	Ole Reitan	Dr. scient. Zoology	Responses of birds to habitat disturbance due to damming
1997	Jon Arne Grøttum	Dr. scient. Zoology	Physiological effects of reduced water quality on fish in aquaculture
1997	Per Gustav Thingstad	Dr. scient. Zoology	Birds as indicators for studying natural and human-induced variations in the environment, with special emphasis on the suitability of the Pied Flycatcher
1997	Torgeir Nygård	Dr. scient Zoology	Temporal and spatial trends of pollutants in birds in Norway: Birds of prey and Willow Grouse used as Biomonitors
1997	Signe Nybø	Dr. scient. Zoology	Impacts of long-range transported air pollution on birds with particular reference to the dipper *Cinclus cinclus* in southern Norway
1997	Atle Wibe	Dr. scient. Zoology	Identification of conifer volatiles detected by receptor neurons in the pine weevil (*Hylobius abietis*), analysed by gas chromatography linked to electrophysiology and to mass spectrometry
1997	Rolv Lundheim	Dr.	Adaptive and incidental biological ice nucleators

		scient Zoology	
1997	Arild Magne Landa	Dr. scient Zoology	Wolverines in Scandinavia: ecology, sheep depredation and conservation
1997	Kåre Magne Nielsen	Dr. scient Botany	An evolution of possible horizontal gene transfer from plants to sail bacteria by studies of natural transformation in *Acinetobacter calcoacetius*
1997	Jarle Tufto	Dr. scient Zoology	Gene flow and genetic drift in geographically structured populations: Ecological, population genetic, and statistical models
1997	Trygve Hesthagen	Dr. philos Zoology	Population responces of Arctic charr (*Salvelinus alpinus* (L.)) and brown trout (*Salmo trutta* L.) to acidification in Norwegian inland waters
1997	Trygve Sigholt	Dr. philos Zoology	Control of Parr-smolt transformation and seawater tolerance in farmed Atlantic Salmon (*Salmo salar*) Effects of photoperiod, temperature, gradual seawater acclimation, NaCl and betaine in the diet
1997	Jan Østnes	Dr. scient Zoology	Cold sensation in adult and neonate birds
1998	Seethaledsumy Visvalingam	Dr. scient Botany	Influence of environmental factors on myrosinases and myrosinase-binding proteins
1998	Thor Harald Ringsby	Dr. scient Zoology	Variation in space and time: The biology of a House sparrow metapopulation
1998	Erling Johan Solberg	Dr. scient. Zoology	Variation in population dynamics and life history in a Norwegian moose (*Alces alces*) population: consequences of harvesting in a variable environment
1998	Sigurd Mjøen Saastad	Dr. scient Botany	Species delimitation and phylogenetic relationships between the Sphagnum recurvum complex (Bryophyta): genetic variation and phenotypic plasticity
1998	Bjarte Mortensen	Dr. scient Botany	Metabolism of volatile organic chemicals (VOCs) in a head liver S9 vial equilibration system in vitro
1998	Gunnar Austrheim	Dr. scient Botany	Plant biodiversity and land use in subalpine grasslands. – A conservtaion biological approach
1998	Bente Gunnveig Berg	Dr. scient Zoology	Encoding of pheromone information in two related moth species
1999	Kristian Overskaug	Dr. scient Zoology	Behavioural and morphological characteristics in Northern Tawny Owls *Strix aluco*: An intra- and interspecific comparative approach
1999	Hans Kristen Stenøien	Dr. scient Bothany	Genetic studies of evolutionary processes in various populations of nonvascular plants (mosses, liverworts and hornworts)
1999	Trond Arnesen	Dr. scient Botany	Vegetation dynamics following trampling and burning in the outlying haylands at Sølendet, Central Norway
1999	Ingvar Stenberg	Dr. scient Zoology	Habitat selection, reproduction and survival in the White-backed Woodpecker *Dendrocopos leucotos*
1999	Stein Olle Johansen	Dr.	A study of driftwood dispersal to the Nordic Seas by

		scient Botany	dendrochronology and wood anatomical analysis
1999	Trina Falck Galloway	Dr. scient Zoology	Muscle development and growth in early life stages of the Atlantic cod (*Gadus morhua* L.) and Halibut (*Hippoglossus hippoglossus* L.)
1999	Marianne Giæver	Dr. scient Zoology	Population genetic studies in three gadoid species: blue whiting (*Micromisistius poutassou*), haddock (*Melanogrammus aeglefinus*) and cod (*Gradus morhua*) in the North-East Atlantic
1999	Hans Martin Hanslin	Dr. scient Botany	The impact of environmental conditions of density dependent performance in the boreal forest bryophytes *Dicranum majus*, *Hylocomium splendens*, *Plagiochila asplenigides*, *Ptilium crista-castrensis* and *Rhytidiadelphus lokeus*
1999	Ingrid Bysveen Mjølnerød	Dr. scient Zoology	Aspects of population genetics, behaviour and performance of wild and farmed Atlantic salmon (*Salmo salar*) revealed by molecular genetic techniques
1999	Else Berit Skagen	Dr. scient Botany	The early regeneration process in protoplasts from *Brassica napus* hypocotyls cultivated under various g-forces
1999	Stein-Are Sæther	Dr. philos Zoology	Mate choice, competition for mates, and conflicts of interest in the Lekking Great Snipe
1999	Katrine Wangen Rustad	Dr. scient Zoology	Modulation of glutamatergic neurotransmission related to cognitive dysfunctions and Alzheimer's disease
1999	Per Terje Smiseth	Dr. scient Zoology	Social evolution in monogamous families: mate choice and conflicts over parental care in the Bluethroat (*Luscinia s. svecica*)
1999	Gunnbjørn Bremset	Dr. scient Zoology	Young Atlantic salmon (*Salmo salar* L.) and Brown trout (*Salmo trutta* L.) inhabiting the deep pool habitat, with special reference to their habitat use, habitat preferences and competitive interactions
1999	Frode Ødegaard	Dr. scient Zoology	Host spesificity as parameter in estimates of arhrophod species richness
1999	Sonja Andersen	Dr. scient Bothany	Expressional and functional analyses of human, secretory phospholipase A2
2000	Ingrid Salvesen	Dr. scient Botany	Microbial ecology in early stages of marine fish: Development and evaluation of methods for microbial management in intensive larviculture
2000	Ingar Jostein Øien	Dr. scient Zoology	The Cuckoo (*Cuculus canorus*) and its host: adaptions and counteradaptions in a coevolutionary arms race
2000	Pavlos Makridis	Dr. scient Botany	Methods for the microbial econtrol of live food used for the rearing of marine fish larvae
2000	Sigbjørn Stokke	Dr. scient Zoology	Sexual segregation in the African elephant (*Loxodonta africana*)
2000	Odd A. Gulseth	Dr. philos Zoology	Seawater tolerance, migratory behaviour and growth of Charr, (*Salvelinus alpinus*), with emphasis on the high Arctic Dieset charr on Spitsbergen, Svalbard
2000	Pål A. Olsvik	Dr.	Biochemical impacts of Cd, Cu and Zn on brown

		scient Zoology	trout (*Salmo trutta*) in two mining-contaminated rivers in Central Norway
2000	Sigurd Einum	Dr. scient Zoology	Maternal effects in fish: Implications for the evolution of breeding time and egg size
2001	Jan Ove Evjemo	Dr. scient Zoology	Production and nutritional adaptation of the brine shrimp *Artemia* sp. as live food organism for larvae of marine cold water fish species
2001	Olga Hilmo	Dr. scient Botany	Lichen response to environmental changes in the managed boreal forset systems
2001	Ingebrigt Uglem	Dr. scient Zoology	Male dimorphism and reproductive biology in corkwing wrasse (*Symphodus melops* L.)
2001	Bård Gunnar Stokke	Dr. scient Zoology	Coevolutionary adaptations in avian brood parasites and their hosts
2002	Ronny Aanes	Dr. scient	Spatio-temporal dynamics in Svalbard reindeer (*Rangifer tarandus platyrhynchus*)
2002	Mariann Sandsund	Dr. scient Zoology	Exercise- and cold-induced asthma. Respiratory and thermoregulatory responses
2002	Dag-Inge Øien	Dr. scient Botany	Dynamics of plant communities and populations in boreal vegetation influenced by scything at Sølendet, Central Norway
2002	Frank Rosell	Dr. scient Zoology	The function of scent marking in beaver (*Castor fiber*)
2002	Janne Østvang	Dr. scient Botany	The Role and Regulation of Phospholipase A_2 in Monocytes During Atherosclerosis Development
2002	Terje Thun	Dr.philos Biology	Dendrochronological constructions of Norwegian conifer chronologies providing dating of historical material
2002	Birgit Hafjeld Borgen	Dr. scient Biology	Functional analysis of plant idioblasts (Myrosin cells) and their role in defense, development and growth
2002	Bård Øyvind Solberg	Dr. scient Biology	Effects of climatic change on the growth of dominating tree species along major environmental gradients
2002	Per Winge	Dr. scient Biology	The evolution of small GTP binding proteins in cellular organisms. Studies of RAC GTPases in *Arabidopsis thaliana* and the Ral GTPase from *Drosophila melanogaster*
2002	Henrik Jensen	Dr. scient Biology	Causes and consequenses of individual variation in fitness-related traits in house sparrows
2003	Jens Rohloff	Dr. philos Biology	Cultivation of herbs and medicinal plants in Norway – Essential oil production and quality control
2003	Åsa Maria O. Espmark Wibe	Dr. scient Biology	Behavioural effects of environmental pollution in threespine stickleback *Gasterosteus aculeatur* L.
2003	Dagmar Hagen	Dr. scient Biology	Assisted recovery of disturbed arctic and alpine vegetation – an integrated approach

2003	Bjørn Dahle	Dr. scient Biology	Reproductive strategies in Scandinavian brown bears
2003	Cyril Lebogang Taolo	Dr. scient Biology	Population ecology, seasonal movement and habitat use of the African buffalo (*Syncerus caffer*) in Chobe National Park, Botswana
2003	Marit Stranden	Dr.scient Biology	Olfactory receptor neurones specified for the same odorants in three related Heliothine species (*Helicoverpa armigera, Helicoverpa assulta* and *Heliothis virescens*)
2003	Kristian Hassel	Dr.scient Biology	Life history characteristics and genetic variation in an expanding species, *Pogonatum dentatum*
2003	David Alexander Rae	Dr.scient Biology	Plant- and invertebrate-community responses to species interaction and microclimatic gradients in alpine and Artic environments
2003	Åsa A Borg	Dr.scient Biology	Sex roles and reproductive behaviour in gobies and guppies: a female perspective
2003	Eldar Åsgard Bendiksen	Dr.scient Biology	Environmental effects on lipid nutrition of farmed Atlantic salmon (*Salmo Salar* L.) parr and smolt
2004	Torkild Bakken	Dr.scient Biology	A revision of Nereidinae (Polychaeta, Nereididae)
2004	Ingar Pareliussen	Dr.scient Biology	Natural and Experimental Tree Establishment in a Fragmented Forest, Ambohitantely Forest Reserve, Madagascar
2004	Tore Brembu	Dr.scient Biology	Genetic, molecular and functional studies of RAC GTPases and the WAVE-like regulatory protein complex in *Arabidopsis thaliana*
2004	Liv S. Nilsen	Dr.scient Biology	Coastal heath vegetation on central Norway; recent past, present state and future possibilities
2004	Hanne T. Skiri	Dr.scient Biology	Olfactory coding and olfactory learning of plant odours in heliothine moths. An anatomical, physiological and behavioural study of three related species (*Heliothis virescens, Helicoverpa armigera* and *Helicoverpa assulta*)
2004	Lene Østby	Dr.scient Biology	Cytochrome P4501A (CYP1A) induction and DNA adducts as biomarkers for organic pollution in the natural environment
2004	Emmanuel J. Gerreta	Dr. philos Biology	The Importance of Water Quality and Quantity in the Tropical Ecosystems, Tanzania
2004	Linda Dalen	Dr.scient Biology	Dynamics of Mountain Birch Treelines in the Scandes Mountain Chain, and Effects of Climate Warming
2004	Lisbeth Mehli	Dr.scient Biology	Polygalacturonase-inhibiting protein (PGIP) in cultivated strawberry (*Fragaria* x *ananassa*): characterisation and induction of the gene following fruit infection by *Botrytis cinerea*
2004	Børge Moe	Dr.scient Biology	Energy-Allocation in Avian Nestlings Facing Short-Term Food Shortage
2005	Matilde Skogen Chauton	Dr.scient Biology	Metabolic profiling and species discrimination from High-Resolution Magic Angle Spinning NMR analysis of whole-cell samples
2005	Sten Karlsson	Dr.scient Biology	Dynamics of Genetic Polymorphisms
2005	Terje Bongard	Dr.scient Biology	Life History strategies, mate choice, and parental investment among Norwegians over a 300-year

			period
2005	Tonette Røstelien	ph.d Biology	Functional characterisation of olfactory receptor neurone types in heliothine moths
2005	Erlend Kristiansen	Dr.scient Biology	Studies on antifreeze proteins
2005	Eugen G. Sørmo	Dr.scient Biology	Organochlorine pollutants in grey seal (*Halichoerus grypus*) pups and their impact on plasma thyrid hormone and vitamin A concentrations
2005	Christian Westad	Dr.scient Biology	Motor control of the upper trapezius
2005	Lasse Mork Olsen	ph.d Biology	Interactions between marine osmo- and phagotrophs in different physicochemical environments
2005	Åslaug Viken	ph.d Biology	Implications of mate choice for the management of small populations
2005	Ariaya Hymete Sahle Dingle	ph.d Biology	Investigation of the biological activities and chemical constituents of selected *Echinops* spp. growing in Ethiopia
2005	Anders Gravbrøt Finstad	ph.d Biology	Salmonid fishes in a changing climate: The winter challenge
2005	Shimane Washington Makabu	ph.d Biology	Interactions between woody plants, elephants and other browsers in the Chobe Riverfront, Botswana
2005	Kjartan Østbye	Dr.scient Biology	The European whitefish *Coregonus lavaretus* (L.) species complex: historical contingency and adaptive radiation
2006	Kari Mette Murvoll	ph.d Biology	Levels and effects of persistent organic pollutans (POPs) in seabirds Retinoids and α-tocopherol – potential biomakers of POPs in birds?
2006	Ivar Herfindal	Dr.scient Biology	Life history consequences of environmental variation along ecological gradients in northern ungulates
2006	Nils Egil Tokle	ph.d Biology	Are the ubiquitous marine copepods limited by food or predation? Experimental and field-based studies with main focus on *Calanus finmarchicus*
2006	Jan Ove Gjershaug	Dr.philos Biology	Taxonomy and conservation status of some booted eagles in south-east Asia
2006	Jon Kristian Skei	Dr.scient Biology	Conservation biology and acidification problems in the breeding habitat of amphibians in Norway
2006	Johanna Järnegren	ph.d Biology	Acesta Oophaga and Acesta Excavata – a study of hidden biodiversity
2006	Bjørn Henrik Hansen	ph.d Biology	Metal-mediated oxidative stress responses in brown trout (*Salmo trutta*) from mining contaminated rivers in Central Norway
2006	Vidar Grøtan	ph.d Biology	Temporal and spatial effects of climate fluctuations on population dynamics of vertebrates
2006	Jafari R Kideghesho	ph.d Biology	Wildlife conservation and local land use conflicts in western Serengeti, Corridor Tanzania
2006	Anna Maria Billing	ph.d Biology	Reproductive decisions in the sex role reversed pipefish *Syngnathus typhle*: when and how to invest in reproduction
2006	Henrik Pärn	ph.d Biology	Female ornaments and reproductive biology in the bluethroat
2006	Anders J. Fjellheim	ph.d Biology	Selection and administration of probiotic bacteria to marine fish larvae
2006	P. Andreas Svensson	ph.d Biology	Female coloration, egg carotenoids and reproductive success: gobies as a model system
2007	Sindre A. Pedersen	ph.d	Metal binding proteins and antifreeze proteins in the

		Biology	beetle *Tenebrio molitor* - a study on possible competition for the semi-essential amino acid cysteine
2007	Kasper Hancke	ph.d Biology	Photosynthetic responses as a function of light and temperature: Field and laboratory studies on marine microalgae
2007	Tomas Holmern	ph.d Biology	Bushmeat hunting in the western Serengeti: Implications for community-based conservation
2007	Kari Jørgensen	ph.d Biology	Functional tracing of gustatory receptor neurons in the CNS and chemosensory learning in the moth *Heliothis virescens*
2007	Stig Ulland	ph.d Biology	Functional Characterisation of Olfactory Receptor Neurons in the Cabbage Moth, (*Mamestra brassicae* L.) (Lepidoptera, Noctuidae). Gas Chromatography Linked to Single Cell Recordings and Mass Spectrometry
2007	Snorre Henriksen	ph.d Biology	Spatial and temporal variation in herbivore resources at northern latitudes
2007	Roelof Frans May	ph.d Biology	Spatial Ecology of Wolverines in Scandinavia
2007	Vedasto Gabriel Ndibalema	ph.d Biology	Demographic variation, distribution and habitat use between wildebeest sub-populations in the Serengeti National Park, Tanzania
2007	Julius William Nyahongo	ph.d Biology	Depredation of Livestock by wild Carnivores and Illegal Utilization of Natural Resources by Humans in the Western Serengeti, Tanzania
2007	Shombe Ntaraluka Hassan	ph.d Biology	Effects of fire on large herbivores and their forage resources in Serengeti, Tanzania
2007	Per-Arvid Wold	ph.d Biology	Functional development and response to dietary treatment in larval Atlantic cod *(Gadus morhua* L.) Focus on formulated diets and early weaning
2007	Anne Skjetne Mortensen	ph.d Biology	Toxicogenomics of Aryl Hydrocarbon- and Estrogen Receptor Interactions in Fish: Mechanisms and Profiling of Gene Expression Patterns in Chemical Mixture Exposure Scenarios
2008	Brage Bremset Hansen	ph.d Biology	The Svalbard reindeer (*Rangifer tarandus platyrhynchus*) and its food base: plant-herbivore interactions in a high-arctic ecosystem
2008	Jiska van Dijk	ph.d Biology	Wolverine foraging strategies in a multiple-use landscape
2008	Flora John Magige	ph.d Biology	The ecology and behaviour of the Masai Ostrich (Struthio camelus massaicus) in the Serengeti Ecosystem, Tanzania
2008	Bernt Rønning	ph.d Biology	Sources of inter- and intra-individual variation in basal metabolic rate in the zebra finch, (*Taeniopygia guttata*)
2008	Sølvi Wehn	ph.d Biology	Biodiversity dynamics in semi-natural mountain landscapes. - A study of consequences of changed agricultural practices in Eastern Jotunheimen
2008	Trond Moxness Kortner	ph.d Biology	"The Role of Androgens on previtellogenic oocyte growth in Atlantic cod (*Gadus morhua*): Identification and patterns of differentially expressed genes in relation to Stereological Evaluations"

2008	Katarina Mariann Jørgensen	Dr.Scient Biology	The role of platelet activating factor in activation of growth arrested keratinocytes and re-epithelialisation
2008	Tommy Jørstad	ph.d Biology	Statistical Modelling of Gene Expression Data
2008	Anna Kusnierczyk	ph.d Bilogy	*Arabidopsis thaliana* Responses to Aphid Infestation
2008	Jussi Evertsen	ph.d Biology	Herbivore sacoglossans with photosynthetic chloroplasts
2008	John Eilif Hermansen	ph.d Biology	Mediating ecological interests between locals and globals by means of indicators. A study attributed to the asymmetry between stakeholders of tropical forest at Mt. Kilimanjaro, Tanzania
2008	Ragnhild Lyngved	ph.d Biology	Somatic embryogenesis in *Cyclamen persicum.* Biological investigations and educational aspects of cloning
2008	Line Elisabeth Sundt-Hansen	ph.d Biology	Cost of rapid growth in salmonid fishes
2008	Line Johansen	ph.d Biology	Exploring factors underlying fluctuations in white clover populations – clonal growth, population structure and spatial distribution
2009	Astrid Jullumstrø Feuerherm	ph.d Biology	Elucidation of molecular mechanisms for pro-inflammatory phospholipase A2 in chronic disease
2009	Pål Kvello	ph.d Biology	Neurons forming the network involved in gustatory coding and learning in the moth *Heliothis virescens:* Physiological and morphological characterisation, and integration into a standard brain atlas
2009	Trygve Devold Kjellsen	ph.d Biology	Extreme Frost Tolerance in Boreal Conifers
2009	Johan Reinert Vikan	ph.d Biology	Coevolutionary interactions between common cuckoos *Cuculus canorus* and *Fringilla* finches
2009	Zsolt Volent	ph.d Biology	Remote sensing of marine environment: Applied surveillance with focus on optical properties of phytoplankton, coloured organic matter and suspended matter
2009	Lester Rocha	ph.d Biology	Functional responses of perennial grasses to simulated grazing and resource availability
2009	Dennis Ikanda	ph.d Biology	Dimensions of a Human-lion conflict: Ecology of human predation and persecution of African lions (*Panthera leo*) in Tanzania
2010	Huy Quang Nguyen	ph.d Biology	Egg characteristics and development of larval digestive function of cobia (*Rachycentron canadum*) in response to dietary treatments -Focus on formulated diets
2010	Eli Kvingedal	ph.d Biology	Intraspecific competition in stream salmonids: the impact of environment and phenotype
2010	Sverre Lundemo	ph.d Biology	Molecular studies of genetic structuring and demography in *Arabidopsis* from Northern Europe
2010	Iddi Mihijai Mfunda	ph.d Biology	Wildlife Conservation and People’s livelihoods: Lessons Learnt and Considerations for Improvements. Tha Case of Serengeti Ecosystem, Tanzania
2010	Anton Tinchov Antonov	ph.d Biology	Why do cuckoos lay strong-shelled eggs? Tests of the puncture resistance hypothesis
2010	Anders Lyngstad	ph.d	Population Ecology of *Eriophorum latifolium*, a

		Biology	Clonal Species in Rich Fen Vegetation
2010	Hilde Færevik	ph.d Biology	Impact of protective clothing on thermal and cognitive responses
2010	Ingerid Brænne Arbo	ph.d Medical technology	Nutritional lifestyle changes – effects of dietary carbohydrate restriction in healthy obese and overweight humans
2010	Yngvild Vindenes	ph.d Biology	Stochastic modeling of finite populations with individual heterogeneity in vital parameters
2010	Hans-Richard Brattbakk	ph.d Medical technology	The effect of macronutrient composition, insulin stimulation, and genetic variation on leukocyte gene expression and possible health benefits
2011	Geir Hysing Bolstad	ph.d Biology	Evolution of Signals: Genetic Architecture, Natural Selection and Adaptive Accuracy
2011	Karen de Jong	ph.d Biology	Operational sex ratio and reproductive behaviour in the two-spotted goby (*Gobiusculus flavescens*)
2011	Ann-Iren Kittang	ph.d Biology	*Arabidopsis thaliana* L. adaptation mechanisms to microgravity through the EMCS MULTIGEN-2 experiment on the ISS:– The science of space experiment integration and adaptation to simulated microgravity
2011	Aline Magdalena Lee	ph.d Biology	Stochastic modeling of mating systems and their effect on population dynamics and genetics
2011	Christopher Gravningen Sørmo	ph.d Biology	Rho GTPases in Plants: Structural analysis of ROP GTPases; genetic and functional studies of MIRO GTPases in *Arabidopsis thaliana*
2011	Grethe Robertsen	ph.d Biology	Relative performance of salmonid phenotypes across environments and competitive intensities
2011	Line-Kristin Larsen	ph.d Biology	Life-history trait dynamics in experimental populations of guppy (*Poecilia reticulata*): the role of breeding regime and captive environment
2011	Maxim A. K. Teichert	ph.d Biology	Regulation in Atlantic salmon (*Salmo salar*): The interaction between habitat and density
2011	Torunn Beate Hancke	ph.d Biology	Use of Pulse Amplitude Modulated (PAM) Fluorescence and Bio-optics for Assessing Microalgal Photosynthesis and Physiology
2011	Sajeda Begum	ph.d Biology	Brood Parasitism in Asian Cuckoos: Different Aspects of Interactions between Cuckoos and their Hosts in Bangladesh
2011	Kari J. K. Attramadal	ph.d Biology	Water treatment as an approach to increase microbial control in the culture of cold water marine larvae
2011	Camilla Kalvatn Egset	ph.d Biology	The Evolvability of Static Allometry: A Case Study
2011	AHM Raihan Sarker	ph.d Biology	Conflict over the conservation of the Asian elephant (*Elephas maximus*) in Bangladesh
2011	Gro Dehli Villanger	ph.d Biology	Effects of complex organohalogen contaminant mixtures on thyroid hormone homeostasis in selected arctic marine mammals
2011	Kari Bjørneraas	ph.d Biology	Spatiotemporal variation in resource utilisation by a large herbivore, the moose
2011	John Odden	ph.d Biology	The ecology of a conflict: Eurasian lynx depredation on domestic sheep
2011	Simen Pedersen	ph.d Biology	Effects of native and introduced cervids on small mammals and birds
2011	Mohsen Falahati-	ph.d	Evolutionary consequences of seed banks and seed

	Anbaran	Biology	dispersal in *Arabidopsis*
2012	Jakob Hønborg Hansen	ph.d Biology	Shift work in the offshore vessel fleet: circadian rhythms and cognitive performance
2012	Elin Noreen	ph.d Biology	Consequences of diet quality and age on life-history traits in a small passerine bird
2012	Irja Ida Ratikainen	ph.d Biology	Theoretical and empirical approaches to studying foraging decisions: the past and future of behavioural ecology
2012	Aleksander Handå	ph.d Biology	Cultivation of mussels (*Mytilus edulis):*Feed requirements, storage and integration with salmon (*Salmo salar*) farming
2012	Morten Kraabøl	ph.d Biology	Reproductive and migratory challenges inflicted on migrant brown trour (*Salmo trutta* L) in a heavily modified river
2012	Jisca Huisman	ph.d Biology	Gene flow and natural selection in Atlantic salmon
2012	Maria Bergvik	ph.d Biology	Lipid and astaxanthin contents and biochemical post-harvest stability in *Calanus finmarchicus*
2012	Bjarte Bye Løfaldli	ph.d Biology	Functional and morphological characterization of central olfactory neurons in the model insect *Heliothis virescens*.
2012	Karen Marie Hammer	ph.d Biology.	Acid-base regulation and metabolite responses in shallow- and deep-living marine invertebrates during environmental hypercapnia
2012	Øystein Nordrum Wiggen	ph.d Biology	Optimal performance in the cold
2012	Robert Dominikus Fyumagwa	Dr. Philos.	Anthropogenic and natural influence on disease prevalence at the human –livestock-wildlife interface in the Serengeti ecosystem, Tanzania
2012	Jenny Bytingsvik	ph.d Biology	Organohalogenated contaminants (OHCs) in polar bear mother-cub pairs from Svalbard, Norway Maternal transfer, exposure assessment and thyroid hormone disruptive effects in polar bear cubs
2012	Christer Moe Rolandsen	ph.d Biology	The ecological significance of space use and movement patterns of moose in a variable environment
2012	Erlend Kjeldsberg Hovland	ph.d Biology	Bio-optics and Ecology in *Emiliania huxleyi* Blooms: Field and Remote Sensing Studies in Norwegian Waters
2012	Lise Cats Myhre	ph.d Biology	Effects of the social and physical environment on mating behaviour in a marine fish
2012	Tonje Aronsen	ph.d Biology	Demographic, environmental and evolutionary aspects of sexual selection
2012	Bin Liu	ph.d Biology	Molecular genetic investigation of cell separation and cell death regulation in *Arabidopsis thaliana*
2013	Jørgen Rosvold	ph.d Biology	Ungulates in a dynamic and increasingly human dominated landscape – A millennia-scale perspective
2013	Pankaj Barah	ph.d Biology	Integrated Systems Approaches to Study Plant Stress Responses
2013	Marit Linnerud	ph.d Biology	Patterns in spatial and temporal variation in population abundances of vertebrates
2013	Xinxin Wang	ph.d Biology	Integrated multi-trophic aquaculture driven by nutrient wastes released from Atlantic salmon (*Salmo salar*) farming
2013	Ingrid Ertshus Mathisen	ph.d Biology	Structure, dynamics, and regeneration capacity at the sub-arctic forest-tundra ecotone of northern Norway

			and Kola Peninsula, NW Russia
2013	Anders Foldvik	ph.d Biology	Spatial distributions and productivity in salmonid populations
2013	Anna Marie Holand	ph.d Biology	Statistical methods for estimating intra- and inter-population variation in genetic diversity
2013	Anna Solvang Båtnes	ph.d Biology	Light in the dark – the role of irradiance in the high Arctic marine ecosystem during polar night
2013	Sebastian Wacker	ph.d Biology	The dynamics of sexual selection: effects of OSR, density and resource competition in a fish
2013	Ragnhild Pettersen	ph.d Biology	Identification of marine organisms using chemotaxonomy and hyperspectral imaging
2013	Angela Mwakatobe	ph.d Biology	Human-Wildlife Interaction in the Western Serengeti: Crop Raiding, Livestock Depredation and Bushmeat Utilisation
2013	Nina Blöcher	ph.d Biology	Biofouling in the Norwegian Salmon Farming Industry
2013	Cecilie Miljeteig	ph.d Biology	Phototaxis in Calanusfinmarchicus - light sensitivity and the influence of energy reserves and oil exposure
2013	Ane Kjersti Vie	ph.d Biology	Molecular and functional characterisation of signalling peptides of the IDA family in *Arabidopsis thaliana*
2013	Marianne Nymark	ph.d Biology	Light responses in the marine diatom *Phaeodactylum tricornutum*
2013	Jannik Schultner	ph.d Biology	*Resource Allocation under Stress - Mechanisms and Strategies in a Long-Lived Bird*
2014	Aravind Venkatesan	ph.d Biology (Semantic Systems Biology)	Application of Semantic Web Technology to establish knowledge management and discovery in the Life Sciences