The 10th Annual Bio-Ontologies Meeting

Robert Stevens, University of Manchester
Phillip Lord, Newcastle University
Robin McEntire, GlaxoSmithKline
Susanna-Assunta Sansone, EMBL-EBI

July 20, 2007
Co-located with ISMB/ECCB 2007
Vienna, Austria

Talks Programme

Start	End	Speaker	Title
09:00	09:10		**Welcome and Intro**
09:10	09:30	Moreira, D.	Experiences from the NCBO OBO to OWL Mapping Effort
09:30	09:50	Jupp, S.	Converting Ontologies to Knowledge Organisation Systems for Document Navigation
09:50	10:10	Yip, L.	Mapping proteins to disease terminologies: from UniProt to MeSH
10:10	10:30	Chabalier, J.	Integrating and querying disease and pathway ontologies: building an OWL model and using RDFS queries
10:30	11:00		**Morning Coffee**
11:00	11:20	Blake, J.	Gene Ontology Annotations: What they mean and where they come from
11:20	11:40	Arunguren, M. E.	Ontology Design Patterns for bio-ontologies
11:40	12:00	Schober, D.	Towards naming conventions for use in controlled vocabulary and ontology engineering
12:00	13:00		**LUNCH**
13:00	14:00		Poster Session
14:00	15:30		Panel Session
15:30	16:00		**Afternoon Coffee**
16:00	16:20	Soldatova, L. N.	ART: an ontology based tool for the translation of papers into Semantic Web format
16:20	16:40	Pesquita, C.	Evaluating GO-based Semantic Similarity Measures
16:40	17:00	O'Neill, K.	OntoDas - integrating DAS with ontology-based queries
17:00	17:20	Spasic, I.	Facilitating the development of controlled vocabularies for metabolomics with text mining
17:20	17:40	Bastos, H.	Using GO terms to evaluate protein clustering
17:40	18:30		Poster Session
18:30			**End of Day**

Poster Programme

Presenter	Title
Aitken, S.	The COBrA-CT Bio-Ontology Tools: Managing the transition from OBO to OWL
Antezana, E.	CCO, a paradigm for knowledge integration
Backhaus, M.	BOWIKI – a collaborative gene annotation and biomedical ontology curation framework
Bada, M.	Identification of OBO Non-alignments and Its Implications for OBO Enrichment
Baker, C.	Ontology-centric knowledge navigation for Lipidomics
Clancy, K.	Experimental Design Ontology
Conesa, A.	Development of a Reference Model for Phenotype Description
Coulet, A.	About the Role of Ontologies in Knowledge Discovery Process: Data Selection and Abstraction for Discovering Genotype-Phenotype Relationships
Day-Richter, J.	Graph-Based Editing in OBO-Edit 2.0
Deus, H.	S3QL: protocol for retrieving experimental data annotated to RDF information systems
Hoehndorf, R.	The role of non-monotonic reasoning for the interoperability of biomedical ontologies
Hoehndorf, R.	GFO-Bio: A faceted biological core ontology
Kraines, S.	A Description Logics Ontology for Biomolecular Processes
Kusnierczyk, W.	What Does a GO Annotation Mean?
Kusnierczyk, W.	The logic of Relations between the Gene Ontology and the Taxonomy of Species
Liu, H.	Building the PRotein Ontology (PRO) Prototype for TGF-beta Signaling Proteins
Merico, D.	NEUROWEB Project: from clinically-based phenotypes towards genomics integration
OBI Consortium	The Ontology for Biomedical Investigations (OBI): the evolution and maintenance of a broad, community-based biomedical ontology development project
Parmentier, G.	Using ontology alignment to add homology relationships between species anatomy ontologies
Rubin, D.	BioPortal: A Web Portal to Biomedical Ontologies and Tools
Ruebenacker, O.	Kinetic modeling using BioPAX ontology
Samwald, M	The Semantic Web Health Care and Life Sciences Interest Group work in progress: A large-scale, OBO inspired, repository of biological knowledge based on Semantic Web technologies
Schlicker, A.	The GOTax Platform: integrating protein annotation with protein families and taxonomy
Shah, N.	Which Annotation did you mean?
Splendiani, A.	Bio-Ontologies in the context of the BOOTStrep project
Whetzel, P.	Pronto – a system for ontology term curation

Panel Session

In this, the tenth annual bio-ontologies meeting, we take the opportunity to review the past ten years and look forward to the next ten years. We have invited the people who organised the first annual meeting and associated tutorial to form a panel for this meeting:

- Russ Altman
- Mark Musen
- Peter Karp
- Steffen Schulze-Kremer

We will ask these originators of the meetings to reflect on the past decade and to look forward to the next ten years. Five minute position statements will be followed by an hour of questions from the floor.

Acknowledgements

We acknowledge the assistance of Steven Leard and all at ISCB for their excellent technical assistance.

We also wish to thank the programme committee for their excellent input and guidance – thanks go to Suzanna Lewis, Chris Mungall, Alan Ruttenberg, and David Benton, with an additional thanks to Jim Butler for two key reviews in the final hours.

We thank Andrew Gibson for the creation and maintenance of our web site.

And, we wish to thank GlaxoSmithKline for their funding in support of the Bio-Ontologies SIG.

Experiences from the NCBO OBO to OWL Mapping Effort

Dilvan A. Moreira[1,2] and Mark A. Musen[1]*

[1]Stanford University, Stanford, USA, [2]University of São Paulo, São Paulo, Brazil.

ABSTRACT

There is a strong need to map the OBO format to OWL and provide tools that enable end users to easily perform this translation. To fulfill this need, the National Center for Biomedical Ontology created the NCBO OBO to OWL mapping and a set of tools to perform the translations for the Protégé and OBO-Edit editors and for command line use. A group of OBO developers and OWL experts worked cooperatively to develop this mapping. This paper summarizes our experiences and main design decisions, so users can better understand the mapping.

Contact: dilvan@gmail.com

1 INTRODUCTION

Bio-ontologies, many written using the OBO format (GO 2004), have become increasingly popular to annotate data from high throughput biological experiments. For that reason, bio-ontologies have grown in size as well as complexity and are becoming the focus of attention of the larger computer science research community.

This larger community has a significant interest in using the life sciences domain as a "focus" for W3C semantic web activity. In this light, biological data annotated using OBO bio-ontologies is a prime resource that can be harvested if access to the ontologies and annotated data is available using an internationally recommended standard.

On the other hand, if biology is to benefit from the rapid progress being made in computer science – especially the support for semantic web ontologies – bio-ontologies need to interoperate with those ontologies which are in the Web Ontology Language - OWL (W3C 2004). As a result, there is a strong need to map the OBO format to OWL and provide tools that enable end users to easily perform this translation, without worrying about underlying formats. To fulfill this need, the National Center for Biomedical Ontology (NCBO) created its OBO to OWL mapping (NCBO 2007) and a set of tools to perform the translations for the Protégé and OBO-Edit editors and for command line use.

A group of OBO developers and OWL experts worked cooperatively to develop this mapping. This paper summarizes their experiences and main design decisions, so users can better understand the mapping and developers can use this knowledge when planning similar efforts.

2 KNOWN LIMITATIONS OF THE OBO FORMAT

The limitations of the OBO Format can be divided in three main areas:

Lack of computational definitions: Ontologies in the OBO format lack computational definitions to determine the meaning of a term (Mungall 2004), which presents problems for tools such as automated reasoners. This lack of definitions leaves the task of maintaining ontology integrity entirely on the ontology developers. This led to the use of *Style Guides* to enforce constraints that should ideally be part of the ontology representation language itself. Aranguren et al (2007) discuss in detail the semantics derived from the Directed Acyclic Graph (DAG) representation implied by the OBO format, using GO as an example. They make the point that it is necessary to rely on text definitions (such as the GO Editorial Style Guide) to accurately interpret OBO relationships such as *part_of*.

Under specified format: Another problem with the OBO format is that it is under specified. It lacks formal syntax or semantic definitions, creating inconsistencies when parsing the format, which are dealt with by undocumented default behavior, such as:

- built-in relationships, such as *is_a*, must not be modified, but the specs allow users to redefine them in an ontology file. OBO parsers silently ignore any redefinition code, but this behavior is not documented.
- The format allows a file to declare the exactly same synonym, as many times as a user wants. However, that is not flagged as a parser error, as the specs do not address the situation. Again the parsers deal with it using implicit undocumented behavior.

Limited scope: It is common to associate the OBO format to the Gene Ontology (GO) project (Ashburner et al. 2000), as the format is an offspring of this project. The goal of the Gene Ontology project is to provide a controlled vocabulary to describe gene and gene product attributes in any organism. The format was created to develop controlled vocabularies and then incrementally evolved to attend the needs of the GO project that kept its scope rather limited. Since GO is not necessarily tied to any particular file for-

* To whom correspondence should be addressed.

mat, it could potentially benefit enormously from using a more expressive format, such as OWL.

OWL does not suffer from any of these three limitations of OBO. It allows the creation of computational definitions using a knowledge representation language based on Description Logic. It is a W3C standard with formal documents describing it. And it has being created for the wider Semantic Web community.

3 MAPPING FROM AN UNDERSPECIFIED FORMAT

As OBO is an underspecified format, there is no document completely describing its syntax or semantics (Horrocks 2007). As the format is loosely specified and it is not expressed as a context-free grammar using Bacus-Naur Form (BNF), it is very difficult to write parsers that will interpret OBO syntax as intended by the developers of OBO. To overcome this problem, the parser built into OBO-Edit (former DAG-edit) was used, as it guarantees that the files will be always parsed correctly because the creators of the parser are also authors of the OBO format. OBO-edit is free software, so its parser can be openly reused. There are currently efforts to write a BNF grammar to describe OBO (Horrocks 2007), but such a grammar has yet to become part of the format. The mapping problem is now confined to establishing semantic correspondences between OBO and OWL constructs.

To overcome the lack of OBO specs, groups creating ontologies using the format have resorted to write Style Guides, such as the GO Style Guide, that clarify (and sometimes, expand) the semantics of OBO constructs used in their ontologies. Also, not all constructs available are used: some of the newer tags never appeared in the sample of OBO ontologies reviewed. So in parallel to the official OBO format, there is a "de facto" format being used.

It was decided that the OBO to OWL mapping would remain faithful to the declared semantics of the OBO format rather than create an intended mapping for a particular ontology to OWL, which would be the case if we incorporate information from a Style Guide. However, when a particular feature was part of the "de facto" format, meaning that it has been adopted by most OBO ontologies, this feature was included in the mapping and eventually added to the official OBO spec. A good example is the use of the format `<ontology id space> : <numeric id>` for OBO term identifiers (i.e. `GO:0003455`), most OBO ontologies use this format, including the GO ontology, but that is not a requirement of the OBO specs.

This strategy meant that the resulting mapping would work with the vast majority of OBO ontologies and that, eventually, the OBO format will include the features from the "de facto" format. We declared the following requirements for the mapping between OBO and OWL:

- The mapping should map all OBO format constructs in use.
- It should not make more assumptions than what is written in the OBO format specification itself, so it can be used for any OBO file.
- It should include features used by most ontologies, as part of the "de facto" OBO format. Actually those feature, although important, formed a very small set.
- It should be lossless, so information will not be lost in the conversion process.
- It should also do the "round trip": generate an OWL file from an OBO file and then regenerate the OBO file back from OWL (assuming no OWL specific edits are made) without losing information.

A mapping like that had to be a collaborative and interactive process involving OBO core developers and OWL experts. The main goals of the mapping were to be able to extract the concepts coded in OBO, encode them in OWL, and ensure that information was not lost during this process. Each one of these goals represents a step in the interactive process of map creation: a set of OBO constructs is interpreted; the information is mapped into OWL; and them transformed back to OBO, to compare to the original set of constructs (to check for loss of information).

For this process to be successful, the last task had to be automated. It was important to know not only what each change meant for a particular ontology, but for a large representative set of OBO ontology (specially so to find actual "de facto" features). So first we identified the OBO ontologies at the OBO Foundry site (http://obofoundry.org) as our large set of representative OBO ontologies. Then a tool was developed, based on the OBO to OWL plugin for Protégé, to read every OBO ontology on this site (every ontology file ending with .obo), translate each into OWL, using the current mapping being tested, translate them back into OBO, and compare the generated OBO file with the original, using the OBO-Diff tool from OBO-Edit (version 1.02). If every ontology passed the test, the current mapping would be considered OK. If some ontologies fail, we would examine them to find if:

- The mapping was doing something wrong.
- The ontologies had a construct not yet mapped.
- They had some "no standard" construct syntax.
- They were not following the OBO format (they had some error).

That "test often" approach gave us the freedom to test many possible mapping options because it allowed us to have an instantaneous assessment of how those options impacted the mapping. That was particularly useful when try-

ing to separate "de facto" features from options that affected just one or two ontologies.

4 OBO TO OWL MAPPING

We distinguish between mapping the lexical components and mapping the semantics, which we explain below.

Traditional medical terminologies and biological ontologies contain both semantic and lexical information. Semantic information provides intrinsic characteristics of classes and instances; it is inherited through subclassing and can be processed by logic-based reasoning systems to determine consistency and completeness. Lexical information provides the intended meaning or interpretation of a class or instances to an end user; it includes names, textual definitions, descriptions, usage notes, etc (Supekar et al 2005).

The OWL language has strong support for representing semantic information but only minimal support for representing lexical (non-semantic) information. OWL provides a particular type of property, `owl:annotationProperty`, to represent non-semantic information: properties of this type and their values are ignored by DL reasoners, so they can have only non-semantic data (Note that in OWL-DL one cannot define subproperties, declare the type or declare domain/range constraints for annotation properties.)

4.1 Lexical mapping

To represent the lexical information present in OBO format, we created a standard set of OWL classes and properties. To determine this set, we used the following five core types of lexical information, generalized and formalized from the study of thirteen public available biomedical terminologies by Supekar et al (2005):

- **Text representations:** text (and markup) that can be used to declare the intended meaning of the concept in a given language, setting or context. They can also be referred to as "terms", "synonyms", "labels", "designations" or "presentations" depending upon the terminology.
- **Definitions:** blocks of text (and markup) that define the intended purpose and meaning of a concept in a given language, setting or context.
- **Usage notes:** text that is intended to inform or instruct users about additional conditions, etc. that need to be considered.
- **Editorial notes:** text that is intended for communication with the authors.
- **Machine instructions:** represent (semi-) formal information that is to be algorithmically or machine processed.

These types were used to help to identify the lexical constructs in the OBO format that had to represent in OWL. These lexical constructs could be represented either by simple binary relations, such as comments (`hasVersion – xsd:string`), or by more complicated n-ary relations ones, such as synonyms (`hasSynonym – Synonym individual`).

The complex relations are mapped using a design pattern where a new class and n new properties are created to represent an n-ary relation. This pattern is recommended in a W3C Working Group Note (http://www.w3.org/TR/swbp-n-aryRelations/). An instance of the relation linking the n individuals is then an instance of this class. Ontologically the classes created in this way are often called "reified relations" and play important roles in many ontologies (e.g. Ontoclean/DOLCE, Sowa, GALEN). In this mapping, those classes are only used in non-semantic constructs, having no impact from a DL point of view.

The set of these new classes and properties – created to represent lexical information from the OBO format in OWL – can be viewed as a sublanguage to represent biological ontologies in OWL. This "OBO in OWL" sublanguage can be seen as the OBO format implemented in a W3C recommended standard. Users of the OBO format can leverage the OWL standard and its related tools without having to give up the freedom of a representation/language that meets their needs and can be changed in an agile manner as those needs changes. To clearly separate the entities used by this sublanguage from the ones used by a particular ontology mapped in OWL, we use the `oboInOwl:` OWL namespace (`http://www.geneontology.org/formats/oboInOwl#`).

4.2 Semantic mapping

To accomplish the semantic mapping we need to map the basic two sets of OBO constructs to OWL:

- OBO terms (`[Term]` stanza) are mapped to OWL classes (`owl:Class`).
- OBO relationship types (`[Typedef]` stanza) are mapped to OWL object properties (`owl:ObjectProperty`). With the exception of `is_a` relations that are represented as subclass relationships in OWL.

The next step is to create a naming scheme, i.e. the steps to compose an OWL id from an OBO id. We require that the id of an OWL class or property (`rdf:ID`) be unique. We follow the design rational (http://protege.cim3.net/cgi-bin/wiki.pl?HidingIdentifiersWithLabels) that an identifier should be used just to identify a concept or relationship without interfering with the name(s) associated with it so that:

- fixing a typo in a name does not make the former concept obsolete
- the text representation of a class or of a relationship can be retired and the same name can be allocated to a new class

- the identifiers are agnostic as to one hard-coded preferred language
- synonyms can be handled more easily.

In OBO, only term ids are guaranteed to be unique but they are usually not meaningful to the user (such as `CL:0000339`) and the colon character is not allowed unencoded in `rdf:ID`s. So we decided to derive OWL ids from OBO ids changing the character ':' to '_'. It was also decided that the OBO format specification should get stricter on term id names, only allowing the format `<ontology id space> : <numeric id>` (this change has since been incorporated into the format). In the case of relationships (`Typedef`) with OBO ids without colon, they are considered as having an undefined OBO id space, and get the `UNDEFINED` prefix in OWL. See Table 1 for examples. More meaningful OBO names (such as "`glioblast (sensu Vertebrata)`") are used as labels (`rdfs:label`).

The OBO relationship ontology (`relationship.obo`) comes embedded in the OBO-Edit tool. It is a very important ontology that defines the basic relationships used by OBO. In the mapping, this ontology is always explicitly included, so the users have a greater incentive to use the defined relations whenever possible and not invent new ones. The OWL prefix `oboRel:` is used to refer to these relations.

Table 1. Mapping OBO ids into OWL `rdf:ID`s, the first line is a term id, the others are relationship ids.

OBO id	OWL rdf:ID
`CL:0000339`	`CL_0000339`
`develops_from`	`UNDEFINED_develops_from`
`CL:develops_from`	`CL_develops_from`
`OBO_REL:develops_from`	`oboRel:develops_from`

5 CONCLUSION

The main goal of this paper was to show the experiences (problems and solutions) faced by the group of people developing the NCBO OBO to OWL mapping, so that other groups facing similar tasks can learn from our experience.

As OWL and other Semantic Web technologies become more mature and mainstream, we will see more efforts trying to map ontologies written in other languages to OWL. So we would like to highlight the main messages of the paper:

OBO format had many shortcomings that the use of OWL would correct.

Working with an underspecified format means that different users interpret it in different way. So we had to find the "de facto" standard.

Automatic testing of the mapping speeds up the mapping process while maintaining quality.

OWL 1.0 does not have good support for lexical representation. A set of classes and properties had to be created.

To those points we would like to add that mapping is a cooperative effort, especially when it involves an underspecified format. The quality and usefulness of this mapping would be very limited if we did not have the support of the OBO developers/community. They made sure that we understood the meaning of OBO constructs and made changes to the format when necessary. To engage them we have to demonstrate the benefits of using OWL over OBO, but with the goal of complementing not substituting OBO. To solve disagreements, during the creation of the mapping, it was important to have software to automatic check the mapping against a large set of ontologies. That allowed us to try different mapping options and choose the best options based in more objective data.

Finally, the question if OWL, or more precisely the OBO in OWL representation, is going to substitute OBO or not is better left to the end users to decide not the mapping team.

ACKNOWLEDGEMENTS

We would like to acknowledge the efforts of the whole NCBO team in making this mapping possible, in special to Chris Mungall, Nigam Shah, John-Day Richter and Suzanna Lewis. This effort was funded by NIH grant U54 HG004028 and a grant from CAPES-Brazil.

REFERENCES

Ashburner, M. et al (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics,* 25, 25-29.

Aranguren M., Bechoffer S., Lord P., Sattler U. and Stevens R. (2007) Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. *BMC Bioinformatics* 2007, 8:57.

GO (2004) The OBO Flat File Format Specification, version 1.2 http://www.geneontology.org/GO.format.obo-1_2.shtml

Horrocks I. (2007) *OBO Flat File Format Syntax and Semantics and Mapping to OWL Web Ontology Language*, http://www.cs.man.ac.uk/~horrocks/obo/syntax.html

Mungall, C.J. (2004) Obol: integrating language and meaning in bio-ontologies, *Comp Funct Genom,* 5, 509-520.

NCBO (2007) *OBO in OWL: Mapping and Tools,* February 12 http://www.bioontology.org/wiki/index.php?title=OboInOwl:Main_Page

Supekar K., Chute C. and Solbrig H. (2005) Representing Lexical Components of Medical Terminologies in OWL, AMIA Annu Symp Proc. 2005 :719-23.

W3C (2004) OWL Web Ontology Language Reference, W3C Recommendation 10 February 2004, http://www.w3.org/TR/owl-ref /

Converting Ontologies to Knowledge Organisation Systems for Document Navigation

Simon Jupp[*1], *Sean Bechhofer*[1], *Patty Kostkova* [2], *Robert Stevens*[1], *Yeliz Yesilada*[1]
[1] *University of Manchester, Oxford Road, UK*
[2] *City eHealth Research Centre, City University, London, UK*

ABSTRACT

Motivation: Bioinformatics relies heavily on web resources for information gathering. Ontologies are being developed to fill the background knowledge needed to drive Semantic Web applications. This paper discusses how ontologies are not always suited for document navigation on the web. Converting ontologies into a model with looser semantics allows cheap and rapid generation of useful knowledge systems. The message is that ontologies are not the only knowledge artefact needed; vocabularies and other classification schemes with weaker semantics have their role and are the best solution in certain circumstances.

1 INTRODUCTION

Navigation via hypertext is a mainstay of the World Wide Web (WWW). The author owned and unary links of standard HTML often do not offer either the links sources or targets needed by a particular group. Conceptual hypermedia provides navigation between web resources, supported by a conceptual model. The content of the model is used to dynamically identify link sources in web documents, and also supply the link targets to relevant web-services. The field of bioinformatics relies heavily on web resources and the community is now rich in bio-medical ontologies that can be used to populate this conceptual model.

The ability to browse documents on the web via hyperlinks embedded in text is still a fundamental part of the information gathering process used by bioinformaticians. As successful as hypertext is, it is not without its limitations;

- **Hard Coding**: Links are hard coded into the HTML source of a document.
- **Ownership**: Ownership of the page is required to place links in pages.
- **Legacy**: Link target can be deprecated leaving invalid links on pages.
- **Unary targets**: The current web links are restricted to point-to-point linking; there is only one target.

Conceptual Open Hypermedia supports the construction of hypertext link structures built using information encoded in ontologies. Dynamic linking, supported by ontologies, offer a mechanism to help overcome some of these restrictions. The Conceptual Open Hypermedia Service (COHSE)[1] (Carr 2001) system enhances document resources through the addition of hypertext links (see Figure 1). These links are generated based on a mapping between concepts found in the document and lexicons available from the ontology. Links can have multiple targets based on the type of concept identified, in addition the structure of the ontology facilitates navigation to further targets based on sub/super concepts asserted in the ontology.

The COHSE architecture has been demonstrated in several fields, the GOHSE (Bechhofer 2005) system was applied to bioinformatics using the Gene Ontology (GO)(Gene Ontology Consortium. 2000) as an ontology and GO associations as link targets. The Sealife project [2] is now looking to extend this work and provide an ontology that integrates many of the ontologies being developed in biomedicine, to aid query by navigation to both scientists and health care professionals in the study of infectious diseases.

One of the major obstacles at this stage is how to integrate all the ontologies into a single model with appropriate semantics that suit navigation. We argue that the strict relationships held between concepts in ontologies are not well suited for navigational purposes. A thesaurus like artefact is better suited for this task, it allows us to capture relationships that are not formal or universal or part of the integral definition of the term. Our goal is to benefit from the work being done in the bio-ontology community i.e. capturing specific domain knowledge, and bring this knowledge into a model that suits the applications needs.

The proposed solution is to convert relevant bio-ontologies, medical vocabularies, thesauri, taxonomies and other concept schemes into one large Knowledge Organisation System (KOS). The Simple Knowledge Organisation System (SKOS) [3] is chosen as a model to hold this information. A use case from the Sealife project is used to demonstrate the application in the study of infectious disease.

* To whom correspondence should be addressed.

[1] http://cohse.cs.manchester.ac.uk/
[2] http://www.biotec.tu-dresden.de/sealife/
[3] http://www.w3.org/2004/02/skos/

2 SEALIFE USE CASE

The Sealife project seeks to develop a series of browsers in the context of the Semantic Web and Semantic Grid. The grid offers an infrastructure for large scale *in silico* science via a large number of computational services. The Grid setting needs to be combined with the continuing presence and use of numbers of Web documents describing knowledge about biology. Ontologies and controlled vocabularies provide great benefits for describing and using their data. The Sealife browser aims to use these vocabularies and ontologies as description of knowledge in the life sciences to flexibly manage the inter-linking of these documents and services.

One example application is to provide dynamic hyper-linking of resources from the National electronic Library of Infection (NeLI) [4] (Kostkova 2003) portal to other related resources on the web. NeLI is a digital library bringing together the best available on-line evidence-based, quality tagged resources on the investigation, treatment, prevention and control of infectious disease. Many documents on the NeLI site contain few, if any, hyperlinks to other resources on the web. It would take a large curational effort and cost to manually mark up these pages with links to other web resources. In addition to this problem, NeLI has a range of users; we want different link targets based on the kind of user browsing the NeLI site. COHSE can help to solve some of these problems, terms from the ontologies can be used to identify concepts in web pages and create the link sources. Each link source can have multiple targets; the targets selected are tailored to suit the needs of each user group.

The ability to identify user groups is important. Users can range from members of the public, molecular biologists to clinicians and GPs. Each group has a different view of the bio-medical domain, and is therefore interested in different kinds of information. By providing alternate vocabularies for different users, the system can identify link sources relevant to that user and also provide multiple targets to relevant web resources. Table 1 shows four different user groups, some questions they might want answering and the different kinds of target sites a Sealife browser would offer them based on the type of user (Madle 2006).

The system is demonstrated with a simple use case involving a news site linking to NeLI. News sites are often the first to report on disease outbreaks via news feeds. Consider the scenario where a traveler is planning a trip to Namibia, only to find an article on the BBC website about a recent outbreak of Polio. COHSE can provide links to relevant resources that had not been included by the original author. Such resources could include information about the poliovirus, its effect on humans, vaccination information and also geographical information about the local area. A family doctor, in contrast, might use a vocabulary skewed to the their interests to link through to sites on drugs, details of symptoms and clinical presentations, treatment and local hospital facilities etc.

User Group	Question	Targets
Family Doctor (GP)	Tuberculosis drugs and side effects?	British National Formulary (BNF)
Clinicians	Tuberculosis treatments guidelines?	Public Health Observatories (PHO)
Molecular Biologists	Drug resistant tuberculosis species?	PubMed
General Public	What is tuberculosis?	Health Protection Agency (HPA) or the NHS direct online website.

Table 1. NeLI users and example targets from the UK.

Figure 1 shows the system in action. The first image shows the original BBC article, the second shows dynamic links that have been added based on concepts held in the ontology. It also shows a link box that is dynamically generated when a link is clicked. The link box contains a textual description of the term and targets to multiple web resources. In addition to targets for the selected link the system can provide targets for broader, narrower and related resources. For example, NeLI has a web-service which takes terms from a NeLI vocabulary as inputs, this service is invoked when the polio link is selected and targets are returned which link to relevant documents from the NeLI portal. This simple demonstration shows how the addition of a navigational layer based on the semantic content of documents can be added to the existing web.

[4] http://www.neli.org.uk/

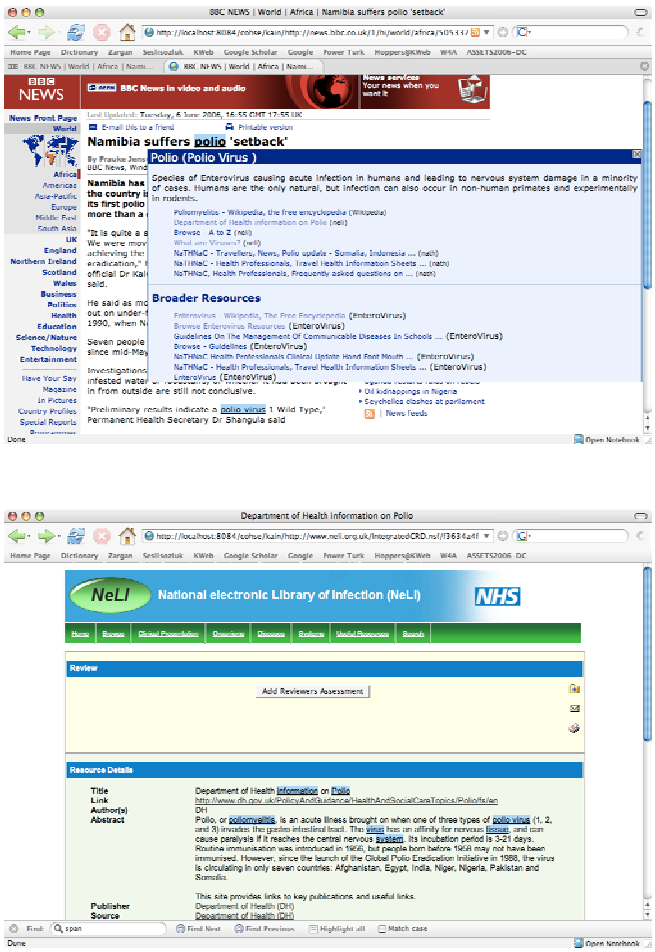


Fig. 1. Dynamic linking in action.

3 GATHERING THE BACKGROUND KNOWLEDGE

For the Sealife browser to be useful across such a diverse subject as biology we need a system to rapidly collect the current available resources together, and place them into a single representation that will facilitate navigation. The bio-medical domain already has a rich collection of vocabularies and ontologies such as MesH[5], UMLS[6], GALEN[7] and the OBO[8] ontologies. There are also classification systems relating to genes, protein, drug and other terminological resources that would be useful to Sealife.

The languages used to represent ontologies vary considerably, and can range from simple taxonomy languages through to rich, formal logic based languages such as OWL. Increasingly strict semantics can remove ambiguity in the representation and facilitate the use of machine processing. Similarly, these languages can be used with varying degrees of ontological formality, not all OWL ontologies make rigorous ontological distinctions. Experience with COHSE has suggested that formal ontological distinctions and strict semantics are not *always* best suited to the task of navigating a collection of resources. Strict sub/super class relationships are not necessarily appropriate for navigation – rather, the looser notions of broader/narrower as found in vocabularies or thesauri provide the user with more appropriate linking.

SKOS is a model for representing classification systems, thesauri, taxonomies and other concept schemes. SKOS is currently undergoing stanardisation by the W3C[9] and has a RDF/XML representation that makes it well suited for semantic web applications. By representing the biological knowledge in SKOS we have a simple model that provides a lexical resource for identifying concepts in our documents, as well as a framework for asserting semantic relationships between concepts. SKOS has a set of properties that are well suited for supporting navigation. These include preferred labels, alternate labels (synonyms) and textual definitions for describing concepts as well as '*broader*', '*narrower*' and '*related*', for representing the relationships between concepts.

4 CONVERTING ONTOLOGIES TO SKOS

The semantics of some biomedical terminologies are already relatively weak. A good example for such a terminology that is commonly used in medicine is the Medical Subject Headings (MeSH). The semantics of **A** *narrower* **B** simply means that users interested in **B** might also be interested in **A**. The MeSH terms found under accident include kinds of accidents – as expected (e.g. Traffic accidents), but also Accident prevention. This is not a good ontological distinction, but a valid one in the context of navigation and retrieval. In contrast in the Open Biomedical Ontologies (OBO), **A** *is-a* **B**, a common type of relationship in OBO ontologies implies that all **A**'s are also instances of **B**. This contrast in semantics means that conversions from MeSH into OBO are not possible without misinterpreting the intended semantics. Despite this, we see that many of the OBO ontologies share concepts with MeSH, especially the Disease Ontology [10]. From a navigation point of view we would like to combine these resources to gain maximum benefit from efforts in MeSH and OBO development. By converting them both into SKOS we can use a single representation and use the lightweight semantics to build a larger and richer vocabulary.

The release of the 10 OBO relations (Smith 2005) gives OBO developers another level of expressivity in their ontologies. These relations have logical definitions with precise semantics; which are used to define relationships

[5] http://www.nlm.nih.gov/mesh/

[6] http://www.nlm.nih.gov/research/umls/

[7] http://www.opengalen.org/

[8] http://obofoundry.org/

[9] http://www.w3.org/

[10] http://diseaseontology.sourceforge.net

between terms in OBO ontologies. When converting OBO ontologies into SKOS we can use these relationships to assert *broader*, *narrower* and *related* relationships between SKOS concepts. Here is an example of the conversion one might make when mapping ontological properties to SKOS properties.

- rel:part_of -> sub-property of skos:narrower (e.g. finger part_of hand)
- rel:contains -> sub-propert of skos:broader (e.g. skull contains brain)
- rel:has_name -> sub-property of skos:related (e.g. Person has_name PersonName)

Another advantage when converting properties from ontologies to SKOS is the ability to assert the inverse. Consider an ontology where **Nucleus** *partOf* **cell**, from an ontological point of view this implies that every **Nucleus** is *partOf* some **Cell**. However, the inverse is not true, every **Cell** does not *havePart* N**ucleus**. When converting to a SKOS model we can assert the inverse using the *narrower* property to say that **Nucleus** is a *narrower* term than **Cell**, which is quite reasonable. When navigating around documents about cells, the system could then also provide links to documents about nuclei – users interested in cells are often also interested in nuclei.

If we use the polio use case example we can show that a great deal of information can be acquired about polio from the various vocabularies alone. When the semantics are strict we have to be very careful how we bring all this related information together. With all this information in SKOS, Sealife can benefit from many different knowledge resources. Table 2 outlines the results from searching polio against a varying set of ontologies and vocabularies alongside the SKOS property used to relate them.

Source	Terms found	SKOS relation to "Poliovirus"
MeSH	Poliomyelitis	skos:altTerm
Disease Ontology	Spinal cord disease	skos:broaderThan
	Postpoliomyelitis Syndrome	skos:narrowerThan
SNOMED	Microorganism	skos:broaderThan
	Enterovirus	skos:broaderThan

Table 2. Searching poliovirus against different resources and converting intended semantics into SKOS semantics for navigation.

There is, however, likely to be some trade off associated when bringing multiple resources together, it is possible that a lot of unwanted terms are returned, especially when using formal ontologies. Ontologies can benefit from an upper-ontology (Rector 2003) that contains abstract categories; these can be used to build formal definitions for classes. A ontological definition which states whether the concept is a physical or non-physical entity may be crucial to the design of a robust ontology, but is largely irrelevant from a navigational point of view. To overcome this we must remove these properties at the stage of conversion into SKOS, how we do the conversion from ontologies to SKOS is something for future discussion.

CONCLUSION

A large community of ontology developers and knowledge engineers is forming in the life sciences. It is hoped that they will deliver the infrastructure needed to realise a real semantic-web, where computers can begin to interpret and interoperate biological data automatically. If applications like Sealife are to demonstrate the early potential of Semantic Web technologies, then the trade off associated with relaxing the semantics in the background knowledge has to be acceptable.

The nature of formal ontologies can sometimes make it difficult to express relationships between concepts that experts from the domain would expect to find under some circumstances. Thesauri are much more suited to represent the way *words* and *language* is used in the field. The Sealife project will demonstrate how the effort and cost associated with building rich formal ontologies can also be used to feed into other knowledge artefacts, like SKOS, which can then be used in different application scenarios.

ACKNOWLEDGEMENTS

Funding by the Sealife project (IST-2006-027269) for Simon Jupp is kindly acknowledged.

REFERENCES

Gene Ontology consortium. (2000). "Gene Ontology: tool for the unification of biology." Nat Genet **25**: 25 - 29.

Carr, L. Bechhofer, S. Goble, C. Hall, W. (2001) *Conceptual Linking; Ontology-based Open Hypermedia.* WWW10, Tenth World Wide Web Conference, Hong Kong.

Bechhofer, S. Stevens, R. Lord, P. (2005) *Ontology Driven Dynamix Linking of Biology Resources.* Pacific Symposium on Biocomputing, Hawaii.

Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C *Relations in Biomedical Ontologies***.** Genome *Biology, 2005, 6:R46*

P. Kostkova et al. (2003) *Agent-Based Up-to-date Data Management in National electronic Library for Communicable Disease.* In SI: "Applications of intelligent agents in health care", J. Nealon, T. Moreno Ed, in Whitestein Series in Software Agent Technologies, pages 103-122.

Rector, A. (2003) *Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL*. Knowledge Capture, ACM, 121-128.

G. Madle, P. Kostkova, J. Mani-Saada, A. Roy. (2006) *Lessons learned from Evaluation of the Use of the National electronic Library of Infection*, Health Informatics Journal, Special Issue, Healthcare Digital Libraries", 12: 137-151

Mapping proteins to disease terminologies: from UniProt to MeSH

Yum Lina Yip[1,2], Anaïs Mottaz[1], Patrick Ruch[2,3], and Anne-Lise Veuthey[,1]*

[1]Swiss Institute of Bioinformatics, [2] University of Geneva, and [3] Medical Informatics Service, HUG, Geneva, Switzerland

ABSTRACT

Although the UniProt KnowledgeBase is not a medical-oriented database, it contains information on more than 2'000 human proteins involved in pathologies. In order to make these data easily accessible to clinical researchers, we have developed a procedure to link diseases described in the UniProtKB/Swiss-Prot entries to the MeSH disease terminology. For this purpose, we mapped disease names extracted either from the entry comment lines or from the corresponding OMIM entry to the MeSH. The method was assessed on a benchmark set, and was shown to have a 100% precision for a recall of 37%. Using the same procedure, the nearly 3'000 diseases in Swiss-Prot were mapped to MeSH with comparable efficiency.

1 INTRODUCTION

Biomedical data available to researchers and clinicians have increased drastically over the last decade because of the exponential growth of knowledge in molecular biology. While this has led to the creation of numerous databases and information resources, the interoperability between the resources remains poor to date. One of the main problems lies in the fact that medical terminologies are scarcely used in molecular biology. Take the example of the UniProt Knowledgebase (UniProtKB), the most comprehensive protein warehouse with extensive cross-references to other database resources (The UniProt consortium, 2007). In UniProtKB, about 2'000 human proteins contain manually curated information related to their involvement in pathologies. While this information is clearly of value, it is not easily accessible for clinical researchers due to the fact that UniProtKB does not use standard medical vocabularies to describe diseases associated to proteins and their variants.

Clearly, in order to increase the interoperability between the biomolecular and clinical resources, one of the key solutions lies in the development or unification of common terminologies capable of acting as a metadata layer to provide the missing links between the various resources. In the medical/clinical domain, there have already been numerous and successful efforts to implement controlled vocabularies for pathologies. Most of these resources are collected and organised into concepts in the UMLS, a major repository of biomedical standard terminologies. The recent integration of the Gene Ontology in the UMLS has further opened new ways of linking biological and medical resources via terminologies. Therefore, terminology and ontology mapping has become an active field of research. The National Library of Medicine (NLM) did an important pioneer effort through the integration of more than 60 medical vocabularies in the UMLS Metathesaurus and the development of lexical tools for this purpose. In parallel, many approaches have been developed which integrate lexical-based, as well as knowledge- and semantic-based methods to map different terminologies (Sarkar et al., 2003, Cantor et al., 2003, Zhang et al., 2007, Lussier et al. 2004, Johnson et al., 2006).

In this study, we have developed an automatic approach to map the disease terms in UniProtKB to MeSH - the controlled vocabulary thesaurus used for biomedical and health-related documents indexing (Nelson et al., 2004). The MeSH thesaurus is structured in a hierarchy of descriptors. Each descriptor includes a set of concepts, and each concept itself contains a set of terms, which are synonyms and lexical variants.

The mapping procedure described below took advantage of the manual annotation in UniProtKB as well as the curated links of UniProtKB entries to OMIM, a comprehensive knowledge base of human genes and genetic diseases (Hamosh et al. 2005). A benchmark set was created for the refinement of term matching algorithm as well as for the definition of matching score and score threshold. This work provides a basis for further work aiming to increase the interoperability between data resources from the medical informatics and the bioinformatics domains.

2 METHODS

Extraction of disease names

In UniProtKB/Swiss-Prot, disease information related to a protein entry is expressed in free text comment lines (category 'Disease'). We proceeded by first manually establishing a list of regular expressions that indicated the presence of disease names within these lines (e.g. 'cause(s)', 'cause of', 'involved in', 'contribute(s) to'). The extraction of complete disease names was relatively easy as they are usually located at the end of a sentence or directly followed by a corresponding OMIM identifier. In parallel, we took advantage of the citations to OMIM phenotypes (#) and genes with phenotypes (+) in the disease comment lines to extract the fields "Title" and "Alternative titles; symbols" from the

* To whom correspondence should be addressed.

corresponding OMIM entries. These two fields provide the disease names in OMIM as well as a set of synonyms.
The UniProtKB/Swiss-Prot release 51.0 and the OMIM version Sept. 2006 were used for this study.

Mapping procedure
We mapped the extracted disease names to the terms from the disease category of the MeSH terminology (version 2006). The complete procedure was summarised in Fig. 1. It consisted of two successive term matching steps:

(1) finding an exact match, where all words composing the name had an identical correspondent in a MeSH term and vice versa. A match was considered as exact if it respected the word order, but not necessarily the case.
(2) when the previous step failed, we looked for partial matches by decomposing the name into its word components and calculate a similarity score with MeSH terms having at least one word in common.

Similarity score
The similarity score is a function of the number of words in common minus the number of words that differ. In order to take into account the informative content of words, we weighted them with an adaptation of the weighting schema, 'Term Frequency X Inverse Document Frequency' (TF X IDF) commonly used in information retrieval techniques (Shatkay, 2005). We calculated the inverse document frequency (IDF) of each word present in the three sources of terms, namely Swiss-Prot disease lines, OMIM Titles and Alternative titles, and disease MeSH terms. The matching score was calculated according to the following formula:

$$S = \frac{\sum_{cw} Log(\frac{1}{freq(cw)}) - \sum_{ncw} Log(\frac{1}{freq(ncw)})}{size(disease)}$$

Where *freq=n/N*, with n the number of occurrence of the word in all OMIM (Titles, Alternative titles), MeSH terms (disease category) and Swiss-Prot disease comment lines, and N the total number of words in these documents. *cw* and *ncw* stand for words in common and not in common, respectively, between the two mapped terms, and *size(disease)* is a normalization factor consisting of the number of words composing the disease name to be mapped.
Hyphenated words were treated in a special way in that each of their components was considered as distinct words. If all components have a matched equivalent, their respective weights were summed up in the score calculation. Otherwise, their weights were subtracted.

Mapping evaluation
In order to evaluate the mapping procedure, 92 disease comments (with 82 references to OMIM) from 43 UniProtKB/Swiss-Prot entries were manually mapped to MeSH by a medical expert. Swiss-Prot entries were selected randomly. However, care was taken so that the chosen sample of entries would be representative and lead to a proportion of exact and partial matches similar to that found in a preliminary mapping attempt.

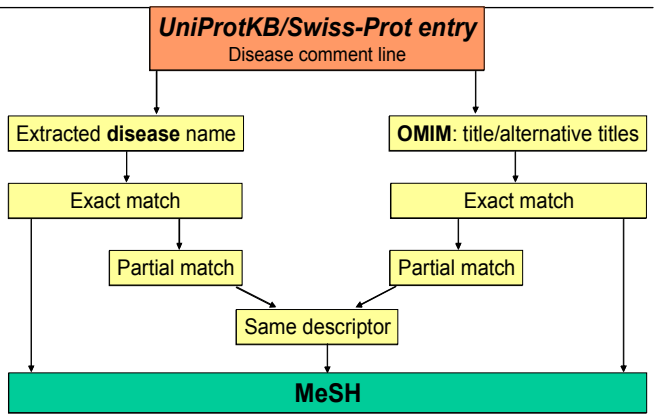


Fig. 1. Procedure of the mapping of Swiss-Prot disease comment lines to MeSH terms.

The mapping procedure was assessed in terms of *precision=TP/(TP+FP)* and *recall=TP/total number of terms*, where *TP* is the number of correct mappings (true positives) and *FP* is the number of incorrect mappings (false positives).

3 RESULTS

3.1 Disease name extraction

In UniProtKB/Swiss-Prot (rel. 51.0), 2'033 human protein entries contained information on the involvement of these proteins in diseases. This corresponded to a total of 2'966 diseases, mainly of genetic causes. The disease names were extracted from the comment lines with a set of regular expressions. The extraction failed in only 7 comment lines where a clear reference to a disease was not expressed, for instance:

"(CBL) can be converted to an oncogenic protein by deletions or mutations that disturb its ability to down-regulate RTKs." (P22681)

By manual assessing the extraction results, we noticed that as the system was constructed to extract only a single disease name per line, it was unable to treat lines such as:

"KRT16 and KRT17 are coexpressed only in pathological situations such as metaplasias and carcinomas of the uterine cervix and in psoriasis vulgaris." (P08779)

We did not investigate further these cases, as the structure of disease lines is planned for a revision in the framework of Swiss-Prot comment standardization efforts.

Among the 2'966 diseases, about 73 % (2'179) had a link to a corresponding phenotype described in OMIM. Extraction of OMIM's disease names from "Title" and "Alternative title; symbols" was straightforward. We kept all words composing a term, even qualifiers such as "included".

3.2 Automatic mapping on the benchmark

We mapped the extracted disease names to the 38'193 terms of the MeSH disease category using two successive procedures. First, we checked for exact matches with MeSH terms. Tested on the benchmark set, this procedure was able to map 16 and 21 diseases from the Swiss-Prot and the OMIM sets respectively, with a precision of 100% (Table 1). As there was an overlap between Swiss-Prot and OMIM terms, the total number of exact matches from the two mappings covered 27 diseases of the benchmark; thus corresponding to a recall of 29%. Overlap of disease mapping did not necessarily mean that the matching terms were the same, but rather that they belonged to the same descriptor in the MeSH terminology.

Table 1. Evaluation of the mapping of the Swiss-Prot disease lines on MeSH terms (92 diseases with 86 references to OMIM). In bold are the combined results which give the highest precision.

	Exact match		Partial match		Total	
	Recall	*Prec.*	*Recall*	*Prec.*	*Recall*	*Prec.*
SP	17%	100%	17%	80%	35%	89%
OMIM	23%	100%	21%	90%	43%	95%
SP ∩ OMIM	11%	100%	**9%**	**100%**	20%	100%
SP ∪ OMIM	**29%**	**100%**	21%	83%	50%	92%

Prec.: précision

The rest of the samples went through the partial matching procedure. As precision is an essential requirement for an automatic mapping, we set up a rather stringent score threshold of +1, above which a partial match was considered as relevant (Fig. 2). Using this threshold, we were able to map an additional set of 20 Swiss-Prot diseases and 21 OMIM terms with a precision of 80% and 90%, respectively. The better performance found with the OMIM terms can be explained by the presence of disease name synonyms in this database. Although the precision obtained by partial mapping is clearly sufficient to aid in the manual curation process, we aimed to further improve the correctness of the mapping so as to automate the mapping procedure completely. For this purpose, we took advantage of the independence of mappings from the two sources, Swiss-Prot and OMIM, and included an additional condition: the respective mappings should point to the same MeSH descriptor. With this condition, the mapping provided by partial matches reduced to 8 diseases, but with a precision of 100%.
In summary, by combining the results of the partial and exact matches provided by both Swiss-Prot and OMIM, we were able to correctly map 35 disease names of the benchmark. This corresponded to a recall of 37% and a 100% precision.

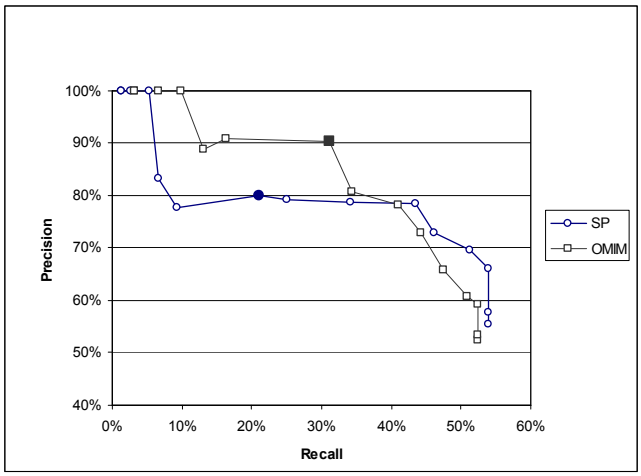


Fig. 2. Recall-precision at each integer unit of the similarity score in the interval [-7,+7]. The black points correspond to the recall/precision at the selected score threshold (+1).

3.3 Automatic mapping of Swiss-Prot disease comment lines

The mapping procedure was used to map the 2'966 disease comment lines present in Swiss-Prot. About 73% of them had corresponding OMIM entry. The results of the mapping were detailed in Table 2. Following the safe combination method described previously, we obtained a global performance of 1031 mapped terms, representing 35% of the total number of disease comment lines. The slight decrease in performance of the mapping with OMIM terms (37% compared to 46% of the benchmark) can be explained by the higher proportion of lines having an OMIM citation in the benchmark (89%). Of course, the precision of the mapping cannot be assessed, and the results are expressed in terms of retrieval instead of recall.
However, as the figures above do not differ significantly from the benchmark, it is likely that the performance is comparable.

Table 2. Mapping of the Swiss-Prot disease lines on MeSH terms

2966 disease comment lines 2173 OMIM	**SP**	**OMIM**	**SP ∩ OMIM**	**SP ∪ OMIM**
Exact match	483 (16%)	610 (21%)	292 (10%)	794 (27%)
Partial match	634 (21%)	483 (16%)	237 (8%)	640 (22%)
Total	1117 (38%)	1093 (37%)	529 (18%)	1434 (48%)

SP: Swiss-Prot

As a first assessment, we simply checked if, in case of exact matches, corresponding Swiss-Prot and OMIM terms mapped to identical MeSH descriptors. This statement was confirmed except in 5 cases. These discrepancies between descriptor matching had two causes. In one case, it was a problem of multiple diseases mentioned in the Swiss-Prot comment line, the one with a OMIM reference being different from the one extracted. For the 4 other cases, one of the exact matches was found with an OMIM synonym (alternative title) which corresponded to a distinct descriptor in MeSH.

4 DISCUSSION

In this study, we designed a mapping procedure to link the UniProtKB/Swiss-Prot human protein entries and the corresponding OMIM entries to the MeSH disease terminology. MeSH was chosen as it is interlinked with many biomedical terminologies within the UMLS. More importantly, its intimate association with literatures will provide us with a valuable means for knowledge discovery using data-mining in the future.

Our procedure that combined exact and partial matches of disease names was able to provide a high precision mapping for more than one third of the total number of disease comment lines in UniProtKB/Swiss-Prot. Although this retrieval could be considered as low for certain applications, it should be noted that stringent conditions were chosen on purpose to provide a high quality fully automated mapping procedure. If manual curation could be solicited, we could accept a reduced precision by joining the mappings of Swiss-Prot and OMIM. This increased the coverage to almost 50% (with a corresponding precision of 92% tested on the benchmark).

One of the main problems encountered in the mapping process lay however in the difference of granularity between the terminologies. MeSH is indeed relatively coarse-grained for genetic diseases. One solution will be to use the hierarchical structure of MeSH to map to less specific concepts. This can be done by increasing the weight of critical words in the partial matches. Indeed, 18 disease names from the benchmark mapped to the correct less specific MeSH terms but they had a score below the threshold. For instance, the term "senile cataract" did not match to "cataract" with a sufficient score, as "senile" appears as rarely as "cataract" in the terminologies. To get rid of insignificant common words, the first method to explore is a natural language processing pre-treatment of terms, such as stop word removal or stemming. Previous studies had shown the efficiency of such methods in terminology mapping processes (Sarkar et al. 2003, Johnson et al., 2006). Second, we can try to improve the word weighting by considering a common vocabulary resource for the word frequency calculation. The third approach would be to explore more sophisticated methods which include some information from the MeSH terminology structure in the score calculation. Such an attempt has been made, for instance, to categorise OMIM phenotypes using MeSH terms (van Driel et al., 2006). Nevertheless, the problem of MeSH granularity will hardly be completely solved by these methods. We need definitely to explore the use of other medical terminology resources, such as ICD – the official disease classification for diagnostic information, and SNOMED-CT – the clinical terminology used for clinical information.

In conclusion, this work represents the first step in standardizing the medical vocabularies in the UniProt Knowledgebase. Through this effort, we provide a bridge for the medical informatic community to explore the genomic and proteomic data present in the biological databases which could be of value for disease understanding.

ACKNOWLEDGEMENTS

This work was funded by the Swiss National Science Foundation (grant No 3100A0-113970).

REFERENCES

Cantor,M.N. Sarkar,I.N. Gelman,R. Hartel.F. Bodenreider,O. and Lussier,Y.A. (2003) An evaluation of hybrid methods for matching biomedical terminologies: Mapping the Gene Ontology to the UMLS. *Stud. Health Technol. Inform.*,**95,** 62-67.

Hamosh, A. Scott, A.F. Amberger, J.S. Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33,** D514-7.

Johnson,H.L. Cohen,K.B. Baumgartner,W.A. Lu,Z. Bada.M. Kester,T. Kim,H. and Hunter,L. (2006) Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. *Pac. Symp. Biocomput.,* 28-39.

Lussier,Y.A. and Li,J. (2004) Terminological mapping for high throughput comparative biology of phenotypes. *Pac. Symp, Biocomput*., 202-213.

Nelson,S.J. Schopen,M. Savage,A.G. Schulman,J.L. and Arluk,N. (2004) The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation. *Proc. Med. Inform.,* 67-69.

Sarkar,I.N. Cantor,M.N. Gelman,R. Hartel,F. and Lussier,Y.A. (2003) Linking biomedical language information and knowledge resources: GO and UMLS. *Pac. Symp. Biocomput*., 439-450.

Shatkay,H. (2005) Hairpins in a bookstacks: Information retrieval from biomedical text. *Brief. Bioinform.,* **6,** 222-38.

The UniProt Consortium, (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.,* **35**, D193-D197.

van Driel,M.A Bruggeman,J. Vriend,G. Brunner,H.G. and Leunissen,J.A. (2006) A text-mining analysis of the human phenome. *Eur J Hum Genet.* **14**, 535-42.

Zhang,S. Mork,P. Bodenreider,O. and Bernstein,P.A. (2007) Comparing two approaches for aligning representations of anatomy. *Artif. Intell. Med.*, **39**, 227-236.

Integrating and querying disease and pathway ontologies: building an OWL model and using RDFS queries

Julie Chabalier, Olivier Dameron, Anita Burgun

EA 3888 – IFR 140, Faculté de Médecine, Université de Rennes 1, 35033 Rennes France

ABSTRACT

Motivation: Our goal is first to create a biomedical ontology relating diseases and pathways, and second to exploit this knowledge. We first created a knowledge source relating diseases and pathways by integrating GO, KEGG orthology and SNOMED CT. We proposed an approach combining mapping and alignment techniques. We used OWL-DL as the common representation formalism. Second, we demonstrated that RDFS queries were expressive enough with acceptable computational performances.

1 INTRODUCTION

Nowadays, there is a need for biology and medical research to be able to compute with the knowledge component (Bodenreider and Stevens, 2006). Use of ontologies within the biomedical domain is currently mainstream (e.g. Gene Ontology (GO Consortium, 2006)). Within a few years, the success of bio-ontologies has resulted in a considerable increase in their number (e.g Open Biological Ontologies). While some of these bio-ontologies contain overlapping information, most of them cover different aspects of life science. However, an application may require a domain ontology which spans several ontologies. Rather than to create a new ontology, an alternate approach consists of reusing, combining and augmenting these bio-ontologies in order to cover the specific domain (Marquet et al., 2006).

A major step for addressing this issue is "ontology integration", which sets up relations between concepts belonging to different ontologies. It encompasses several notions: 1) *merging* consists in building a single, coherent ontology from two or more different ontologies covering similar or overlapping domains, 2) *aligning* is achieved by defining the relationships between some of the terms of these ontologies (Klein, 2001), 3) *mapping* corresponds to identifying similar concepts or relations in different sources (Lambrix and Tan, 2006).

The automatic exploitation of the knowledge represented in integrated ontologies requires an explicit and formal representation. Description logics, and OWL (Web Ontology Language) in particular, offer a compromise between expressivity and computational constraints (Horrocks et al., 2003). However, for leveraging its expressivity, ontologies should contain features such as necessary and sufficient definitions for classes whenever possible, as well as disjointness constraints. While recent works put forward a set of modeling requirements to improve the representation of biomedical knowledge (Rosse et al., 2005; Stevens et al., 2007), current biomedical ontologies are mostly taxonomic hierarchies with sparse relationships. Even though, dedicated reasoners are hardly able to cope with them.

Associations between classes of genes and diseases as well as associations between pathways and diseases are key components in the characterization of diseases. Different phenotypes may share common pathways and different biological processes may explain the different grades of a given disease. However, this information remains absent in most existing disease ontologies, such as SNOMED CT®. Pathway related information is present in other knowledge sources. The KEGG PATHWAY database is a collection of pathways maps representing our knowledge on the molecular interaction and reaction networks for metabolism and cellular processes (Kanehisa et al., 2004). As the GO does not provide direct association with pathways, Mao et al. have proposed to use the KEGG Orthology (KO) as a controlled vocabulary for automated gene annotation and pathway identification (Mao et al., 2005). Recently, information about the pathways involved in human diseases has been added to KO.

The objective of this study is to infer new knowledge about diseases by first integrating biological and medical ontologies and finally querying the resulting biomedical ontology. We hypothesize that most typical queries do not need the full expressivity of OWL and that RDFS is enough for them. In this paper, we use the term 'pathway' for metabolic pathways, regulatory pathways and biological processes. The approach presented here consists in developing a disease ontology using knowledge about pathways as an organizing principle for diseases. We represented this disease ontology in OWL. Following an integration ontology methodology, pathway and disease ontologies have been integrated from three sources: SNOMED CT, KO, and GO. To investigate how information about pathways can serve disease classification purposes, we compared, as a use case, glioma to other neurological diseases, including Alzheimer's disease, and other cancers, including chronic myeloid leukemia.

2 INTEGRATION FRAMEWORK

Adapted from the Pinto's methodology for ontology integration (Pinto and Martins, 2001), we followed three main steps to build an integration framework: 1) identify candidate ontologies and their relevant parts, 2) get these candidate ontologies in an adequate form and 3) apply integration operations.

2.1 Identify candidate ontologies

The KEGG PATHWAY database was used as the reference database for biochemical pathways. It contains most of the known metabolic pathways and some regulatory pathways. KO is a further extension of the ortholog identifiers, and is structured as a DAG hierarchy of four flat levels. The top level consists in the following five categories: metabolism, genetic information processing, environmental information processing, cellular processes and human diseases. The second level divides the five functional categories into finer sub-categories. The third level corresponds to the pathway maps, and the fourth level consists in the genes involved in the pathways. The first three levels of this hierarchy were integrated in the disease ontology.

Gene Ontology is the most widely used bio-ontology. It presents three independent hierarchies relative to biological processes, molecular functions and cellular components. A biological process is an ordered set of events accomplished by one or more ordered assemblies of molecular functions (e.g. "cellular physiological process" or "pyrimidine metabolism"). Since we consider all pathways as biological processes, the biological process hierarchy was used to enrich the pathway definitions.

SNOMED CT was used as reference source for disease definitions because it is the most comprehensive biomedical terminology recently developed. We used SNOMED to enrich the definitions of human diseases provided by KO.

2.2 Get candidate ontologies in an adequate form

The three ontologies that we identified are all in a specific format. In order to integrate them, it was necessary to convert the relevant parts into a common formalism. This formalism has to be compatible with our requirements of generating new knowledge through the combination of elements from various ontologies, and of supporting a uniform query mechanism.

OWL is a W3C recommendation for representing ontologies. The elements of the ontologies can be referred to by their Unique Resource Identifier (URI). Therefore, integrating several ontologies is simple through the use of namespaces for avoiding potential ambiguities. It is also possible to combine the elements of several ontologies. Moreover, the OWL language has a precisely defined semantics, and its OWL-Lite and OWL-DL variants support advanced classification-based reasoning capabilities. This is particularly relevant as our integration effort encompasses a new view on pathologies through genomic information, which implies the creation of new hierarchies.

We selected the OWL language. The OWL representation of GO is available on the GO website. We automatically converted in OWL the relevant parts of KO and SNOMED-CT. KO hierarchy is available in HTML format. We extracted the three upper levels of this hierarchy. Each KO class was represented by an OWL class respecting the subsumption hierarchy. SNOMED CT is not freely available. However, it is part of the UMLS knowledge Sources (Bodenreider, 2004). Therefore, we extracted the relevant concepts and their parents, as well as their relations, from the SNOMED CT part of the UMLS. The concepts and relations were respectively represented as OWL classes and properties.

2.3 Ontology integration

The ontology integration process was based on ontology alignment, which defines relationships between terms, and on ontology mapping, which is a restriction of ontology alignment by taking into account only equivalence relationships between terms.

First, the KO terms were mapped to GO biological process terms using lexical mapping. Then, as the KEGG pathways often correspond to composite terms, e.g. "Fructose and mannose metabolism", we segmented and reconstructed these terms according to the coordination conjunctions. For example, the segmentation-reconstruction operation on the KO term "Fructose and mannose metabolism" resulted in two terms, "Fructose metabolism" and "Glucose metabolism", which are both present in GO. We used the MetaMap Transfer program (version 2.4.B) (Aronson, 2001) to map a list of terms to UMLS concepts with the possibility to restrict the output to selected sources (restrict_to_sources option). As GO is part of the UMLS source vocabularies, we used this option to restrict the mappings to GO. The resulting GO terms were aligned to the initial KO terms by subsumption relations. Both equivalence and subsumption relations enriched the disease ontology.

The lexical alignment KO-GO aimed to link the disease and pathway classes through the hasPathway/isPathway relationships. We aligned KEGG diseases and GO biological processes by using the KEGG PATHWAY database and the Gene Ontology Annotation files (GOA) (Camon et al., 2004) to retrieve these relations. First, for each disease present in KO, we extracted the relations between the genes and the disease pathways from the fourth level of the KO hierarchy. We retrieved the biological pathways in which these genes are involved by mining the KO hierarchy and the GOA files. We have considered that a pathway is related to a disease if a gene is involved in both the disease and the pathway according to KEGG and GOA.

The KEGG pathways were mapped to the concepts of SNOMED CT using lexical mapping. As for the mapping to

GO, we segmented and reconstructed these terms. Like GO, SNOMED CT is part of UMLS, we used the restrict_to_sources option of Metamap to perform this mapping.

3 QUERYING THE DISEASE ONTOLOGY

Queries can be used either for checking the consistency or for exploiting the resulting integrated bio-ontology. Typical consistency queries consist in detecting if a specific pathway and a more general one are associated with a same disease. Such an imprecision of granularity can either come from one faulty ontology or from the integration of the knowledge from two ontologies with different granularity. Typical queries for exploiting the ontology involve 1) retrieving the pathways common to several diseases, 2) retrieving the pathways associated with one disease but not with another one, or 3) retrieving the diseases associated with the pathways associated with one class of diseases.

Computing the solutions for both kinds of queries only requires following explicit relations. It does not require OWL-based classification, and can be performed using only the RDFS semantics. We loaded the ontology in a Sesame RDF repository, and represented the queries using the SeRQL language. The disease ontology and some examples of typical queries are available on the project website[1].

4 BIOMEDICAL USE CASE

Of the 16 disease entities modeled in the KO hierarchy (third-level subclasses of the Human Disease class), 15 disease terms were successfully mapped to SNOMED CT concepts through MetaMap. The remaining term, considered as disease in KEGG, “Epithelial cell signaling in Helicobacter pylori infection”, was partially mapped to the concept “Helicobacter pylori infection (Helicobacter Infections)”.

Of the 252 KO classes from the three first levels (excluding disease related classes), 123 have an exact correspondence in GO. This relatively low number is due to the complexity of the KO classes which are sometimes represented by composite terms. From the 39 composite terms, we identified three composition patterns, which were used for term reconstruction. This resulted in 83 terms. Among them, 68 are biological consistent (manually validated). The strict mapping of the validated terms results in 21 supplementary classes aligned to GO terms. Finally, 144 of the 252 KO classes (57%) were successfully associated withGO terms.

For being able to manually check that our queries returned correct results, we considered three diseases: chronic myeloid leukemia, glioma, and Alzheimer's disease. First, we performed some RDFS queries for checking the consistency of the integrated ontology. Among the pathways associated with one disease, 87 are more general than some other pathway associated with this disease (47 for leukemia, 29 for glioma and 10 for Alzheimer’s disease). We removed the least specific pathways. We then performed some RDFS queries for comparing diseases by their associated pathways. First, we compared two neurological disorders, namely glioma and Alzheimer’s disease. 8 direct pathway classes involved in glioma were also associated to Alzheimer’s disease (86 indirect classes). Then we compared glioma and leukemia; 44 direct pathway classes were shared by these two cancers (165 indirect classes). Finally, 37 pathways are specific to these two cancers (97 indirect classes). Furthermore, the three diseases are associated with pathways themselves associated with glioma. Due to space limitation, the results of the comparison of the three diseases are exposed online.

5 DISCUSSION

This study focused on the integration of biological and medical ontologies. The first step of the integration methodology consisted in identifying candidate ontologies. We selected SNOMED CT, KO hierarchy and GO. Bio-ontology integration was complicated due to the different representation formalisms of the ontologies. We chose OWL as a common ground and had to convert two of them. The bio-ontologies natively represented in OWL are still scarce, and most of the OWL ontologies fail to exploit its full semantic richness. Until this situation improves, the integration of bio-ontologies will still be limited to lexical techniques similar to those we used. Formalization of at least some bio-ontologies is work in progress, as for instance the GONG project (Aranguren et al., 2007) for GO. We plan to incorporate more formal ontologies in our work and evaluate their contribution.

While our KO to GO mapping approach focused on biological processes, the original KEGG to GO mapping, available on the KEGG website is mainly restricted to GO molecular functions. As the three GO hierarchies are independent, we could not access the processes associated with molecular functions through this mapping. Therefore, we had to develop a standalone mapping method.

The GO integration increased the number of pathways in the bio-ontology. The SNOMED ontology supplied intrinsic knowledge about integrated diseases.

The representation of the relevant parts of the ontologies was done in OWL in order to be able to represent some background knowledge that would otherwise remain implicit in RDFS (e.g. that the siblings are disjoint). The work overload for doing so was not significant. Moreover, because of the OWL semantics, all the OWL classes are RDFS classes. Therefore, the OWL ontologies we produced are also valid RDFS ontologies. This is why we were able to perform RDFS queries on these ontologies with no

[1] http://www.ea3888.univ-rennes1.fr/biomed_ontology/

additional cost, while retaining the possibility to use the power of OWL-reasoning. This last point is of particular importance with respect to our approach based on ontology reuse: others may need to reuse the ontologies we produced in another context for performing reasoning tasks beyond the expressivity of RDFS. Eventually, our approach showed (1) that the full power of OWL is not necessary for performing some queries on bio-ontologies and that RDFS queries are sometimes enough, and (2) that these RDFS queries can still be performed on OWL ontologies. The combination of these two points is encouraging, as even though most of the available bio-ontologies are represented in OWL, they currently do not use all the richness of the OWL operators and are consequently useless for leveraging its expressivity. As the quality of these ontologies increases by the application of guidelines and best practices, OWL reasoning will gradually become possible.

The biomedical domain complexity as well as the size and the heterogeneity of the available resources justified our automatic knowledge integration approach. Moreover, it supports both the evolution of the ontologies we used and the integration of additional ones.

For validating our RDFS queries, we also modelled them as OWL classes. However, computing the results of these queries using standard reasoners such as Pellet or RacerPro, turned out impossible as of three diseases. We simplified the bio-ontology to consider only two diseases and verified that the results of the RDFS queries matched those obtained through OWL classification.

The goal of our study was to use knowledge about pathways for organizing diseases. This required the definition of the hasPathway relationship (and its inverse isPathwayOf) between a disease and its associated pathways. We could not reuse any relation from existing bio-ontologies. In order to combine bio-ontologies, we will need to represent additional relationships between their specific views on biology, for example the isRegulatedBy relationship. Such relations do not exist in the OBO foundry nor in the BioTop ontology (Schulz et al., 2006). In order to combine the existing bio-ontologies, a common domain-specific set of relationships is needed. It could be a specialization of the OBO relation ontology.

Similarly, we were interested in view of pathways as biological processes as proposed by the KO hierarchy. Combining our ontology with BioPax (Strömbäck and Lambrix, 2005) would allow to enrich it by providing a view of pathways as biochemical reactions. This could result in additional organizing principles for diseases.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Region Bretagne (PRIR).

REFERENCES

Aranguren, M.E., Bechhofer, S., Lord, P., Sattler, U. and Stevens, R.(2007) Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. *BMC Bioinformatics*.8:57.

Aronson, A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, AMIA Symp., 17-21.

Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Res*., 32, Database issue:D267-70.

Bodenreider, O. and Stevens, R. (2001) Bio-ontologies: current trends and future directions. *Brief Bioinform.*, 7:256–274.

Camon, E., Magrane, M., Barrell, *et al.* (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, 32:D262-266.

GO Consortium (2006), The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*, 34:D322-326.

Horrocks, I., Patel-Schneider P.F. and van Harmelen F. (2003) From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics*,1(1):7-26.

Kanehisa, M, Goto, S., Kawashima, S., Okuno, Y and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, 32:D277-280.

Klein, M. (2001) Combining and relating ontologies: an analysis of problems and solutions, *Proceedings of the Workshop on Ontologies and Information Sharing* (IJCAI'01), Seattle, USA.

Lambrix, P. and Tan H. (2006) SAMBO - A System for Aligning and Merging Biomedical Ontologies. *Journal of Web Semantics*, 4.

Mao, X., Cai, T., Olyarchuk, J.G. and Wei, L. (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, 1;21(19):3787-93.

Marquet, G., Mosser, J. and Burgun, A (2006) Aligning biomedical ontologies using lexical methods and the UMLS: the case of disease ontologies. *Stud Health Technol Inform.*,124:781-6.

Pinto, H.S. and Martins J.P. (2001) Ontology Integration: How to perform the Process, Proceedings of IJCAI2001, *Workshop on Ontologies and Information Sharing*, Seattle, USA, 71-80.

Rosse, C., Kumar, A., Mejino, J.L., Cook, D.L., Detwiler, L.T. and Smith B (2005) A strategy for improving and integrating biomedical ontologies. *AMIA Annu Symp Proc*. 639-43.

Schulz, S., Beisswanger, E., Hahn, U., Wermter, J., Stenzhorn, H. and Kumar, A. (2006) From GENIA to BioTop: Towards a top-level Ontology for Biology. *International Conference on Formal Ontology in Information Systems* (FOIS 2006).

Stevens, R., Aranguren, M.G., Wolstencroft, K., Sattler, U., Drummond, N., Horridge, M., Rector, A (2007) Using OWL to model Biological Knowledge. Int J Hum Comput Stud. in Press.

Strömbäck, L. and Lambrix, P. (2005) Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPax. Bioinformatics, 15;21(24):4401-7.

Wroe, C.J, Stevens, R., Goble, C.A. and Ashburner, M. (2003) A Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL. *Pacific Symposium on Biocomputing*. 8:624-635.

Gene Ontology Annotations: What they mean and where they come from

Judith Blake[1], David P. Hill[1], Barry Smith[2]*

1 Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME

2 Department of Philosophy and Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, NY

ABSTRACT

The computational genomics community has come increasingly to rely on the methodology of creating annotations of scientific literature using terms from controlled structured vocabularies such as the Gene Ontology (GO). We here address the question of what such annotations signify and of how they are created by working biologists. Our goal is to promote a better understanding of how the results of experiments are captured in annotations in the hope that this will lead to better representations of biological reality through both the annotation process and ontology development, and in more informed use of the GO resources by experimental scientists.

1 INTRODUCTION

The PubMed literature database contains over 15 million citations and it is clearly beyond the ability of anyone to comprehend information in such amounts without computational help. One avenue to which bioinformaticians have turned is the discipline of ontology, which allows experimental data to be stored computationally in such a way that it constitutes a formal, structured representation of the reality captured by the underlying biological science. An ontology of a given domain represents types and the relations between them in a formal way that is designed to support automatic reasoning about the instances of these types. From the perspective of the biologist, the development of bio-ontologies has enabled and facilitated the analysis of very large datasets. This utility comes not from the ontologies *per se*, but from the use to which they are put in a data curation process that results in *annotations*, which are statements of associations of genes or gene products with the types of biological entities designated by terms in ontologies.

One prime use of an ontology such as the GO (The Gene Ontology Consortium, 2006) is in the creation of annotations by the curators of model organism databases (e.g., Blake *et al.*, 2006; Cherry *et al.,* 1997; Grumbling *et al*., 2006) and genome annotation centers (Camon *et al*., 2004) designed to capture information about the contributions of gene products to biological systems as reported in the scientific literature. Because the annotations are so integral to the use of bio-ontologies, it is important to understand how the curatorial process proceeds. We here use the GO annotation paradigm to illustrate important aspects of this process.

To help in understanding, we provide a glossary of the terms most important to our discussion:

An ***instance*** is a particular entity in spatio-temporal reality, which instantiates a type (for example, a type of gene product molecule, a type of cellular component).

A ***type*** (aka "universal") is a general kind instantiated by an open-ended totality of instances that share certain qualities and propensities in common.

A ***gene product instance*** is a molecule (usually an RNA or protein molecule) generated by the expression of a nucleic acid sequence that plays some role in the biology of an organism.

A ***molecular function instance*** is the enduring potential of a gene product instance to perform actions, such as catalysis or binding, on the molecular level of granularity.

A ***biological process instance*** (aka "occurrence") is a change or complex of changes on the level of granularity of the cell or organism that is mediated by one or more gene products.

A ***cellular component instance*** is a part of a cell, such as a cellular structure or a macromolecular complex, or of the extracellular environment of a cell.

For each of the *instance* terms in the above, there is a corresponding *type* term (defined in the obvious way; thus a *molecular function type* is a type of molecular function instance, and so on).

An ***annotation*** is the statement of a connection between a type of gene product and the types designated by terms in an ontology such as the GO that is created on the basis of the observations of the instances of such types in experiments and of the inferences drawn from such observations.

An ***evidence code*** is a three-letter designation used by curators during the annotation process that describes the type of

* Corresponding author. All authors made equal contributions to this work.

experimental support linking gene product types with GO types, for example: ***IDA*** (Inferred from Direct Assay: used when someone has devised an assay that directly measures the execution of a given molecular function and the experimental results show that instances of the gene product serve as agents in such executions), and ***IGI*** (Inferred From Genetic Interaction: used when an inference is drawn from genetic experiments using instances of more than one gene product type to the effect that molecules of one of these types are responsible for the execution of a specified molecular function).

2 THE CURATOR PERSPECTIVE

A GO annotation represents a link between a gene product type and a molecular function, biological process, or cellular component type (a link, in other words, between the gene product and what that product is capable of doing, what biological processes it contributes to, or where it is capable of functioning in the natural life of an organism). Formally, a GO annotation consists of a row of 15 columns. For the purpose of this discussion, there are 4 primary fields: i) the ID for the gene product being annotated; ii) the ID for the ontology term being associated with the gene product; iii) an evidence code, and iv) the reference/citation for the source of the information that supports the particular annotation. For example, the sonic hedgehog gene has an MGI_ID MGI:98297, has been assigned a GO_ID GO:0043237 using the IDA evidence code, and the experiment was reported in PMID: 15056720 [http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=markerGO&key=13433] Additional details about the annotation structure and GO-defined annotation processes are available at the GO website [http://www.geneontology.org/GO.annotation.shtml].

The annotation process can be described as a series of steps. First, specific experiments, documented in the biomedical literature, are identified as relevant to the curation process responsibilities of a given curator. Second, the curator applies expert knowledge to the documentation of experimental results to decide on appropriate data associations with the ontology in question. Finally, annotation quality control processes are employed to ensure that the annotation has a correct formal structure, to evaluate annotation consistency among curators and curatorial groups, and to harvest the knowledge emerging from the activity of annotation for the contributions it might make to the refinement and extension of the GO itself.

2.1 Identification of relevant experimental data

The main goal of the GO annotation effort is to create genome-specific annotations supported by evidence obtained in experiments performed in the organism being annotated. In fact, however, many annotations are *inferred* from experiments performed in other organisms, or they are inferred not from experiments at all but rather from knowledge about sequence features for the gene in question. Such information, too, is captured in the GO annotation, with corresponding evidence codes. It is thus important for the user of GO Annotations to understand not only what an experimental annotation actually reflects, but also how sequence and/or protein architecture similarities can be utilized to transfer GO annotations between closely-related species or among paralog groups within species. Each of these sources of information for GO annotations can be identified in the annotation file. This complexity, if not taken account of by the user, can confound data analyses and undermine the goal of hypothesis generation on the basis of GO annotation sets.

2.2 Identification of the appropriate ontology annotation term

The decision as to what GO term to use in an annotation depends on several factors. First, the experiment itself often brings some limit on the resolution of what can be understood from its results. For example, cell fractionation might localize a protein to the nucleus, but immunolocalization experiments might localize a protein to the nucleolus. As a result, the same gene may have annotations to different terms in the same ontology based on different experiments. Second, curators have different degrees of expertise both in domains of biology and in the full content of the GO. Thus there may be some slight but measurable variation between curators in the selection of GO terms that they might employ given the same experimental data. As a result of these factors, users of GO annotations are best served by taking advantage of the graph structure of GO's three constituent hierarchies to combine annotations for several levels of the GO in their analyses.

Corresponding to the three GO hierarchies are three different kinds of annotations:

3 MOLECULAR FUNCTION ANNOTATION

In the simplest biological situation, molecules of a given type are associated with a single molecular function type. A specific molecule *m* is an instance of a molecule type *M* (represented for example in the UniProt database), and its propensity to act in a certain way is an instance of the molecular function type F (represented by a corresponding GO term). So, a molecule of the gene product type *Adh1*, alcohol dehydrogenase 1 (class I), has as its function an instance of the molecular function type *alcohol dehydrogenase activity*. This means that such a molecule has the potential to execute this function in given sorts of contexts.

If we say that instances of a given gene product type have a potential to execute a given function, for example to catalyze a certain reaction, this does not mean that every instance of this type will in fact execute this function. Thus molecules of the mouse gene product type *Zp2* are found in the oocyte and have the propensity to bind molecules of the

gene product type *Acr* during fertilization. If, however, an oocyte is never fertilized, the molecules still exist and they still have the propensity to execute the binding function, but the function is never executed.

The experimental evidence used to test whether a given molecular function type *F* exists comes in the form of an 'assay' for the execution of that function type in molecules of type *M*. If instances of *F* are identified in such an assay, this justifies a corresponding *molecular function annotation* asserting an association between *M* and *F*.

4 BIOLOGICAL PROCESS ANNOTATION

A biological process instance is made up of the *executions* of one or more such molecular function instances working together to accomplish a certain biological objective. Like molecular functions, biological processes, too, are detected experimentally. When instances of biological process type *P* are detected, either by direct observation or by experimental assay, as being associated with instances of a given gene product type *M*, then this justifies the assertion of an association between *M* and *P* where a molecule of gene product type *M* can execute a molecular function in an instance of *P*. This is called a biological process annotation.

We can discover how gene product, function and process types are related together by examining instances as they interact in some context in which a biological process is being performed. If we observe in such a context molecules of a given type, and if we observe also their coordinated execution of a functions of a given types, then these types can be associated in their turn with the corresponding overarching biological process type.

5 CELLULAR COMPONENT ANNOTATION

In a large majority of cases, annotations linking gene product with cellular component types are made on the basis of a direct observation of an instance of the cellular component in a microscope, as for example in (MacPhee *et al*., 2000), which reports an experiment in which an antibody that recognizes gene products of the *Atp1a1* gene is used to label the location of instances of such products in preimplantation mouse embryos. The fluorescent staining shows that the gene products are located at the plasma membrane of the cells of these embryos. In this case, the instances of the gene products are the actual molecules that are so labeled, and the instance of the cellular component is the plasma membrane that is observed under the microscope. A curator has accordingly used the results of this experiment to make an annotation of the ATP1A1 gene product to the GO cellular component *plasma membrane*, which asserts that a molecule of ATP1A can be found in an instance of the cellular component type *plasma membrane*.

6 ONTOLOGIES AND ANNOTATIONS

The development of an ontology reflects a shared understanding of the domain being represented on the part of domain scientists. This understanding, for biological systems, is the result of the cumulation of experimental results reflecting that iterative process of hypothesis generation and testing for falsification which is the scientific method. The process of annotation brings new experimental results into relationship with the existing scientific knowledge that is captured in the ontology. There will, then, necessarily be times when new results stand in conflict with the then current version of the ontology. One of the strengths of the GO development paradigm is that development of the GO has been primarily a task of biologist-curators who are experts in understanding specific experimental systems: as a result, the GO is continually being updated in response to new information. GO curators regularly request that new terms be added to the GO or suggest rearrangements to the GO structure, and the GO has an ontology development pipeline that addresses not only these requests but also submissions coming in from external users. In addition, the GO community works with scientific experts for specific biological systems to evaluate and update GO representations for the corresponding parts of the ontology.

7 DISCUSSION

Gene Ontology annotations report connections between gene products and the biological types that are represented in the GO using GO evidence codes recording the process by which these connections are established via experimental analysis of actual instances of gene products in a laboratory or of inferential reasoning from such analysis. We believe that an understanding of the role of instances in spatiotemporal reality at the beginning of this sequence – instances of gene product types and of the types designated by Gene Ontology terms – can provide for a more rigorous analysis of the knowledge that is conveyed by the annotations and by the ontologies themselves. While each annotation rests ultimately on the observation of instances in the context of a scientific experiment, it is not about such instances. Rather it is about the corresponding types. This is possible because annotations are derived by scientific curators from the published reports of scientific experiments describing general cases, cases for which we have scientific evidence to believe that they are typical. If such beliefs are falsified through further experimentation, then the corresponding annotations will need to be revised.

It is to us obvious that our cumulative biological knowledge should represent how instances relate to one another in reality, and that any development of bio-ontologies and of relationships between such ontologies should take into account information of the sort that is captured in annotations. While we are still at an early stage in the process of creating truly adequate algorithmically processable representations of biology reality, we believe that the GO methodology of allowing ontology development and creation of annotations to influence each other mutually represents an evolutionary path forward in which both annotations and ontology are being enhanced, in a cumulative process, in both quality and reach.

ACKNOWLEDGEMENTS

The authors would like to thank Monica S. McAndrews-Hill, Cynthia Smith, Terry Hayamizu and Waclaw Kusnierczyk for their help with earlier versions of this manuscript. This work was supported by NIH grants HG02273 (DPH, JB) and HG00330 (JB) and by the NIH Roadmap for Medical Research, Grant 1 U 54 HG004028 (BS).

.

REFERENCES

Gene Ontology Consortium (2006) The Gene Ontology. (GO) project in 2006. *Nucl. Acids Res*, **34**, D322-D326

Blake JA, Eppig JA, Bult CJ, Kadin JA, Richardson JE and Mouse Genome Database Group. (2006) The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res*, **34**, D562-7.

Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, Botstein D. (1997) Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, **387(6632 Suppl)**, 67-73

Grumbling G, Strelets V, and The FlyBase Consortium. (2006) FlyBase: anatomical data, images and queries. *Nucleic Acids Res*, **34**, D484-D488.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, **32**, D262-D266.

MacPhee DJ; Jones DH; Barr KJ; Betts DH; Watson AJ; Kidder GM. (2000) Differential involvement of Na(+),K(+)-ATPase isozymes in preimplantation development of the mouse. Dev Biol, **222(2)**, 486-498.

Ontology Design Patterns for bio-ontologies

Mikel Egaña Aranguren[1], Robert Stevens[2] and Erick Antezana[3]*

[1,2] (mikel.eganaaranguren@cs.man.ac.uk) University of Manchester Computer Science, Oxford Road, M13 9PL, UK

[3] Department of Plant Systems Biology, Flanders Institute for Biotechnology (VIB)/ Ghent University., Technologiepark 927, Ghent, Belgium

ABSTRACT

Conceptual modeling in bio-ontologies is often a difficult task and as a result most current bio-ontologies are not axiomatically rich. A contribution to the solution of this problem is to provide biologists with "modeling recipes", in the form of Ontology Design Patterns (ODPs) that could be used for an easier and more efficient modeling of bio-ontologies in OWL or OBO. This paper describes what ODPs are and why they should be used, also exploring documentation and application methodologies for ODPs.

1 INTRODUCTION

Bio-ontologies represent knowledge about the different domains of biology in a computer amenable way, and they are used as integrators of resources or as knowledge bases (KBs). They have now become a central element of knowledge management in bioinformatics.

For a bio-ontology to be as useful as possible, the knowledge represented by it must reflect the domain with the highest resolution possible. However, conceptual modeling is a difficult task for many bio-ontologists, since they usually lack the necessary training on conceptual formalisms. Knowledge Representation (KR) languages like OWL [1] or OBO [2] offer a plethora of modeling primitives, but that power is not often used to its full potential by modelers when building bio-ontologies. Richer axiomatic descriptions of the domain being modeled mean a higher-fidelity ontology and also one that allows for a wider range of inferences to be made, both over the ontology and the objects it describes. To make things more difficult, bio-ontologists usually do not see the benefits of doing such thorough and fine-grained modeling, often leaving much of the expressivity in the labels used for classes or properties.

A step towards solving this problem is to provide bio-ontologists with ready-made modeling solutions in the form of Ontology Design Patterns (ODPs). Bio-ontologists should be able to model more richly in bio-ontologies with less effort, by applying ODPs, and hence improve the knowledge management in bioinformatics.

We start this paper by laying out our definition of ODPs, comparing it with prior efforts, and describing the advantages of using ODPs. Then we discuss the documentation system and application of ODPs. Finally a simple example of an ODP and pointers for future research are presented.

2 ONTOLOGY DESIGN PATTERNS

2.1 What are Ontology Design Patterns?

The concept of ODPs is analogous to the concept of Software Design Patterns (SDPs) in Object Oriented Programming [3]: they are solutions to modeling problems that appear repeatedly in different developments. These solutions are demonstrated to be efficacious, as they have been tested on plenty of different systems and they are well documented.

ODPs are solutions to modeling problems that help bio-ontologists to make better use of the expressivity of their KR language. ODPs are examples of solutions, rather than abstract solutions that are instantiated in different systems, as opposed to SDPs. Thus, a bio-ontologist could use an ODP as a guide and be able to recreate it in the concrete ontology in which it is being developed. Some ODPs, however, can have directly re-usable or generic parts and others do not.

Even though ODPs have already been documented in the literature, they have not been explicitly mentioned as such until recently [4,5]. In [5] they are mentioned as a part of some ontology building methodologies, without further analysis such as documentation and application. The concept of CODePs (Conceptual Design Patterns) [4] is close to ODPs, but they differ in the level of *resolution* of the proposed solution; CODePs are necessarily less fine grained than ODPs, as they represent conceptual and general patterns, whereas ODPs offer patterns in a given KR language with full semantic coverage (OWL or OBO). CODePs and ODPs are complementary: a CODeP will incarnate itself in an ODP in a concrete KR language: the ODP will inform bio-ontologists how to implement a given CODeP in a concrete KR language. This in fact happens with the CODeP Description-Reification [4] and the ODP N-ary Relationship [8]. Therefore, the application

*To whom correspondence should be addressed.

procedures, documentation system as well as the representation of ODPs and CODePs are different.

"Knowledge Patterns" [6] are conceptual general patterns that are "morphed" into a given KB by a set of axioms. Thus, the knowledge pattern can not be applied directly by bio-ontologists: this is a drawback since the application of the pattern needs to be as easy as possible. The semantics of the pattern and the pattern needed in the KB can be different or even incompatible. The same argument applies for the "Semantic Patterns" described in [7]. The ODPs presented in this paper are solutions to real biological modeling problems rather than theoretical propositions of general patterns; the value of ODPs is that they are ready to be used by bio-ontologists when building ontologies. ODPs are presented in OWL or OBO to make full use of the languages' semantics: the semantics can be mapped to other languages or formalisms for interoperability, but the opposite is less likely to happen: it is difficult for biologists, given a pattern in an abstract formalism, to apply that pattern to an actual ontology with a concrete KR language, via transformations.

Ontology Engineering and Knowledge Representation have been subjected to some formal analyses in terms of general best practices and design patterns, which in some cases are semantically equivalent to ODPs. Some of those efforts have been collected (albeit not as a systematized collection) in the W3C Semantic Web Best Practices and Deployment Working Group web [8]. Other efforts have been published as self-contained patterns in single publications, for example regarding partonomy [9]. In all those efforts documentation, graphical representation and application of the ODPs have not been mentioned as such (if mentioned at all), and only implicitly and partially used. Some of those ODPs are collected in the catalogue presented herein, documented using the documentation system proposed in Section 3.

2.2 Why use Ontology Design Patterns?

These are the main advantages that the use of ODPs can offer:

- **Expressive and granular modeling**. ODPs produce more richly axiomatised ontologies by allowing a more fine-grained modeling of the knowledge domain. They help in making the implicit knowledge found, for example, in term names, explicit, encoding it in the semantics of the ontology. On the other hand, bio-ontologies are going deeper in the knowledge they hold, and models to represent that deeper knowledge with the adequate granularity are needed.
- **Focused development**. Having an ODP as an engineering artifact reduces development time.An "off-the-shelf2 solution to a modeling need also allows a modeler to concentrate upon the needs of the domain itself.
- **Semantic encapsulation**. ODPs provide bio-ontologists with an easy way of dealing with the complexity of semantics in conceptual modeling by encapsulating it in the ODP.
- **Computationally explicit**. ODPs can be codified programmatically, providing a means for automatically building sectors of an ontology that are complex, regular or guide the bio-ontologist in the process.
- **Robustness and modularity**. Some ODPs help modelers in creating more robust and modular ontologies.
- **Good communication**. The use of ODPs improves communication between ontology developers. The developers can easily recognize the different features of the ontology produced by the ODP, as it represents a well known and thoroughly documented abstraction.
- **Documented modeling**. When creating bio-ontologies the process is more precisely documented by simply mentioning which ODPs were used.
- **Reasoning**. The richer axioms needed for efficient and productive reasoning are reached more easily using ODPs. Therefore, more sophisticated inferences can be performed.
- **Rapid prototyping.** Having prototypes allows developers to discuss complete models of ontologies in early stages and hence making sounder ontologies. It also allows for faster development. ODPs are ideal for rapidly developing prototypes.
- **Alignment.** More and more bio-ontologies are being developed and efficient ways for comparing/aligning them are necessary. The consistency of modeling inherent in the use of ODPs should support semantic matching between different ontologies.
- **Re-engineering.** ODPs may be applied in the beginning of an ontology development process as well as during the life cycle of it, providing, for instance, valuable insights for refactoring some components which may hold an inconsistency or which may violate design principles.

3 DOCUMENTING ONTOLOGY DESIGN PATTERNS

The documentation system proposed in this paper is inspired by the original SDPs documentation system, with

some changes. The system is based on sections that are mandatory for the description of an ODP (See Table 1).

ODPs can be broadly classified according to their usage:

Extensional ODPs: ODPs that can be used to overcome the limitations of a KR language (e.g. exceptions in OWL).

Good practice ODPs: ODPs that are used to ensure a modeling good practice. These ODPs are used to produce more modular, efficient and maintainable ontologies, tackling already known pitfalls of ontology engineering.

Domain Modeling ODPs: ODPs that are used to model a concrete part of the knowledge domain. They can be defined as "signature" ODPs: each knowledge domain has got its peculiarities and these ODPs are used to model those peculiarities.

Table 1. Documentation system sections and their explanation. R: required, O: optional.

Section name	*Explanation*
Name (R)	A unique name for the ODP
Also known as (O)	Any other name that is given to this ODP
URL (R)	An URL where the ODP can be obtained
Classification (R)	"Extensional", "Good practice" or " Domain Modeling"
Motivation (R)	The scenario where the ODP might be needed
Aim (R)	The concrete solution the ODP provides
Elements (R)	The ontology elements (e.g. classes) that build the ODP
Structure (R)	How the elements relate to each other to build the ODP
Implementation (R)	Explanation of how to build or apply the ODP in an actual system
Result (R)	The structure that should appear in the ontology after applying the ODP (and often after reasoning)
Side effects (R)	Any non obvious consequences of applying the ODP
Sample (R)	An extract of the main semantics of the ODP
Known uses (O)	Any system where the ODP has been applied
Related ODPs (O)	Any ODP that uses or is used by this ODP, or any ODP that has anything in common with this one
References (O)	Any publications or web page where this ODP has been previously mentioned
Additional information (O)	Any extra information needed

There is an implementation of an actual catalogue of real ODPs available as an alpha version in [10], with the following ODPs: Exception, N-ary Relationships, Normalisation, Value Partition, Upper Level Ontology, List and Adapted SEP triples. A future stable version of this catalogue will be implemented directly in OWL, with scripts for translations to XHTML and other languages. The catalogue will be implemented in OWL to ensure consistency of the whole catalogue, better querying and to provide the ODP semantics directly in the catalogue.

4 APPLYING ONTOLOGY DESIGN PATTERNS

4.1 DIRECT APPLICATION

The main method is to directly apply the ODP, thus to recreate completely or in part the structure of the ODP in the ontology, sometimes reusing parts of the example ODP. The user can be guided in the process with wizards, for example using the wizards provided by the CO-ODE project [11] for the Protégé ontology editor [12].

4.2 APPLICATION BY CONDITION MATCHING

ODPs can be applied to an ontology by defining a condition for matching classes of that ontology and adding the ODP when the match happens. The condition can be of two types:

Syntactic condition: the condition relies on a string value. Thus, the class name or any annotation value (label, comment, and so forth) can be used to try to match the condition. The condition can be a given value (e.g. *cell differentiation*.) or a regular expression (e.g. *(.+?) (differentiation)*).

Semantic condition: the condition relies on the structure of the ontology. For example, a condition can be defined so as to allow any class that happens to have a concrete class as its superclass to be matched. A semantic condition can be as complex as the user wishes (using boolean operators, restrictions, etc.) as the reasoner will try to match it against the ontology and retrieve any matched classes.

An implementation of the method is available with the name of Ontology Processing Language (OPL) [10]. OPL is a syntax partially based in the Manchester OWL syntax [13], that allows classes to be selected from an ontology in OWL according to a condition, such as the ones described above, and adding or removing axioms to/from the selected classes. The OPL commands are written in a flat file by the user and the OPL program parses the file, selecting classes and doing the changes defined, creating a new ontology. ODPs can be codified in the defined changes, and human-readable explanations can be written in comments. Thus, the ODPs are stored in a flat file for direct application, altogether with any comments bio-ontologists could find of

interest. Therefore, ODPs can be applied at any time, to any ontologies, by running the OPL program, and are stored persistently (See Table 2).

Table 2. Extract of an OPL flat file, to be processed by the OPL program. The statements to be processed are a SELECT statement followed by some ADD statements. The statements end with a semicolon and the comments (starting with hash) are not processed. This will result in the ODP Value Partition being added to the ontology (See Section 5).

```
SELECT ?x WHERE ?x label regulation;
ADD ?x equivalentTo (positive or negative);
ADD positive disjointWith negative;

# When parsing, the program will interpret the following, and perform it:
# "Select any class with the label 'regulation' and add an axiom that sets
# that class to be equivalent to the union of the classes 'negative' and
# 'positive'. Make 'positive' and 'negative' disjoint "
```

5 A SIMPLE ODP: THE VALUE PARTITION

A simple example of an ODP is the one known as Value Partition. This ODP is used to represent values that a parameter might take: for example, regulation can only be either positive or negative (Fig. 1). It is built using a covering axiom and a disjoint axiom. The covering axiom is used to ensure that if a new instance is added to the covered class, it must be an instance of one of the subclasses, which are mutually disjoint.

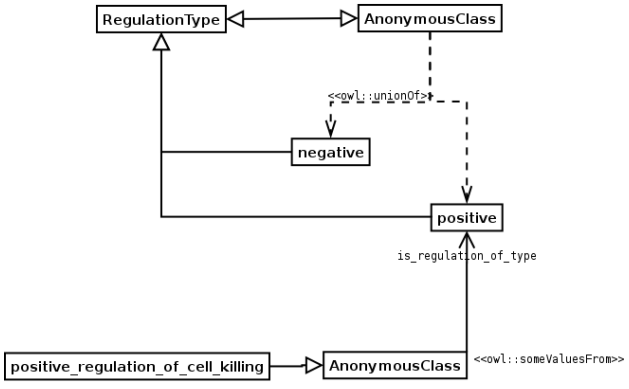


Fig. 1. The Value Partition is a very simple example of an ODP.

6 CONCLUSION AND FUTURE RESEARCH

ODPs offer advantages to modelers when creating and maintaining bio-ontologies; semantic encapsulation, better communication and documentation, more robust results, more efficient reasoning and querying, and faster and more reliable development.

There are open issues in the documentation system presented: the sections may change in the future as ODPs are more widely used, including the classification. A "graphical ontology metalanguage" *a la* UML for representing the structure of the ODPs is still needed, and there are ODPs that have not been completely explored.

Providing tools for bio-ontologists so they can easily create and manage ODPs is vital. A possible solution is the creation of a Protégé plugin that allows for the creation of ODPs graphically.

There are plenty of areas of biology that have not been explored to find possible ODPs, like taxonomy, phylogeny, molecular interactions, and many more. Those areas need the creation of new ODPs to make it possible for biologists to create better bio-ontologies (or bio-ontologies at all) on those domains. This will ultimately provide a more robust and fine grained representation of the knowledge in biology, allowing for a more efficient knowledge management in the field.

ACKNOWLEDGEMENTS

MEA is funded by Manchester University and EPSRC. EA is funded by the EU (FP6, contract number LSHG-CT-2004-512143).

REFERENCES

1. http://www.w3.org/2004/OWL/
2. http://www.obo.sf.net
3. E. Gamma, R. Helm, R. Johnson, and J. Vlissides. **Design Pattern - Elements of Reusable Object-Oriented Software**. Addison-Wesley (1995).
4. Gangemi A: **Ontology Design Patterns for Semantic Web Content**. *ISWC 2005.* LNCS 1729.
5. Svatek V: **Design Patterns for Semantic Web Ontologies: Motivation and Discussion.** *7th Conference on Business Information Systems, Poznan.* 2004.
6. Clark P, Thompson J and Porter B: ***Knowledge Patterns.*** **Handbook on Ontologies.** 2003: 121-134.
7. Staab S, Erdmann M and Maedche A: **Engineering Ontologies Using Semantic Patterns.** *IJCAI 2001.*
8. http://www.w3.org/2001/sw/BestPractices/
9. Rogers J, Rector A: **GALEN's Model of Parts and Wholes: Experience and Comparisons.** *AMIA 2000.*
10. http://www.gong.manchester.ac.uk
11. http://www.co-ode.org
12. http://protege.stanford.edu/
13. Horridge M, Drummond N, Goodwin J, Rector A, Stevens R, Wang H: **The Manchester OWL syntax.** *OWLed 2006*

Towards naming conventions for use in controlled vocabulary and ontology engineering

Daniel Schober[1], Waclaw Kusnierczyk[2], Suzanna E Lewis[3], Jane Lomax[1], Members of the MSI, PSI Ontology Working Groups[4,5], Chris Mungall[3], Philippe Rocca-Serra[1], Barry Smith[6] and Susanna-Assunta Sansone[1*]*

[1]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [2]Department of Information and Computer Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, [3]Berkeley Bioinformatics and Ontologies Project, Lawrence Berkeley National Labs, Berkeley, CA 94720 USA, [4]http://msi-ontology.sourceforge.net, [5]http://www.psidev.info/, [6]Center of Excellence in Bioinformatics and Life Sciences, and National Center for Biomedical Ontology, University at Buffalo, NY, USA

ABSTRACT

Motivation: For most people, the term "standard" generates an immediate impulse to run in the opposite direction. We all know that this means someone is bent upon the "one, true capitalization style", thereby fomenting an instantaneous rebellion. While it is somewhat audacious to propose standards, the adoption of a few shared simple conventions is an important strategy to improve quality in controlled vocabularies and ontologies we build. Ontologies should not only satisfy computational requirements, but also meet the needs of human readers who are trying to understand them. When confronted by the full complexity of an ontology, logical coherence and predictable naming is important, then our guesses about where something may be found, or what it is called, are right more often than wrong. Conforming to naming conventions in ontology construction will help consumers more readily understand what is intended and avoid the introduction of faults, and it is here where its value lies.

1 INTRODUCTION

A wide variety of controlled vocabularies, ontologies, and other terminological artifacts relevant to the biological or medical domains are already available through open access portals, such as the Ontology Lookup Service (OLS) [1] and the NCBO BioPortal [2], and the number of such artifacts is growing rapidly. One of the goals of the Open Biomedical Ontologies (OBO) Foundry [3] is facilitating integration among these diverse resources. Such integration, however, demands considerable effort [4] and differences in format and appearance can only add obstacles to the realization of this task [5]. Heterogeneity derives from the diversity of representation languages and ontology engineering methodologies [6] and it is manifested in the adoption by different communities of Description Logic or First Order Logic formalisms. Diversity also derives from the wide spectrum of syntaxes used to express these formalisms, such as Ontology Web Language (OWL) and OBO, and the commitment of the communities to conceptualist or realism-based philosophical approaches. As diverse as these backgrounds are the naming schemes applied. Even here, in this relatively straightforward area, no convention has been agreed upon or accepted by a wider community [7]. While the other sources of diversity are tremendously complex and challenging, it is our belief that establishing a set of naming conventions is tractable, particularly if we base our conventions on lessons we have drawn from actual experience.

There is, of course, no shortage of naming conventions. One significant barrier is that many of them are domain specific conventions and limited in coverage, and thus are not generally applicable to other domains. For example, the Human Genome Organization (HUGO) nomenclature [8] is restricted to gene names. In other cases conventions refer exclusively to programming languages or to natural language documents [9]. A second impediment is accessibility. While a naming convention may exist, the documentation may be dispersed in multiple documents or document sections, e.g. the BioPAX manual [10], or is primarily commercial in nature, e.g. the ISO standards [11].

A concerted activity involving some members of the Metabolomics Standards Initiative (MSI) [12] and the Proteomics Standards Initiative (PSI) [13] ontology working groups has been directed towards the review of existing documentations in an effort to distill universally valid aspects of these multiple threads. The aim of this analysis is to overcome the present diversity and fragmentation of naming schemes and determine what conventions can be commonly applied in the biological domain. In this article we describe the results of this synthesis: naming conventions that, we believe, could provide robust labels for controlled vocabulary terms and ontology classes.

* To whom correspondence should be addressed.

2 NAMING CONVENTIONS

In this section we rely on the reference terminology proposed by Smith *et al*. [14] to refer to the **representational units** out of which ontologies and similar artifacts are composed, with the expectation that a common lexicon will be agreed upon by a wider community in the future. A **term** is a single word or combination of words. A term used to designate some entity is called a **name**. Entities that represent structures or characteristics in reality and that appear e.g. as general terms in scientific text books are called **universals.** Universals are exemplified, or instantiated, in particulars which we call **instances**.

***Explicit and concise names*:** Each name should be chosen with care and should be meaningful to human readers. In order to be effective and usable, names should be kept short, easy to remember and self-explanatory. Names should be precise, concise and linguistically correct and should conform to the rules of the language used. However, in most cases articles can be omitted.

Context independence: The name should as far as possible capture the intrinsic characteristics of the universal to be represented, rather than extrinsic characteristics or roles an entity may potentially play in a particular context. Names should be meaningful, even when viewed outside the immediate context of the ontology. Therefore one should avoid names that require knowledge of context, either because they are truncated or are colloquialisms. For example, the truncated name 'two dimensional J-resolved' out of context is undecipherable, but if 'two dimensional J-resolved pulse sequence' is used instead, the reader at least will know that it is a 'pulse sequence'. As another example, a NMR instrument is colloquially referred to as 'the magnet', the magnet being an important component of these instruments. However, in other situations 'the magnet' may be a reference to persons who are extremely attractive to others, and one would not want to confuse these two universals. Sometimes hyphens, abbreviations, or acronyms hint at names that have such an omission, For example, 'gene-technology', might conceivably be replaced by the name 'gene modification technology'.

***Compound names*:** To be sufficiently explicit and clear to human readers, it is often necessary to apply composite multi-word names, e.g. 'high resolution-magic angle NMR probe'. For computability, this named entity ideally should capture such qualifiers (or differentia) through additional relationships to other named entities (e.g. to 'high-resolution') to create a computationally interpretable definition. Failing this, whenever names are composed from multiple terms, efforts should be made to use the exact name strings of entities that are defined elsewhere: in this or other ontologies. Developing this habit will make it feasible in the future to retrofit the ontology with these relationships by the simple expedient of string-matching. For example, when used as an affix in a compound term, 'calcium' should always be written out as 'calcium', and never as 'Ca', 'C++', or 'Ca(2+)'. Consistent use of affixes throughout an ontology is a simple expedient to keep developers sane.

Homonyms: Names that are ambiguous, sharing the same spelling but which differ in meaning are best to avoid for obvious reasons. The word 'set', for example, is one of the more ambiguous words in English, having around thirty meanings in English alone. A 'parameter set' could refer to a collection of parameters or to the process of setting the parameters in an instrument. Using multiple homonyms within an ontology creates confusion, since readers may not always realize immediately which is the intended meaning in any particular case.

Consistency of language: Our experience has shown that it is beneficial to be consistent in naming universals in the language of choice. For example, one may choose to use either vernacular English or the Latin form. If a conscious decision is not made to choose one style over the other, then both 'gut' and 'intestinum' are equally valid and ultimately confusing to developers and users alike. The main point is that these decisions should be made in advance, and strictly adhered to in implementations to insure internal consistency. This choice is by no means restrictive for consumers because alternative forms may (and should) be readily included as synonyms. This will also safeguard the ability of search engines to perform efficiently using whichever alternate form is supplied in the query. A solution based on both preferred and alternative names also allows to address the issue of differences in accepted spelling (e.g. between British and US English). Ontology builders may opt to always use the US form 'polymerizing', and provide the UK form 'polymerising' and translations into other languages as a synonym. Likewise, an effective use of synonyms can also address inconsistent translations from other alphabets or character sets. For example, the German "ü" (u-Umlaut) is often unavailable and may be substituted by either "u" or by "ue". A single, most appropriate choice of form should be made for the primary name, and the other forms made available as synonyms. Such consistency and documentation of the chosen conventions helps to avoid irregularities in terminology.

Noun and verb forms: In building an ontology one must be continuously on guard, and recognize precisely what entity one wishes to represent. For example, the name 'NMR measurement' may be slightly ambiguous to a human reader. It might be used to describe a value (an instance) that is an NMR measurement, or it might possibly be construed as referring to the act of taking an NMR measurement. To describe the first usage the noun form is most suitable, while to describe the latter the verb form

applies. In practice, most controlled vocabularies and ontologies refer to universal entities that are nouns (e.g. a person, place, thing, state, quality, or action that a verb acts upon).

Abbreviations and acronyms: These should be resolved in the names and included as synonyms, e.g. high resolution probe' should be used instead of the totally unintuitive 'HRP' acronym or 'high res. probe.' abbreviation. The point at which an abbreviation or acronym becomes more commonly in everyday language than its full name, for example 'LASER, it should be used as the name, and its fully spelled out name made a synonym. Community interaction is necessary to assess frequency usage. Acronyms, which employ expressions with other meanings, should generally be avoided. For instance, the acronym for 'Chronic Olfactory Lung Disorder' is 'COLD', and this is clearly too easily confused with 'cold'.

Singularity: Every name in an ontology refers to a single universal. Hence every name in an ontology should be a singular noun or noun phrase. This rule helps to prevent redundancy and misclassification. To represent an aggregate of protocols one could use 'protocol collection'. An instance of 'protocol' is a protocol and an instance of 'protocol collection' is a collection of protocols. There are other possibilities for indicating collections, such as: aggregate, collective or population, where each may be used according to the case in hand, but used consistently.

Positive names: Names should be formulated to be positive not negative. For instance, one should avoid a name like 'non-separation device' because logically this will include everything in the universe that is a non-separation device: including you, and me, and the bunny-rabbit in the backyard. Negative names do not sufficiently constrain meanings, and are thus strongly discouraged.

Conjunctions: Words that are used to join other words, such as the logical connectives 'and' and 'or' are a red flag. In ontologies built according to the realist perspective, a name that includes a conjunction, such as 'rabbit or whale', is nonsensical because such a universal would never exist. Sometimes hyphens and slashes hint at logical connectives and should to be avoided for this reason.

Taboo words: Words from the representation formalism should not be used within names for representational units. Affixes reflecting epistemological claims do not belong in the names. Since each class 'protocol' implicitly means 'the class protocol', either prefixes or suffixes designating the type of the representational unit, e.g. as in 'protocol class' or 'protocol type', should be avoided. The same applies to suffixes like 'entity' and 'relation'. This is implicit anyway and therefore would be redundant. Metadata should be excluded from term names as far as they can be archived within the expressivity of the representational artifact. If representational units for administrative metadata, e.g. term-versioning, exist, the corresponding data should be factored out of the names and into suitable separate representational units.

Typography: Typographic differences may be computationally irrelevant. If someone queried a database with either "MixedCase", "MIXEDCASE", or "mixedcase", a single record should be returned. However, for legibility and familiarity to humans, case is often a consideration and lower case is recommended. Acronyms, such as 'DNA', that are widely understood by readers can be used as names and should be capitalized. We can relinquish CamelCase because we recommend using a separator, either the space (' ') or underscore ('_') character, to delimit words in compound terms. Using word separators is closest to natural language and does not prevent you to have names like 'CapNMR probe' or 'pH value'. Full stops, exclamations and question marks do not belong in class names. Names should be as computationally pliant as possible. For this reason, subscripts, superscripts or accents should be avoided and Greek symbols should be spelled out (e.g. 'cm^3' should be 'cm3', and 'α' should be 'alpha'). This would ease translation between syntaxes that allow, or disallow a certain formatting.

Registered product, brand and company-names: Proprietary names should be captured as they are. For example, there can be an 'AVANCE II spectrometer', starting with a capital letter, and there can be a CamelCase brand name like 'SampleJet'. Since product names often get very cryptic (e.g. a Bruker NMR magnet has the product name 'US 2'), we recommend a convention that renders these more understandable: Use the company name as prefix, the product name as infix and the product type (superclass) as headword/suffix, e.g. use 'Bruker US 2 NMR magnet' instead of 'US 2'.

3 CONCLUSION

The lesson we have learned from our work is that formulation of universally applicable naming conventions is an exceptionally difficult task, due to the complex dimensionality of the area. Our experience, however, within the PSI and MSI Ontology working groups indicate that the application of common naming guidelines can maximize the communication among geographically distributed developers, simplifies ontology development and helps in subsequent administration tasks. While providing a rigorous and common framework for the developmental process, naming conventions do not place restrictions on the use of less formal terms which can be listed as synonyms.

By increasing the robustness of controlled vocabulary term and class names, we anticipate that a standard naming

convention will assist in the integration, e.g. comparison, alignment and mapping of terminological artifacts. They can facilitate access to ontologies through meta-tools, e.g. PROMPT related ontology merging tools as currently developed by the NCBO BioPortal, by reducing the diversity with which these tools have to deal, thus reducing the burden on tool- and ontology developers alike. Further more explicit naming conventions will ease the use of context-based text-mining procedures used for automatic term-recognition and annotation. On the user side, naming conventions can increase term accessibility and increase exportability and term re-use, reducing development time and costs. Therefore we foresee that such conventions could benefit the overall management of the final resources.

These naming conventions should be seen as an initial step and **straw man proposal**. Currently these are under review by our collaborators in the Ontology for Biomedical Investigation (OBI) project [15]. Although a statistical evaluation on how these conventions would improve ontology editing and integration steps has just started, we hope that the benefit of such common naming conventions are evident and we encourage potentially interested parties to further evaluate and refine these. We are in the process of creating a webpage to gather feedback and suggestions, further details will be available at [16]. Ultimately we hope that in its final form these naming conventions will be widely endorsed by larger umbrella organizations and recognized authorities such as OBO, becoming part of *best practice* design principles akin to those endorsed by the OBO Foundry.

ACKNOWLEDGEMENTS

We kindly acknowledge Robert Stevens, Luisa Montecchi-Palazzo, Frank Gibson, Judith Blake and the members of the OBI working group for their comment and contributions in fruitful discussions. We also gratefully thank the BBSRC e-Science Development Fund (BB/D524283/1), the EU Network of Excellence NuGO (NoE 503630) and the EU Network of Excellence Semantic Interoperability and Data Mining in Biomedicine (NoE 507505).

REFERENCES

1. RG Cote, P Jones, R Apweiler, H Hermjakob: **The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries**. *BMC Bioinformatics* 2006, **7**:97.
2. DL Rubin, SE Lewis, CJ Mungall, S Misra, M Westerfield, M Ashburner, I Sim, CG Chute, H Solbrig, MA Storey, et al: **National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge**. *Omics* 2006, **10**:185-98.
3. B Smith, M Ashburner, J Bard, W Bug, W Ceusters, LJ Goldberg,et.al., **The OBO Foundry: The Coordinated Evolution of Ontologies to Support Biomedical Data Integration**, Nature Biotechnology (in review)
4. K Rickard, J Mejino, RJ Martin, A Agoncillo, C Rosse: **Problems and solutions with integrating terminologies into evolving knowledge bases.** *Medinfo.* 2004, **11**:420-4.
5. S Zhang, O Bodenreider: **Law and order: assessing and enforcing compliance with ontological modeling principles in the Foundational Model of Anatomy**. *Comput Biol Med* 2006, **36**:674-93.
6. AC Yu: **Methods in biomedical ontology**. *J Biomed Inform* 2006, **39**:252-66.
7. O Tuason, L Chen, H Liu, JA Blake, C Friedman: **Biological nomenclatures: a source of lexical knowledge and ambiguity**. *Pac Symp Biocomput* 2004:238-49.
8. MW Wright, EA Bruford: **Human and orthologous gene nomenclature**. *Gene* 2006, **369**:1-6.
9. SH Brown, M Lincoln, S Hardenbrook, ON Petukhova, ST Rosenbloom, P Carpenter, P Elkin: **Derivation and evaluation of a document-naming nomenclature**. *J Am Med Inform Assoc* 2001, **8**:379-90.
10. GD Bader, PC Michael: **BioPAX—biological pathways exchange language, Documentation**. **2007**.
11. **ISO, International Organization for Standardization,** http://www.iso.org/ , URL, last accessed 01.05.07
12. SA Sansone, D Schober, HJ Atherton, O Fiehn, H Jenkins, P Rocca-Serra, et al., **Metabolomics Standards Initiative - Ontology Working Group Work in Progress**, Metabolomics (in press)
13. H Hermjakob: **The HUPO Proteomics Standards Initiative - Overcoming the Fragmentation of Proteomics Data**. *Proteomics* 2006, **6**:34-38.
14. B Smith, W Kusnierczyk, D Schober, W Ceusters: **Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain**. In: *KR-MED 2006*; 2006.
15. **Ontology for Biomedical Investigations (OBI),** http://obi.sourceforge.net/ ,URL, last accessed 01.05.07
16. **The metabolomics standard initiative ontology working group (MSI-OWG)**, http://msi-ontology.sourceforge.net, URL, last accessed 01.05.0

ART: An ontology based tool for the translation of papers into Semantic Web format

Larisa N. Soldatova[1], Colin R. Batchelor[2], Maria Liakata[1], Helen H. Fielding[3], Stuart Lewis[1] and Ross D. King[1]*

1) University of Wales, Aberystwyth, UK; 2) Royal Society of Chemistry Publishing, UK; 3) University College London, UK

ABSTRACT

The paper describes initial work on an ontology based tool, ART, for the semantic annotation of papers stored in digital repositories. ART is intended for the annotation not only of data and metadata about a paper, but also the main elements of the described scientific investigation, such as goals, hypotheses, observations. ART will also be able to aid in the expression of research results directly in a semantic format, through the composition of text using ontology-based templates and stored typical key phrases for the description of basic elements of the research. ART's system design, its functionality, and related projects are discussed. An example annotated paper is presented in order to demonstrate the expected output of the tool.

1 INTRODUCTION

Semantic Web technologies use semantic metadata to improve information retrieval and knowledge representation. Metadata provides semantic clarity, explicitness, and facilitates the reusability of represented information and knowledge [Soldatova & King, 2006a]. Ontology based semantic annotation of papers and data promotes the sharing of research results, and reduces the duplication and loss of knowledge. It also facilitates text mining and knowledge discovery applications.

We are developing ART (an ontology based ARticle preparation Tool), a practical annotation tool which can be used to add value to repository papers and data. ART will generate annotations containing not only metadata about the paper (title, author, etc.), but also generic scientific concepts, such as the type of investigation (theoretical or experimental), its goal, results, the reliability of the results, etc. ART also aims to automate the recognition of those generic concepts in a text. The tool will use a number of Open Biomedical Ontologies (OBO) (http://obi.sourceforge.net/) to find in the text domain specific concepts and to link them to external sources. The result will be an article in OWL-DL (http://www.w3.org/TR/owl-guide/) format that can be submitted to a digital repository along with the original article free-text. The OWL version of the article could then be used for a variety of computational applications (e.g. data mining); or by researchers to check explicit explanations of some terms from the text, or to get more details about experiments.

We also envisage ART being used by authors at the time of manuscript submission to generate annotations, expressed in OWL, that describe the paper and related data. The tool will lead the author through a process where: experimental goals, hypotheses, methodologies, and results, are described and linked to the text and external data files. ART will check that all necessary information is present, and if necessary give examples of formulating concepts.

As a part of the project we will create a digital repository of papers in OWL format. This repository will be an example of an intelligent digital repository. It will be possible to use it for investigation of advanced text mining and knowledge discovery techniques. Since all the papers will be represented in enriched semantic format and directly linked to data sources, new intelligent queries, like: “find evidence for the given hypothesis”, “is the research conclusion consistent with the evidence and the assumptions?” will be possible.

The rest of the paper is organised as follows: section 2 has a brief description of the related work; section 3 presents the ART project, its goals, tasks and principles; section 4 describes the design of the ART tool, main modes and functions; an example of an annotated paper is considered in section 5; whereas section 6 discusses the current state of the project and future plans.

2 RELATED PROJECTS AND LINKS

The ART project aims to build on the experience of eBank. The later project aims to provide a technological solution to the access and curation of digital resources (http://www.ukoln.ac.uk/projects/ebank-uk/). The project is being led by UKOLN in partnership with the Intelligence, Agents & Multimedia Group, Department of Electronics & Computer Science, and the Department of Chemistry, University of Southampton and the Digital Curation Centre (DCC) (http://www.dcc.ac.uk/). eBank/e-Print already uses Dublin Core Metadata (DC) (http://dublincore.org/) to index the e-prints and enable searching. The ART project will contribute to the DCC services. The DCC priorities related to ART are:

- Metadata extraction and curation (investigating standards and tools for the curation of scientific metadata).
- Semantic data curation ('meaning' and 'machine process-ability' foundations of the Semantic Web and Ontological communities).
- Data transformation, integration and publishing (manipulation of data formats, metadata conversion).

The ART project will help to ensure that metadata for various scientific domains are stored and updated in one place. It can provide consistency in managing the digital resources.

ART will build on the experience of semantic enrichment with the RSC's (Royal Society of Chemistry) Project Prospect (http://www.projectprospect.org/). Here journal articles are marked up with chemical structures and domain terms from the IUPAC Gold Book [International Union of Pure and Applied Chemistry, 1997] and terms from the OBO ontologies GO (Gene Ontology) [The Gene Ontology Consortium, 2000], SO (Sequence Ontology) [Eilbeck *et al*., 2005], and CL (Cell Type ontology) [Bard *et al*., 2005]. Mark up of terms from ChEBI (Chemical Entities of Biological Interest) [Matos *et al*., 2006], FIX (ontology of physico-chemical methods and properties) (http://obo.sourceforge.net/cgi-bin/detail.cgi?fix) and REX (ontology of physico-chemical processes) (http://obo.sourceforge.net/cgi-bin/detail.cgi?rex) is in preparation. ART will also incorporate generic scientific concepts from EXPO (Ontology of scientific EXPeriments) [Soldatova & King, 2006b], OBI (Ontology for Biomedical Investigations) (http://obi.sourceforge.net/), ECO (Evidence Code Ontology) (http://obo.sourceforge.net/cgi-bin/detail.cgi?evidence_code).

A similar ontology based format is going to be used for the related ROAD (Robot-generated Open Access Data) project (http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/road.aspx). This project will be investigating the issues involved with the automatic routine deposit of data generated by the Robot Scientist [King *et al*., 2004].

ART aims to advance digital repositories technology, and provide an ontological foundation for manuscript annotation and formalisation.

3 ART PROJECT

The aim of the JISC (Joint Information Systems Committee, UK) funded ART project is to develop an ontology based tool to assist in:

- Translating scientific papers into a format with an explicit semantics.
- Explicit linking of repository papers to data and metadata.
- Creation of an example intelligent digital repository.

We would like to stress that the generic approach and further extensibility of the tool are core principles for the system design. The article translation and preparation tool ART is intended to eventually be a general purpose for any domain where there are available ontologies. The domain independent parts of the system will be fully reusable, and the domain dependent part must be provided with the list of external domain sources. The restriction is that for some domains formalized representations do not exist yet. However ontology development is a rapidly progressing area.

To develop ART we are currently focusing on physical chemistry as the application domain. The rationale for this is that chemistry publications are among the most formalized of all the sciences, and the eBank project has already used chemistry as an exemplar. In addition, physical chemistry papers employ many concepts that have already been formalized in a number of OBO ontologies, i.e. <CHEBI: molecular entity>, <OBI: solid state>, <SBO: concentration>.

4 DESCRIPTION OF THE SYSTEM

The ART system will have two main parts: domain independent and domain dependent (see fig. 1). Each of these will use corresponding ontologies to annotate text. Domain independent sources such as DC Metadata, EXPO, OBI and ECO will be incorporated into the system. The tool will import domain dependent sources after identification of the domain. For physical chemistry these sources are: ChEBI, FIX, REX, and IUPAC. The system can be easily extended by including more internal and/or external ontologies and other sources.

ART will use natural language processing techniques to support both domain-specific and domain-independent mark up. ART itself will use SciXML [Rupp et al., 2006] to represent scientific articles. There exists a framework [Hollingsworth et al., 2005] for converting PDF, which is the most likely format for article submissions, into SciXML, which will provide bibliographic metadata such as titles (<DC: title>), authors (<DC: creator>). The domain-specific mark up of OBO concepts can be partly achieved through named entity recognition. [Batchelor and Corbett, 2007]. The automatic identification of domain-independent concepts is significantly more challenging. However, they frequently occur in well-defined 'zones' within articles [Teufel et al., 1999], and are often introduced by meta-discourse markers or 'cue phrases' articles [Teufel, 1998]. The system will use this information to attempt to identify generic concepts such as <EXPO: goal> or <OBI: conclusion> and will ask users to confirm or to correct the identified concepts in interactive mode. The system will be able to provide a user with explanations why these concepts are necessary, and give definitions and examples. The outcome will be a se-

mantically enriched paper in a text format and OWL paper annotation. If the user wishes, the tool can automatically generate a summary of the article and RSS feed.

ART will also be able to help the author represent his/her research results directly in OWL format. The system will ask for input of the required metadata and data about the research, and will provide examples and explanations where necessary. ART will be designed to assist in composing a paper reporting the results of the investigation. The system will have ontology-based templates of papers. After collecting all the metadata and data about the investigation, the system will propose a paper structure and give examples of key phrases for the description of the main research components.

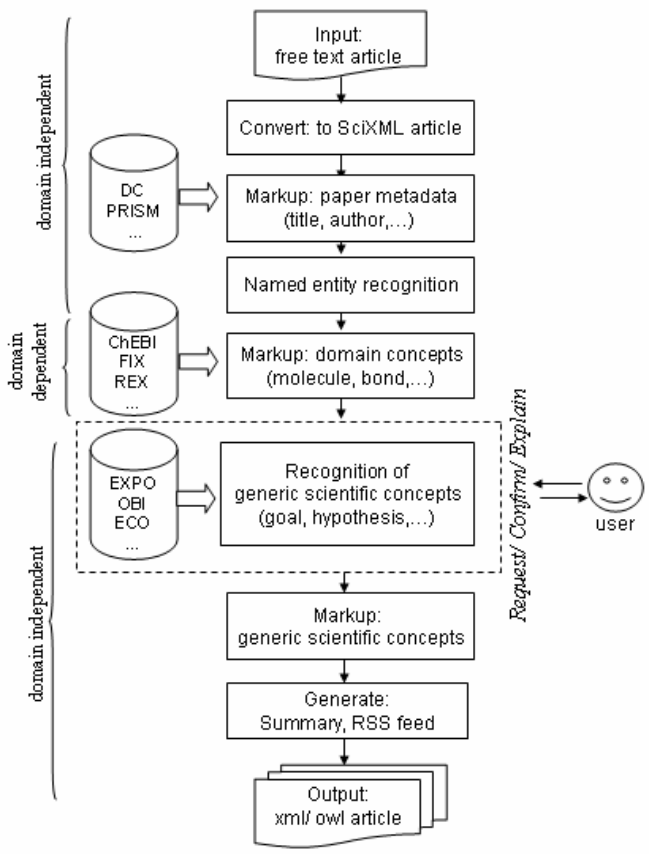


Fig. 1. A flow diagram for the ART system for the physical chemistry domain.

5 EXAMPLE

Let us consider an example physical chemistry paper from the Faraday Discussions [Hiberty *et al*., 2006]. This paper discusses the notion of a charge shift (CS) bond, and is particularly interesting for our purposes because it is a theoretical computational investigation. This area is at present poorly covered by existing ontologies and we will have to extend EXPO to handle these investigations, which are common in many areas of science. Figure 2 shows a fragment of the annotated scientific concepts in the example paper in order to demonstrate the outcome of the ART tool:

<EXPO: investigation> *map to*
<DC: title> *map to*
<OBI: investigation>
The physical origin of large covalent-ionic resonance energies in some two-electron bonds.

<EXPO: goal>
"studying in detail all the aspects of bond formation in a series of molecules that each display a range of bonding features: H_2 and C_2H_6 as members of the classical family of covalent bonds, Cl_2 as a bond exhibiting significant CS character, and the series N_2H_4, H_2O_2, and F_2 as molecules exhibiting increasing CS character from left to right of the periodic table."

<EXPO: object of investigation> *map to*
<OBI: investigation object role>
characteristics of CS bond in H_2, C_2H_6, Cl_2, N_2H_4, H_2O_2, F_2 molecules

<EXPO: method>
valence bond calculation on two levels:
valence-bond-self consistent field (VBSCF)
breathing-orbital valence bond (BOVB)

<EXPO: method assumption> *map to*
<ECO: traceable author statement>
"all the orbitals, including the inactive set, are kept strictly localized, and the ionic components are described as simple closed-shell VB functions" for H_2, C_2H_6, N_2H_4, H_2O_2, molecules".

<EXPO: experimental equipment>
<OBI: software>
XMVB Program
Gaussian 98 series Program

<EXPO: experiment results>
<EXPO: computational data>
Dissociation energy curves of the purely covalent VB structure for H_2, C_2H_6, Cl_2, N_2H_4, H_2O_2, F_2 molecules
(calculated with VBSCF method)
VB-3 three structure ground state for H_2, C_2H_6, Cl_2, N_2H_4, H_2O_2, F_2 molecules
(calculated with BOVB method)

<EXPO: conclusion> *map to*
<OBI: conclusion>
"CS bonding is characterised by the following features: (i) a covalent dissociation curve with a shallow minimum situated at long interatomic distance, or even a fully repulsive covalent bond; (ii) a large covalent-ionic resonance energy RE_{cs} that is responsible for the major part, or even for totality, of the bonding energy."

Fig. 2. A fragment of the annotated article in a text format.

The system will provide mappings between the incorporated representations. The same element in the text can be linked to a number of internal and/or external resources. For instance in the example considered, the title of the paper is marked as <EXPO: investigation> which corresponds to the class <OBI: investigation> and to the term <DC: title>. These mappings are not equivalent, but ART will contain formalized description of the semantics involved.

Some generic scientific concepts can be automatically recognized by the system using the cue phrases. For instance the phrase in the considered paper "this paper is aimed at…" indicates the goal of the investigation and the metadata about the text structure <section: conclusion> points out to the list of the conclusions. The system will be able to identify more cue phrases for the recognition of more elaborate concepts. However, it will not always be possible to automatically identify concepts in free text. However the system 'knows' what should be in a scientific paper and can ask a user, for example "What are the results of the investigation?" The user can then indicate them in the text or input them directly.

We expect that the ART tool can also help with automated ontology construction. For example in the paper considered above, the new concept <CS bond> is discussed, and features of such bonds are investigated. This concept is absent from all existing ontologies. The system could collect such missing terms and then ontology developers would consider them for inclusion to the corresponding ontology.

6 DISCUSSION

The potential users of the ART tool are: curators of digital repositories who would like to semantically enhance the papers stored in the repositories; researchers who would like to represent their research results in semantic machine readable format for various computer applications; publishers and reviewers. Reviewing is a time-consuming process and any tool to facilitate the process will be of significant value.

ART is an ongoing project and the ART tool is in a development stage. The authors would like to use this opportunity and invite potential users for feedback on the proposed functionality of the system.

ACKNOWLEDGEMENTS

The work was funded by JISC (Joint Information Systems Committee, UK).

REFERENCES

Batchelor, C.R. and Corbett, P.T. (2007) Semantic enrichment of journal articles using chemical NER. *In proc. of ACL* (in press).

Hiberty, Ph.C., Ramozzi, R., Song, L., *et al*. (2006) The physical origin of large covalent-ionic resonance energies in some two-electron bonds. *Faraday Discuss*., **135**, 261-272.

King, R. D., Whelan, K.E., Jones, M.F., Reiser, P.G.K, Bryant, C.H. (2004) Functional Genomics Hypothesis Generation by a Robot Scientist. *Nature*, **427/6971**, 247-252.

Soldatova, L.N. and King R.D. (2006) Ontology Engineering for Biological Applications. *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences.* Christopher J.O. Baker and Kei-Hoi Cheung (Eds). Springer, NY., 121-137.

Soldatova, L.N. and King, R.D. (2006) An Ontology of Scientific Experiments. *Journal of the Royal Society Interface* **3/11**, 795-803.

Teufel, S., Carletta, J., Moens, M. (1999) An annotation scheme for discourse-level argumentation in research articles. *In Proc. of EACL*.

Teufel, S. (1998) Meta-discourse markers and problem-structuring in scientific articles, *Workshop on Discourse Structure and Discourse Markers, ACL*, Montreal.

International Union of Pure and Applied Chemistry (1997) Compendium of Chemical Terminology. 2nd edition. Blackwell Science, Oxford.

The Gene Ontology Consortium (2000) Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, **25**, 25-29.

Eilbeck K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., Ashburner, M. (2005) The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology* **6**, R44.

Bard, J., Rhee, S.Y. and Ashburner, M. (2005) An ontology for cell types. *Genome Biology* **6**, R21.

Matos, P., Ennis, M., Darsow, M., Guedj, M., Degtyarenko K. and Apweiler R. (2006) ChEBI - Chemical Entities of Biological Interest. *Nucleic Acids Research*, Database Summary paper 646.

Rupp, C.J., Copestake, A., Teufel, S. and Waldron, B. (2006) Flexible Interfaces in the Application of Language Technology to an eScience Corpus. *In Proc. of the 4th UK E-Science All Hands Meeting*. Nottingham, UK.

Hollingsworth W., Lewin I. and Tidhar D. (2005) Retrieving Hierarchical Text Structure from Typeset Scientific Articles – a Prerequisite for E-Science Text Mining. *In Proc. of the 4th UK E-Science All Hands Meeting*, Nottingham, UK, 267-273.

Evaluating GO-based Semantic Similarity Measures

Catia Pesquita[*], *Daniel Faria, Hugo Bastos, André O. Falcão and Francisco M. Couto*

University of Lisbon, Department of Informatics, Campo Grande, Lisboa – Portugal

ABSTRACT

Motivation: While several efforts have been made in measuring GO-based protein semantic similarity, it is still unclear which are the best approaches to measure it and furthermore whether electronic annotations should be used.

Results: We studied the behaviour of 8 distinct semantic similarity measures as function of sequence similarity with and without electronic annotations. We found that 5 of these measures shared a cumulative normal distribution pattern, which is likely inherent to the relation between functional and sequence similarity. We also present a novel graph-based measure for protein semantic similarity, which produced better results than the other measures studied.

1 INTRODUCTION

Since its foundation, the Gene Ontology (GO) has had a high impact in gene-product annotation, leading to its adoption by an increasing number of sequence databases. This fact, combined with the quality and structure that GO adds to annotation, has enabled its use as a background for functional comparison of gene-products. This type of comparison, called semantic similarity, is usually based on comparing the GO terms to which gene-products are annotated.

To calculate protein semantic similarity, Lord *et al.* (2003[a,b]) used three semantic similarity measures developed for WordNet and based on the notion of information content (*IC*): Resnik´s (1999), Lin´s (1998), and Jiang and Conrath´s (1999). The authors identified a correlation between semantic similarity and sequence similarity, which was stronger in the GO *molecular function* aspect. However, as these three measures were developed for comparing single terms in a hierarchy, some issues arise when applying them to GO-based protein similarity.

One issue is that GO terms can have several disjoint common ancestors. Lord *et al.* (2003[a]) dealt with this by considering only the most informative common ancestor between two terms, whereas Couto *et al.* (2005) proposed the GraSM approach, to account for all disjoint common ancestors.

Another issue is that proteins can be annotated with several GO terms, so computing the semantic similarity between two proteins requires a way of combining the semantic similarity between their terms. To address this, Lord *et al.* (2003[b]) used the arithmetic average of all term pairs, Sevilla *et al.* (2005) opted for their maximum, and Schlicker *et al.* (2006) introduced a composite average where only the best matching term pairs are used.

A different, graph-based approach was proposed by Gentleman (2005), who developed two measures for GO-based protein semantic similarity, both comparing the portion of the GO-graph shared by a pair of proteins.

Despite several studies, it is still unclear which are the best measures and/or approaches to calculate protein semantic similarity, and whether electronic annotations should be used for this purpose or ignored.

In this paper, we investigate the behaviour of several semantic similarity measures as function of sequence similarity, using both the whole annotation space and the subset of non-electronic annotations. We also introduce a novel graph-based measure for protein semantic similarity and compare its performance with that of the other measures.

2 METHODS

2.1 Semantic similarity measures

We used three term semantic similarity measures: Resnik´s (1999), Lin´s (1998), and Jiang and Conrath´s (1999); and combined them with two different approaches to compute protein similarity: the average and the best-match average (*BMA*). The former was applied as described by Lord *et al.* (2003[b]), and the latter was applied as described by Schlicker *et al.* (2006) except that only *molecular function* GO terms are being used. IC and similarity measures were calculated as previously described (Faria *et. al*, 2007).

We also use two graph-based similarity measures: *simUI* (Gentleman, 2005) and the novel *simGIC* (for Graph Information Content). *simUI* calculates similarity as the number of GO terms shared by two proteins divided by the number of GO terms they have together. *simGIC* is an expansion of *simUI* where instead of counting the terms we sum their *IC*. For two proteins *A* and *B* with terms *t*, *simGIC* is given by:

$$simGIC(A,B) = \frac{\sum_{t \in A \cap B} IC(t)}{\sum_{t \in A \cup B} IC(t)} \quad (1)$$

2.2 Dataset

The full protein dataset used was a subset of 22,067 proteins from the Swiss-Prot database, having at least one *molecular function* GO term of *IC* 0.65 or higher. The goal was to have

[*] To whom correspondence should be addressed.

a dataset that was well characterized functionally but large enough to provide meaningful results.

An all-against-all BLAST search was performed, considering a threshold *e*-value of 10^{-4}. For each protein pair {A,B} with A?B, sequence similarity was defined as:

$$simSeq(A,B) = \log_{10}\left(AVG\left(B_{score}(A,B), B_{score}(B,A)\right)\right) \quad (2)$$

where B_{score} is BLAST's bit-score (which is not symmetric).

For the resulting 618,146 protein pairs, functional semantic similarity was computed with the measures described in 2.1, using *molecular function* GO terms.

A second dataset of proteins with only non-electronic GO annotations was also used. It contained 8,377 proteins which lead to 49,480 protein pairs.

The source data came from the UniProt database (release 2007-02-20), the GO database (release 2007-02) and the GOA-UniProt dataset (release 2007-02).

2.3 Semantic vs. Sequence Similarity

Due to large size and high dispersion of the semantic vs. sequence similarity raw data, discrete intervals of sequence similarity were taken, and average similarity values were calculated for each interval. Intervals had constant size except where the number of protein pairs in an interval was too small (under 200). The procedure was applied to all measures for both datasets.

A cumulative normal distribution curve was fitted to the discrete averaged semantic similarity vs. sequence similarity data. Non-linear regression was done applying the Newton optimization algorithm to solve the least squares method. Besides the normal parameters (mean and standard deviation), two additional parameters were required: a multiplicative scale factor and an additive translation factor (Figure 1).

3 RESULTS AND DISCUSSION

The measures using the average approach (Resnik's, Lin's, and Jiang and Conrath's) were clearly those which performed worse, being the only measures whose behaviour was not monotonically increasing (Figure 1F). This is not unexpected since this approach is biased, penalizing protein pairs which have several distinct functional aspects in common. In fact these measures only became decreasing for high sequence similarity scores, which correspond to protein pairs of larger sequence size, likely to have more than one functional aspect.

The remaining five measures (those with the *BMA* approach, *simUI* and *simGIC*) all showed a crescent behaviour with a similar topology (Figure 1A-E). We found that topology to be well modelled by a scaled cumulative normal distribution (Table 1) despite the higher dispersion visible for the non-electronic dataset, likely due to its smaller size.

What is most striking in the fitted curves is that the parameters for the normal distribution (mean and standard deviation) are nearly identical between measures, within each dataset (Table 1). Considering that these are five distinct measures, one of which (*simUI*) doesn't even rely on the notion of IC, we postulate that a normal distribution curve with these parameters is characteristic of the GO term *molecular function* annotations themselves. What this means is that the ability of *molecular function* GO terms to distinguish different levels of sequence similarity is given by a normal probability density function, which is not altogether surprising. It reflects the fact that sequence pairs with either very low or very high sequence similarity are hard to distinguish functionally, being nearly all unrelated or identical respectively.

Table 1. Regression parameters for the fitted normal distribution curves

	Measure	Regression Parameters				Scaled
		Mean	stdev[1]	scale[2]	trans[3]	Residual[4]
full	simUI	2,2	0,25	0,45	0,48	0,0026
	simGIC	2,2	0,27	0,65	0,26	0,0026
	BMA-R	2,2	0,27	0,51	0,43	0,0029
	BMA-L	2,2	0,27	0,41	0,53	0,0025
	BMA-JC	2,2	0,27	0,25	0,73	0,0029
non-electronic	simUI	2,4	0,31	0,43	0,37	0,0084
	simGIC	2,4	0,30	0,56	0,21	0,0078
	BMA-R	2,4	0,30	0,51	0,21	0,0084
	BMA-L	2,4	0,30	0,46	0,30	0,0091
	BMA-JC	2,4	0,29	0,28	0,58	0,0091

1 – standard deviation; 2 – multiplicative scale factor; 3 – additive translation factor; 4 – average residual by point divided by the scale factor, to dilute scale differences.

By analyzing the regression parameters (Table 1), we see that all these measures are capturing the normal behaviour with similar accuracy (they have similar scaled residuals within each dataset) but with different resolutions, as shown by the different scale factors (see Figure 1).

It is important to note that, while the fitted curves are isomorphic (they are inter-convertible through a linear transformation using the scale and translation factors), the actual semantic similarity measures are not: only their average behaviour is modelled by the curves. One example of this is that all 5 measures produce an equal value (of 1) if two proteins have exactly the same GO terms, of which there are occurrences in several intervals of sequence similarity. Such equality would not be maintained when applying the isomorphism between the measures' curves.

The choice of the best similarity measure therefore should fall to the measure which has the highest resolution, since on average that measure translates differences in annotation to higher differences in semantic similarity, allowing their clearer perception. In this context, the results support the choice of the novel *simGIC* measure, which showed a higher resolution than the other measures with both datasets.

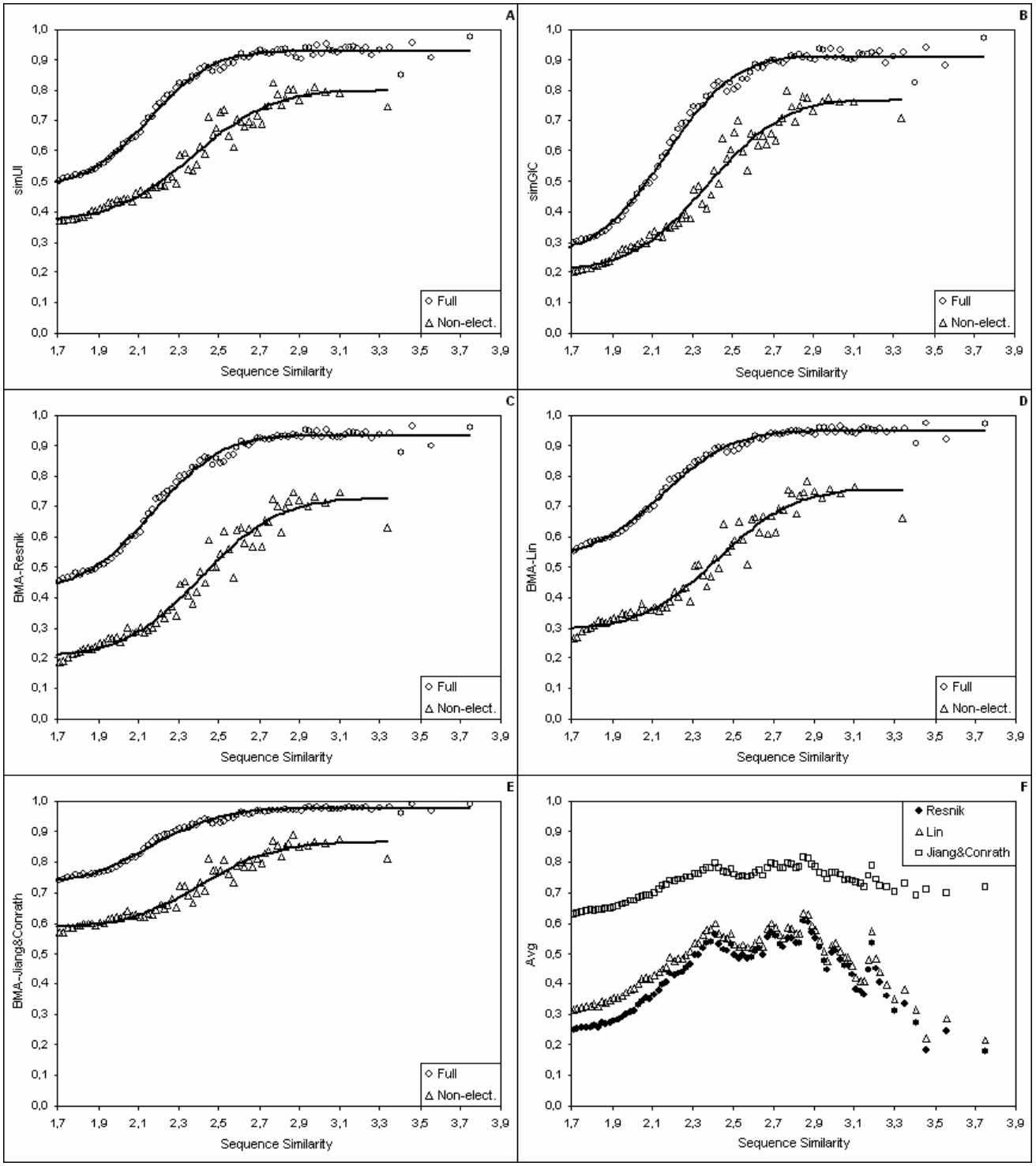


Fig. 1. Semantic similarity vs. sequence similarity for the 8 measures tested, with both full and non-electronic datasets. **A** – *simUI* measure; **B** – simGIC measure; **C** – Resnik's measure with BMA approach; **D** – Lin's measure with BMA approach; **E** – Jiang and Conrath's measure with BMA approach; **F** – Resnik's, Lin's and Jiang and Conrath's measures with the average approach (full dataset only); lines in **A**-**E** correspond to fitted cumulative normal distribution curves. In addition to mean and standard deviation, which determine the inflexion point and width of the curve respectively, two parameters were used to fit the curves: a multiplicative scale parameter to account for the measures not covering the whole 0-1 scale, and an additive translation parameter to account for their minimum value being greater than 0.

However, the only measure which showed a clearly low resolution was Jiang and Conrath's measure. The remaining measures have a resolution not much below that of *simGIC*, with Resnik's measure being second best.

As for the differences in the normal distribution parameters between the two datasets, they reflect the fact that the non-electronic annotation space is different from the full space. For instance, the average number of annotations per protein is smaller in the non-electronic dataset than in the full one (4.8 and 5.5 respectively). Also relevant is the fact that there are much less proteins with non-electronic annotations (8% of the full set), and these could not be representative of the whole protein similarity space.

Despite these differences, the fact remains that the behaviour of the two datasets is similar, which suggests that electronic annotations can not only be reliably used in semantic similarity calculations, but also improve their precision by providing a richer annotation space.

4 CONCLUSIONS

We studied the averaged behaviour of several distinct semantic similarity measures as function of sequence similarity, uncovering an underlying normal distribution-like pattern with constant shape parameters (mean and standard deviation). We postulate that this pattern is characteristic of the variation of functional similarity (as measured by GO *molecular function* annotations) with sequence similarity.

We developed a novel graph-based semantic similarity measure for proteins, which performed better than the remaining measures by translating sequence similarity into a greater coverage of the semantic similarity scale.

We also compared the performance of the similarity measures with and without electronic annotations, concluding that electronic annotations do not significantly affect the behaviour of the similarly measures, and actually increase their precision. While they may lack the reliability of curated annotations, electronic annotations are the present and future of bioinformatics, constituting an increasingly important portion of the annotation space (currently amounting to 97%). What is more, their precision is improving, with values of 91-100% having been reported (Camon *et al.*, 2005).

Future work will include comparing semantic similarity with other aspects, such as protein families (Pfam) and Enzyme Commission classes, as well as using a sequence similarity measure independent of sequence length. We will also investigate other semantic similarity measures, such as the GraSM approach.

ACKNOWLEDGEMENTS

This work was partially supported by the Portuguese 'Fundação para a Ciência e Tecnologia' with the grant ref. SFRH/BD/29797/2006.

REFERENCES

Camon, E., Magrane, M., Barrell, D., *et al.* (2004) The Gene Ontology Annotation (GOA) database: sharing knowledge in UniProt with Gene Ontology. *Nucleic Acids Research* 32 D262.

Camon, E., Barrell, D., Dimmer, E. C., *et al.* (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* 6(Suppl 1):S17.

Couto, F. M., Silva, M. J., and Coutinho, P. M. (2005) Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. *In Proc. of the ACM Conference in Information and Knowledge Management as a short paper.*

Devos, D., and Valencia, A. (2000) Practical limits of function prediction. *Proteins: Structure, Function, and Genetics* **41**, 98–107.

Faria, D., Pesquita, C., Couto, F. M. and Falcao, A. O. (2007) ProteInOn: A Web Tool for Protein Semantic Similarity, *DI/FCUL TR 07-06*, Dpt. Informatics, Univ. Lisbon.

Gentleman, R. (2005) Visualizing and Distances Using GO, Retrieved Jan. 10th, 2007: http://bioconductor.org/packages/2.0/bioc/vignettes/GOstats/inst/doc/GOvis.pdf

GO-Consortium. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32**, Database issue, D258–D261.

Jiang, J., and Conrath, D. (1997) Semantic similaritybased on corpus statistics and lexical taxonomy. *In Proc. of the 10th International Conference on Research on Computational Linguistics.*

Lin, D. (1998) An information-theoretic definition of similarity. *In Proc. of the 15th International Conference on Machine Learning*.

Lord, P., Stevens, R., Brass, A., and Goble, C. (2003)[a] Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**, 10, 1275–1283.

Lord, P., Stevens, R., Brass, A., and Goble, C. (2003)[b] Semantic similarity measures as tools for exploring the gene ontology. *In Proc. of the 8th Pacific Symposium on Biocomputing.*

Resnik, P. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Artificial Intelligence Research* **11**, 95–130.

Schlicker, A., Domingues, F. S., Rahnenfhrer, J., and Lengauer, T. (2006) A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics* **7**, 302.

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martnez-Cruz, L. A., Corrales, F. J., and Rubio, A. (2005) Correlation between gene expression and GO semantic similarity. *In IEEE/ACM Transactions on Computational Biology and Bioinformatics.*

Bodenreider, O., Stevens, R., (2006) Bio-ontologies: current trends and future directions. *Briefings In Bioinformatics.* **7**. No 3. 256-274

Wu, C., Apweiler, R., Bairoch, A., *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* D187–D191.

OntoDas - integrating DAS with ontology-based queries

Kieran O'Neill[*1], *Anita Schwegmann*[1,2], *Rafael Jimenez*[1,3], *Dan Jacobson*[1] *and Alexander Garcia*[1,4]

[1] *Central Node, National Bioinformatics Network, Cape Town, 7405, South Africa*
[2] *Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, South Africa*
[3] *Proteomics Services Group, European Bioinformatics Institute, Hinxton, United Kingdom*
[4] *Centro Internacional de Agricultura Tropical, Cali, Colombia*

ABSTRACT

We present OntoDas, an extension to the Dasty2 DAS client which visually enables the construction of ontology-based queries for retrieving sets of related proteins. OntoDas is based on AJAX and makes extensive use of web services, including Distributed Annotation System (DAS), Ontology Lookup Service (OLS) and a custom-built web service for executing the queries. By integrating multiple web services and making use of AJAX technology, we have created a tool that provides a unified view for creating dynamic, visual, ontology-based queries that is easy to use and install in other web-based systems. OntoDas facilitates the discovery of sets of proteins annotated with specific sets of Gene Ontology (GO) terms. The tool makes use of query previews to enable users to rapidly find results and to explore the query space. Other information visualization techniques are used to provide cognitive support to biologists when building queries.

1 INTRODUCTION

The ability to collect, store, and manage data is increasing quickly, but our ability to understand it remains constant. In an attempt to gain better understanding of data, fields such as information visualization, data mining and graphic design are employed (Fry 2004). Ontologies provide a means of integrating biological information from diverse sources. Gene Ontology is a controlled vocabulary consisting of three distinct ontologies intended to describe the roles of genes and gene products in any organism (Ashburner et al, 2000). GO has been used to annotate gene products in numerous model organisms, and thus can and has been used as a platform for cross-database queries. One useful type of query which can be performed against GO and other ontologies is to find sets of gene products sharing annotations. An example of such a query might be:

Retrieve gene products that *participate in* **blood coagulation** and *are located in* the **extracellular space** and *have* **protease inhibitor activity**.

These queries can be executed using scripts, such as the Perl API to the GO MySQL database. However, both the task of constructing queries in a scripting language and the task of making sense of the results place high cognitive load on the user. Cognitive support, defined as the augmentation of human cognition using artifacts external to the human mind (Walenstien, 2002), would potentially be useful to biologists. Information visualisation aims to find computer-supported, interactive, visual representations of abstract data to amplify cognition (Card et al, 1999), and could provide this support.

GViewer (Shimoyama et al, 2005) is an example of a visual tool for executing these kinds of queries. It is a powerful tool for querying rat genome information using multiple ontologies, and visualizing the location of the results on a graphical representation of the rat genome. However, the interface for constructing queries is form-based, and requires the ontology terms to be known in advance by the user. This presents two problems: Firstly, there is no guarantee that the text entered will actually match an ontology term, and secondly that, even if it does match, the combination of terms will return any hits (Plaisant et al, 1999). This can be overcome by providing a query building interface, in which query previews (summarized previews of query results) are provided (Plaisant et al, 1999). In a query preview interface, options for modifying the query are provided, but constrained to valid terms in the ontology, and to those terms which, when added to the query, return some results.

For instance, for the term *"blood coagulation"*, in the April 2007 release of the GO MySQL database, only 1837 (of 24021) GO terms can be combined with this term to produce two-term AND queries which return results. For a more complex query, such as finding genes annotated with the terms "*blood coagulation*", "*protease inhibitor activity*" and "*extracellular space*", only 196 terms can be added to the query and still produce non-empty sets of gene products. By only displaying the 196 combinable terms, rather than all 24019 valid GO terms, the search space to find a term to refine the query with is reduced by two orders of magnitude,

* To whom correspondence should be addressed.

and the user is guaranteed that the term they choose will produce results.

Another feature which could aid users when constructing these kinds of queries is to enable them to build the query by using genes or gene products. For example, if the terms used to annotate a single protein are displayed, users may select combinations of those terms to specify queries to find related proteins. A tool such as Dasty2 for viewing the details for a single protein could be extended to be an entry point for an ontology based query system, and be used to view details on results returned.

OntoDas is a tool for visually constructing ontology based queries using GO. It makes use of Dasty2 as an entry point to query construction, and as a viewer for details on individual results. OntoDas employs information visualization techniques to assist biologists in creating and exploring queries to find sets of related gene products.

2 DESIGN OF ONTODAS

OntoDas was developed as an extension to Dasty2, an AJAX (asynchronous JavaScript and XML)-based DAS client for viewing the sequence annotations of a single protein. As such, it makes extensive use of the existing paradigm and visual style of Dasty2, and employs Dasty2 as an entry and exit point to queries. In order to enable the lookup of ontology terms annotated to the protein being viewed, and to construct a query from a combination of these terms a panel was added to Dasty2. This takes the user to the main OntoDas view, which provides the user with the ability to retrieve query results, and to modify the query in useful ways that will produce nonempty result sets.

2.1 An Example

Figure 1 illustrates three main steps when constructing a query, starting from the Dasty2 view. During the first step, the terms annotated to an instance of the protein *tissue factor pathway inhibitor 2*, are displayed. When selecting the terms *"blood coagulation"*, *"extracellular space"* and *"protease inhibitor activity"*, the query is being built in an executable way. By using terms annotated to at least one gene product, the user is guaranteed that this will return results, in this case 14.

During the second step, the user examines *"blood coagulation"*. The parent child, neighbouring and lexically similar terms that produce results are displayed along with the size of the result set for the potential new query. The user decides to substitute with *"wound healing"*. Making this change refreshes the interface.

During the third step, the user decides to narrow down the query by adding another term. OntoDas provides the user with a list of all terms that can be added to the query so non-empty results sets are produced. In order to assist the user when finding those terms of interest, the list can be grouped by ontology, sorted alphabetically or by the number of genes returned. It is also possible to display the list as a tree, using a file explorer style representation, similar to that used by AmiGO. The user hovers the mouse over the term *"proteinaceous extracellular matrix"*, and is presented with the full definition of the term as a tooltip. At any point, the user can view the result set, and view details on an individual gene product in Dasty2.

2.2 Visual Considerations

OntoDas aims to provide as much useful information as possible, without overwhelming the user. Queries are framed in natural language so as to make them easier to understand. This representation uses phrases based upon the formal relations laid out by Smith et al (Smith et al, 2005). The phrase "participate in" is used to refer to terms from the biological process ontology, to express the relation "has_participant". The phrase "are located in" refers to terms from the cellular component ontology, in order to express the relation "has_location". For terms from the molecular function ontology, since no relation had been agreed on, the phrase "have" is used, expressing a simple, non-specific property relation. Examples are shown in the introduction and in the screenshots in figure 1.

Additionally, "information scent" has been provided to guide users in choosing terms: Information scent is defined as *"the (imperfect) perception of the value, cost, or access path of information sources obtained from proximal cues"* (Pirolli and Card, 1999), and has been shown to be a highly important factor when finding terms in an ontology (Pirolli et al, 2003). OntoDas provides information by displaying complete term definitions when the user hovers the cursor over a term, and by showing the size of potential result sets. Full term definitions are used because terms themselves are often ambiguous, whereas their natural language definitions are more likely to ensure terms' appropriate interpretation (Bodenreider and Stevens, 2006). Finally, grouping and sorting of terms is provided to reduce the cognitive load on users in finding specific terms of interest (Card et al, 1999).

2.3 Technical Aspects

OntoDas uses other web services in addition to DAS, specifically a custom-made web service acting as a query execution engine, and Ontology Lookup Service (OLS) (Côté et al, 2006), a powerful support vector machine-based text search tool for finding ontology terms. OLS provides the

lexically similar neighbours which are displayed when a user is modifying a query term. The custom web service currently works with the GO relational database, using a small Python script front-end. It is designed to be modular, and potentially to use the DGB graph database developed by the NBN (Otgaar et al, 2005). By using multiple web services, powerful functionality can be provided from multiple sources in a single view. Similarly, by using JavaScript OntoDas enables visual manipulations without round trips to the web server, thereby improving the responsivity of the interface.

3 CONCLUSIONS AND FUTURE WORK

Previewing queries constrains query spaces. The use of visualisation techniques, as well as the use of informative cues such as query summaries and natural language ontology definitions support users when constructing ontology-based queries. OntoDas makes use of these techniques, together with semantic web technologies to provide visual support in the construction of complex biological queries.

In the future, it is intended that OntoDas will make greater use of the DAS and OLS web services. DAS could be better integrated into the OntoDas query interface, so that result sets of gene products could be narrowed further based on the location and type of protein or other sequence annotations within the results. OLS could be used for finding a broader range of lexically related terms when substituting a term in a query.

A major issue to be considered is how to enable the construction of queries containing additional operators such as OR and NOT, as currently only AND is supported. However, this could require substantial changes to the interface. OntoDas could also be extended to work with more ontologies annotating the same set of gene products, such as those used by the rat and mouse genome projects. This would require the extension of the web service backend. This functionality would enable the construction of more sophisticated queries.

As OntoDas is at an early stage of development, it is likely that unforeseen usability problems will have to be dealt with. Usability testing could help to discover some of the issues that biologists have when using the interface.

ACKNOWLEDGEMENTS

Support for this work was provided by student scholarships from the National Bioinformatics Network and the National Research Foundation (of South Africa).

REFERENCES

Ashburner, A.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M. and Sherlock, G. (2000), Gene Ontology: tool for the unification of biology, *Nature Genetics* **25**, 25 - 29

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

Card, S.; Mackinlay, J. and Shneiderman, B. (1999) *Readings in Information Visualization: Using Vision to Think,* Morgan Kaufmann

Côté, R.; Jones, P.; Apweiler, R. and Hermjakob, H. (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries, *BMC Bioinformatics,* **2006**, 97

Fry, B. J. (2004), *Computational Information Design*, PhD thesis, Massachusetts Institute of Technology, School of Architecture

Otgaar, D.; Dominy, J.; Maclear, A.; Gamieldien, J.; Martinez, F. and Jacobson, D. (2006) DigraBase: A Graph-theoretic Framework for Semantic Integration of Biological Data. Poster. *Bio-Ontologies SIG, ISMB 2006*

Pirolli, P. and Card, S. (1999) Information foraging, *Psychological Review,* **106** , 643-675

Pirolli, P.; Card, S.K. and Wege, M.M.V.D. (2003) The effects of information scent on visual search in the hyperbolic tree browser *ACM Trans. Comput.-Hum. Interact., ACM Press,* **10** , 20-53

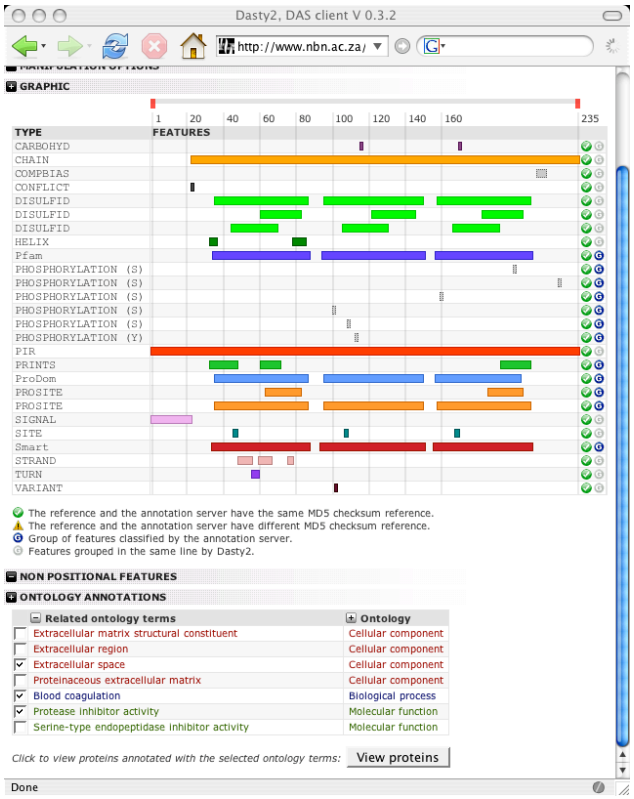
Plaisant, C.; Shneiderman, B.; Doan, K. and Bruns, T. (1999) Interface and data architecture for query preview in networked information systems *ACM Trans. Inf. Syst.*, **17** , 320-341

Shimoyama, M.; Petri, V.; Pasko, D.; Bromberg, S.; Wu, W.; Chen, J.; Nenasheva, N.; Twigger, S. and Jacob, H. (2005), Using Multiple Ontologies to Integrate Complex Biological Data, *Bio-Ontologies SIG, ISMB 2005*

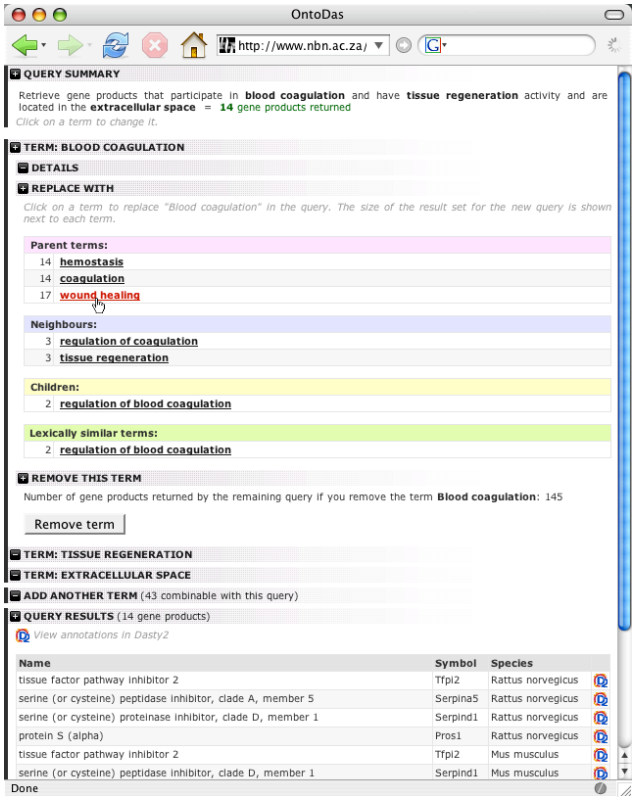
Smith, B.; Ceusters, W.; Klagges, B.; Köhler, J.; Kumar, A.; Lomax, J.; Mungall, C.; Neuhaus, F.; Rector, A. and Rosse, C. (2005) Relations in biomedical ontologies *Genome Biology, **6:R46***

Walenstein, A. (2002) *Cognitive Support in Software Engineering Tools: A Distributed Cognition Framework*, PhD thesis, Simon Fraser University

Step 1:



Step 2:



Step 3

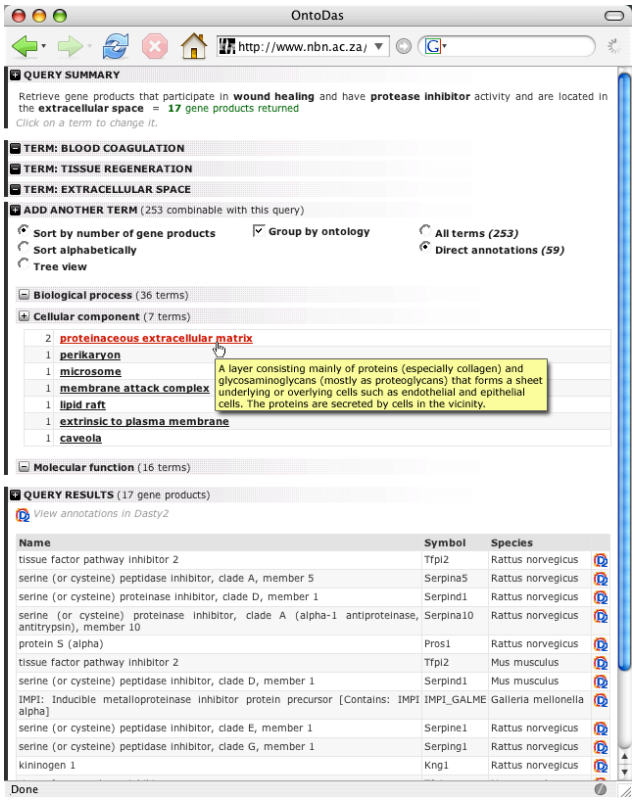


Fig. 1. An illustration of the example used throughout section 2.1 Step 1: The user views the ontology annotations for a protein, choosing three to form a query from. Step 2: The user chooses the term *"wound healing"* to substitute for *"blood coagulation"*, Step 3: The user decides to add the term *"proteinaceous extracellular matrix"* to the query. At all points the results are available in a collapsible tab. Whenever ontology terms are displayed, the full definition can be obtained by hovering the cursor over the term.

Facilitating the development of controlled vocabularies for metabolomics with text mining

Irena Spasic[1,], Daniel Schober[2], Susanna-Assunta Sansone[2], Dietrich Rebholz-Schuhmann[2], Douglas B. Kell[1], Norman Paton[1] and the MSI Ontology Working Group Members*

[1]*Manchester Centre for Integrative Systems Biology, The University of Manchester, 131 Princess Street, Manchester M1 7ND, UK*
[2]*The European Bioinformatics Institute, EMBL Outstation – Hinxton, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK*

ABSTRACT

Many bioinformatics applications rely on controlled vocabularies or ontologies to consistently interpret and seamlessly integrate information scattered across disparate public resources. Experimental data sets from metabolomics studies need to be integrated with one another, but also with data produced by other types of omics studies in the spirit of systems biology, hence the pressing need for vocabularies and ontologies in metabolomics. Here we describe the development of controlled vocabularies for metabolomics investigations. Manual term acquisition approaches are time-consuming, labour-intensive and error-prone, especially in a rapidly developing domain such as metabolomics, where new analytical techniques emerge regularly, thus often compelling domain experts to use non-standardised terms. We suggest a text mining method for efficient corpus-based term acquisition as a way of rapidly expanding a set of controlled vocabularies with the terms used in the scientific literature.

1 INTRODUCTION

The lack of suitable means of formally describing the semantic aspects of omics investigations presents challenges to the effective exchange of information between biologists (Quackenbush, 2004; Field and Sansone, 2006; Shulaev, 2006). The inherent imprecision of free-text descriptions of experimental procedures hinders computational approaches to the interpretation of experimental results. Representational artefacts such as controlled vocabularies and/or ontologies can be used to add an interpretive annotation layer to the textual information (Schulze-Kremer, 1998; Spasic et al., 2005; Cimino and Zhu, 2006). A controlled vocabulary (CV) is a structured set of terms[1] and definitions agreed by an authority or a community. Ontologies are semantically rich representations which include CV terms to refer to concepts at the linguistic level and logical statements to characterise the ways in which these concepts are interrelated (Smith, 2006). Many scientific communities, including those operating in the metabolomics domain (Castle et al., 2006), have started developing an appropriate set of ontologies for the primary purpose of data annotation (Bodenreider and Stevens, 2006). The Metabolomics Standards Initiative (MSI, 2007) Ontology Working Group (MSI OWG, 2007) has been established in an effort to develop a common semantic framework for metabolomics studies so to describe the experimental process consistently and to ensure meaningful and unambiguous data exchange. The MSI OWG has been conceived as a 'single point of focus' for communities where independent activities – to develop terminologies and databases for metabolomics investigations – are underway. Interoperability among these systems is the key driving force behind this endeavour. In this article we describe the text mining method we use to rapidly expand the set of MSI CVs with the terms acquired from the scientific literature.

2 BACKGROUND

The first step in developing an ontology involves the identification of its purpose and scope, followed by the knowledge acquisition, manual and/or automatic, from sources such as domain specialists, literature, databases, existing ontologies (Stevens et al., 2000), etc. Such knowledge includes domain concepts, their relations and terms that represent the concepts linguistically, and it is encoded in a formal representation language such as e.g. Web Ontology Language (OWL, 2007). The quality of an ontology is evaluated in terms of its consistency, completeness and conciseness. While providing a framework for flexible yet coherent and rigorous structuring of domain-specific knowledge, it is also necessary for an ontology to be easily extensible especially in an expanding domain such as metabolomics. The new knowledge generated by high-throughput screening is communicated through biotechnical literature, which can be exploited by text mining (TM) tools in order to facilitate the process of keeping ontologies up to date (Mack and Hehenberger, 2002). Such approaches typically involve automatic term recognition (ATR), estimation of similarity between terms and their subsequent clustering. The similarity measures can be applied directly to terms, but may consider their

* To whom correspondence should be addressed.

[1] Terms are means of conveying scientific and technical information (Jacquemin, 2001). More precisely, terms are linguistic representations of domain-specific concepts (Kageura & Umino, 1996).

contexts too. Term clustering suggests term associations, which can be used to verify or update instances of semantic relations in an ontology.

Still, such automatically constructed ontologies are limited to the corpus of source documents and different types of relations are often conflated into undefined and diffuse associations based on term co-occurrences. The former can be alleviated through information retrieval (IR), which gathers and filters relevant documents (Baeza-Yates and Ribeiro-Neto, 1999). The latter is often remedied by relying on linguistic indicators (in the form of lexico-syntactic patterns) of different types of relations. For example, Hahn et al. (2002) followed patterns such as "*...the symptom <term>...*" and "*...symptoms like <term>...*" to positively identify the given term as an instance of a *Symptom*.

3 SCOPE, COVERAGE AND STRATEGY

The proposed scope of the metabolomics ontology as developed by the MSI OWG aims to support the activities of other MSI WGs: Biological Context Metadata, Chemical Analysis, Data Processing and Exchange Format. The minimal reporting requirements identified by the first three WGs will inform the development of data exchange standards and the ontology in order to provide a common mode of transporting information between systems.

The coverage of the domain has been divided according to the typical structure of metabolomics investigations: general components (investigation design; sample source, characteristics, treatments and collection; computational analysis) and the technology-specific components (sample preparation; instrumental analysis; data pre-processing). The semantic framework for the general aspects of metabolomics investigations largely overlaps with the ongoing standardisation efforts in other omics domains, such as the Human Proteome Organization Proteomics Standards Initiatives (HUPO-PSI, 2007; Taylor et al., 2006), the Microarray Gene Expression Data Society (MGED, 2007) and other ontology communities under the Open Biomedical Ontologies (OBO, 2007; Rubin et al., 2006) umbrella. Therefore, the MSI OWG has decided to focus initially on the technology-specific components. Further, the development activities in this sub-domain have been prioritised according to the pervasiveness of the analytical platforms used.

An array of analytical technologies has been employed in metabolomics studies (Dunn and Ellis, 2005). Mass spectrometry (MS) is the most widely used analytical technology in metabolomics, as it enables rapid, sensitive and selective qualitative and quantitative analyses with the ability to identify individual metabolites. In particular, the combined chromatography-MS technologies have proven to be highly effective with this respect. Gas chromatography-mass spectrometry (GC-MS) uses GC to separate volatile and thermally stable compounds prior to detection via MS. Similarly, liquid chromatography-mass spectrometry (LC-MS) provides the separation of compounds by LC, which is again followed by MS. On the other hand, nuclear magnetic resonance (NMR) spectroscopy does not require the separation of the compounds, thus providing a non-destructive, high-throughput detection method with minimal sample preparation, which has made it highly popular in metabolomics investigations despite being relatively insensitive in comparison to the MS-based methods. For the MS the MSI OWG will leverage on previous work by the PSI MS Ontology WG. An ontology for chromatography, shared by both proteomics and metabolomics domains, is being developed in close collaboration with the PSI Sample Processing Ontology WG. The technologies the MSI OWG is currently focusing on are NMR and GC.

The MSI OWG efforts are divided into two key stages: (1) reaching a consensus on the CVs, and (2) developing the corresponding ontology as part of the Ontology for Biomedical Investigations (OBI, 2007). In this paper, we focus on the first stage. Each CV is compiled in an iterative process consisting of the following phases:

(1) Create an initial CV by re-using the existing terminologies from database models, glossaries, etc. and normalise the terms according to the common naming conventions.
(2) Expand the initially created CV with other frequently co-occurring terms identified automatically using a TM approach over a relevant collection of scientific publications.
(3) Discuss the CV within the MSI OWG and circulate it to the practitioners in the relevant metabolomics area for the validation in order to ensure the quality and completeness of the proposed CV.

The result of the first phase is an initial draft CV encompassing terms of different types: methods, instruments, parameters that can be measured, etc. In the highly dynamic metabolomics domain, experts often use non-standardised terms. Therefore, in order to reduce the time and cost of compiling a CV and sustain its completeness, we propose a TM approach to acquire terms automatically from the scientific literature.

4 AUTOMATED TERM ACQUISITION

A set of relevant tasks regarding term acquisition has been identified, including: IR, ATR and term filtering. First, the IR module is used to gather and filter relevant documents from the literature. The resulting domain-specific corpus of documents is subjected to ATR in order to extract terms as domain-specific lexical units, i.e. the ones that frequently occur in the corpus and bear special meaning in the domain. In order to reduce the number of terms not directly related to a given technology, we filter out typically co-occurring classes of terms denoting substances, organisms, organs,

diseases, etc., which in contrast to the considered analytical techniques have more established CVs.

4.1 Information retrieval

We used two relevant sources of information: Medical Literature Analysis and Retrieval System Online (MEDLINE, 2007) and PubMed Central (PMC, 2007), henceforth referred to together as PubMed, which provide peer-reviewed literature and make it freely accessible in a uniform format. MEDLINE distributes *abstracts* only, while PMC provides *full-text articles*.

Documents available in PubMed are indexed by Medical Subject Headings (MeSH, 2007), a CV consisting of hierarchically organised terms that permit direct access to relevant documents at various levels of specificity. For example, *Magnetic Resonance Spectroscopy* is a relevant MeSH term for NMR spectroscopy. However, any analytical technique employed in metabolomics is not likely to be the main focus of a study. Consequently, the corresponding documents may not necessarily be annotated with technology-related MeSH terms. Moreover, it is more likely for the results discovered to be reported in an abstract than for the technology-specific experimental conditions leading to these results. Such parameters are usually reported in the *Materials and Methods* section or in the supplementary material. Hence, it is important to: (1) go beyond MeSH terms in query formulation, and (2) query the full-text articles as opposed to abstracts only. We used the terms from an initially compiled CV as search terms to retrieve additional documents that describe research that utilises a technology, but which do not necessarily deal with the technology per se (and therefore may not be indexed by technology-related MeSH terms):

(1) For each CV term obtain the number of matching documents. Sort the terms by the number of documents they return and set a cut-off point to remove the terms that return too many documents, as they are likely to be broad terms not limited to a specific technology and consequently introducing unwanted noise into the domain-specific corpus.
(2) For each remaining term retrieve the matching documents. Sort the retrieved documents according to the number of search terms they match and set a cut-off point to remove the ones that do not contain a sufficient number of the seed CV terms.

These methods encoded in Java take advantage of E-Utilities (Entrez, 2007), a web service which enables the users to run Entrez queries and download data using their own applications. Information about terms and query results are stored in a local database (DB) hosted on PostgreSQL. By storing the mappings between terms and documents, we can combine the querying ability of the DB management system with that of Entrez. The DB is also accessible via Java applications (using the Java Database Connectivity– a standard SQL DB access interface). Hence, all implemented modules can be incorporated into customised workflows (Oinn et al., 2007).

4.2 Term recognition

The IR results include two domain-specific corpora, one for abstracts and one for full-text articles. The *Methods and Materials* sections are extracted automatically relying on the PMC's XML format in which full papers are distributed. It is important to focus on the sections that are likely to contain terms relevant for an analytical technology as a preparation step for ATR. Namely, we employed the C-value method (Frantzi and Ananiadou, 1999), publicly accessible at (NaCTeM, 2007), which extracts terms using linguistic knowledge (term formation patterns) and statistical analysis (various types of occurrence frequencies). It relies primarily on the frequency of term usage and their general syntactic properties rather than exploiting orthographic, lexical and syntactic features of specific named entities. The latter would significantly increase the time and cost of the development of CV term acquisition as these would have to be tuned for specific (semantic) classes of terms. Moreover, the type of terms sought may not necessarily exhibit any discriminatory textual properties. On the other hand, focusing the C-value on specific sections combined with subsequent filtering offers an alternative for rapid development of TM workflows for CV expansion.

4.3 Term filtering

Manual inspection of initially extracted terms revealed the main types of concepts studied in metabolomics: substances, organisms, organs, conditions/diseases, etc. Unlike analytical technologies many of which are relatively recent, some of these classes are fairly stable with respect to the number of new concepts/terms being introduced thus having more established CVs, e.g. Linnaean taxonomy (Linnaeus, 1753) classifies living organisms in a systematic manner. We relied on the Unified Medical Language System (UMLS, 2007) as it merges information from over 100 biomedical source vocabularies (Bodenreider, 2004). The following semantic types in the UMLS proved relevant to our problem: *Organism*, *Anatomical Structure*, *Substance*, *Biological Function* and *Injury or Poisoning*. We used them to select the corresponding terms from the UMLS thesaurus. Then, we applied a simple pattern matching technique to filter out these terms and their variations.

5 RESULTS

The initial CVs were compiled manually, providing a total of 243 and 152 seed terms for NMR and GC respectively. The MeSH terms relevant for the techniques of interest are: *Magnetic Resonance Spectroscopy* and *Chromatography, Gas*. Table 1 provides the term acquisition results for the two case studies and IR approaches using the MeSH terms and the seed CV terms (at least 3 and 7 matching terms for

abstracts and full papers respectively) including: (1) the number of documents retrieved, (2) the size of collected corpora, (3) the number of terms extracted by C-value and (4) the final number of terms remaining after filtering. Although freely available for browsing, for most articles in PMC the publisher does not allow downloading the text in XML format, neither does PMC allows bulk downloading in HTML format. Hence, we were able to process only a small portion of full papers (the numbers in brackets refer to these papers). A total of 5,699 and 2,612 new terms were acquired for NMR and GC respectively. The average ratio between the number of acquired technology-specific terms and the corpus size was 16.25 for full papers and only 0.13 for abstracts. The overlap between the terms acquired form abstracts and those from the full papers was on average only 2%. This comparison confirms that the *Materials and Methods* sections represent a significant source of technology-specific terms and also emphasises the need of making full-text articles available to TM applications for the benefits of the overall biomedical community.

The preliminary results are available at (MSI OWG, 2007), where the potential CV terms are accessible to the metabolomics community for comments and curation. The official version of the NMR CV has been made publicly available at (OBO, 2007).

Table 1. Term acquisition results

		MeSH terms		seed CV terms	
		abstracts	**full papers**	**abstracts**	**full papers**
documents	NMR	122,867	6,125 (141)	1,613	758 (29)
	GC	60,338	1,351 (79)	3,948	1,383 (58)
corpus size	NMR	113,191	663	2,047	270
	GC	42,418	68	3,012	97
C-value terms	NMR	5,602	6,215	124	2,601
	GC	2,708	811	2,442	1,114
final terms	NMR	2,298	3,257	61	1,385
	GC	567	348	1,323	526

ACKNOWLEDGEMENTS

We kindly acknowledge the MSI Oversight Committee, other MSI WGs, NaCTeM, the OBI WG, the OBO Foundry leaders and the Ontogenesis Networks members for their contributions in fruitful discussions. We gratefully acknowledge the support of the BBSRC/EPSRC via "The Manchester Centre for Integrative Systems Biology" grant (BB/C008219/1: DBK, NP and IS), the BBSRC e-Science Development Fund (BB/D524283/1: SAS and DS) and the EU Network of Excellence Semantic Interoperability and Data Mining in Biomedicine (NoE 507505: IS and DS).

REFERENCES

Baeza-Yates,R. and Ribeiro-Neto,B. (1999) *Modern information retrieval*. ACM Press, New York.

Blake,J. (2004) Bio-ontologies–fast and furious. *Nat Biotechnol*, **22**, 773-4.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, **32**, D267-70.

Bodenreider,O. and Stevens,R. (2006) Bio-ontologies: current trends and future directions. *Brief Bioinform*, **7(3)**, 256-74.

Castle,A.L. et al. (2006) Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results. *Brief Bioinform*, **7**, 59-65.

Cimino,J.J. and Zhu,X. (2006) The practical impact of ontologies on biomedical informatics. *Methods Inf Med*, **45 Suppl 1**, 124-35.

Dunn,W.B. and Ellis,D.I. (2005) Metabolomics: current analytical platforms and methodologies. *Trends Anal Chem*, **24(4)**, 285-294.

Entrez (2007) http://www.ncbi.nlm.nih.gov/Entrez

Field,D. and Sansone,S.A. (2006) A special issue on data standards. *OMICS*, **10(2)**, 84-93.

Frantzi,K. and Ananiadou,S. (1999) The C-value/NC-value domain independent method for multiword term extraction. *J Nat Lang Proc*, **6(3)**, 145-80.

Hahn,U., Romacker,M. and Schulz,S. (2002) Creating Knowledge Repositories from Biomedical Reports: The MEDSYNDIKATE Text Mining System. Pac Symp Biocomput, 338-349.

HUPO-PSI (2007) http://psidev.sf.net

Jacquemin,C. (2001) *Spotting and Discovering Terms through Natural Language Processing*. MIT Press.

Kageura,K. and Umino,B. (1996) Methods of automatic term recognition – a review. *Terminology*, **3(2)**, 259-89.

Linnaeus,C. (1753) *Species plantarum*. Stockholm.

Mack,R. and Hehenberger,M. (2002) Text-based knowledge discovery: search and mining of life-sciences documents. *Drug Discov Today*. **7**(11), S89-98.

MEDLINE (2007) http://www.pubmed.gov

MeSH (2007) http://www.nlm.nih.gov/mesh

MGED (2007) http://www.mged.org

MSI (2007) http://msi-workgroups.sf.net

MSI OWG (2007) http://msi-ontology.sf.net

NaCTeM (2007) http://www.nactem.ac.uk

OBI (2007) http://obi.sf.net

OBO (2007) http://obo.sourceforge.net

Oinn,T. et al. (2007) Taverna/myGrid: aligning a workflow system with the life sciences community. In *Workflows for e-Science: scientific workflows for Grids* (Ed. Taylor,I.J. et al.), Springer, Guildford, 300-19.

OWL (2007) http://www.w3.org/TR/owl-features

PMC (2007) http://www.pubmedcentral.nih.gov

Quackenbush,J. (2004) Data standards for 'omic' science. *Nat Biotechnol*, **22(5)**, 613-4.

Rubin,D.L. et al. (2006) National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS*, **10**(2), 185-98.

Shulaev,V. (2006) Metabolomics technology and bioinformatics. *Brief Bioinform*, **7**, 128-39.

Schulze-Kremer,S. (1998) Ontologies for molecular biology. Pac Symp Biocomput, 695-706.

Smith,B. (2006) From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. *J Biomed Inform*, **39**, 288-98.

Spasic,I. et al. (2005) Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform*, **6(3)**, 239-51.

Stevens,R., Bechhofer,S. and Goble,C. (2000) Ontology-based knowledge representation for bioinformatics. *Brief Bioinform*, **1**(4), 398-414.

Taylor,C.F. et al. (2006) The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). *OMICS*, **10(2)**, 145-51.

UMLS (2007) http://umlsinfo.nlm.nih.gov

Using GO terms to evaluate protein clustering

Hugo Bastos, Daniel Faria, Catia Pesquita and André O. Falcão*

University of Lisbon, Department of Informatics, Campo Grande, 1749-016 Lisbon, PORTUGAL

ABSTRACT

Motivation: Protein sequence data is growing at an exponential rate. However a considerable portion of this data is redundant, with many new sequences being very similar to others in the databases. While clustering has been used to reduce this redundancy, the influence of sequence similarity in the functional quality of the clusters is still unclear.

Results: In this work, we introduce a greedy graph-based clustering algorithm, which is tested using the Swiss-Prot database. We study the topology of the protein space as function of the threshold BLAST *e*-values, and the functional characterization of the clusters using the Gene Ontology. Initial results show that seemingly the cluster centers alone can capture a large portion of the information content of the database, therefore largely reducing its redundancy. Also it was found an expected increase of cluster functional coherence and characterization with the stringency of the threshold, as well as the amount of information captured by the cluster centers.

1 INTRODUCTION

The Universal Protein Resource (UniProt) provides a central hub for the collection of protein sequences, with accurate, consistent and rich annotation (The UniProt Consortium, 2007). As such, it can be considered a subset of the universal protein space. The protein space is a metric space containing all protein sequences that uses sequence similarity as a distance function. While it is true that the number of sequences in UniProt is growing exponentially over time, it is also true that a significant proportion of these sequences is redundant, being very similar to others already present in the database.

One way to cope with this large amount of partially redundant data, is through clustering methods, which ideally may reduce the dimension of the protein space without loss of information. In fact, UniProt itself provides clustered versions of its full database with several levels of sequence similarity (UniRef) to facilitate information retrieval and accelerate the querying process. Other purposes for clustering biological sequences include functional annotation, comparative genomics and structural genomics (Petryszak *et al.*, 2005, Nikolski and Sherman, 2007, Yan and Moult, 2005).

* To whom correspondence should be addressed.

Protein sequence clustering methods can be classified according to two aspects: the clustering algorithm type used, which is usually either hierarchical or graph-based; and the criterion used to group sequences, which is usually either domain-based or family-based.

On structural genomics studies, Charette *et al.* (2006) used protein clustering for protein ligand-docking and molecular dynamics, whereas Shen *et al.* (2005) used it for protein class prediction, and Yan and Moult (2005) searched for representative family structural templates.

In the field of functional genomics, Pellegrini *et al.* (1999) used clustering to functionally assign proteins, whereas Nikolski and Sherman (2007) used a consensus algorithm tailored for comparative genomics projects. Hierarchical algorithms include ProtoMap (Yona *et al.*, 1999), ProtoNet (Sasson *et al.*, 2003), and CLUGEN (Ma *et al.*, 2005) which aim at constructing a comprehensive view of the protein space by means of family-based classification. On the other hand, CluSTr's (Petryszak *et al.*, 2005) main objective is automated protein annotation. As for graph-based algorithms, Abascal and Valencia (2003) used this type of clustering to identify families for comparative genomics and protein functional inference, while the Cluster-C algorithm (Mohseni-Zadeh *et al.*, 2004) is used for protein family construction within whole proteomes. SEQOPTICS (Chen *et al.*, 2006) used a different clustering algorithm which performs density-based ordering.

One way to evaluate the biological quality of the clustering process, is by analyzing the amount of information (functional or other) conserved in the cluster centers vs. the reduction in dimension achieved. In this context, a unified and structured vocabulary to describe proteins' functional aspects would provide a unique background for evaluation, facilitating the clusters' functional characterization.

Being already extensively used to annotate several biological databases (including UniProt), the Gene Ontology (GO) (Ashburner *et al.*, 2000) is one of the best choices for this end. Indeed, GO has already been used as a background for functional comparison of proteins (Lord *et al.*, 2003) and to understand the relation between protein sequence and function (Duan *et al.*, 2006).

In this paper we present a graph-based protein sequence clustering algorithm which was tested using the Swiss-Prot database, with a discrete range of BLAST *e*-value cut-offs. We study the reduction in protein space and its topology as

function of sequence similarity, and evaluate the biological quality of the clusters using three GO-based parameters which measure different aspects: functional coherence, functional characterization and representativeness of the cluster center.

2 DATA AND IMPLEMENTATION

In order to evaluate the cluster algorithm all the Swiss-Prot portion of the UniProt Knowledgebase (Release 9.6) was used. After filtering out all segment sequences, an all-against-all BLAST homology search was performed with the remaining 2.51×10^5 sequences. The maximum *e*-value accepted in the BLAST step was 10^{-4}, which resulted in about 5.48×10^7 pairwise BLAST comparisons. A simple greedy graph partitioning clustering algorithm was then applied to the BLAST results, as is described next.

For a given *e*-value threshold an edge (weighted by *e*-value) between two nodes (sequences) is said to exist if the *e*-value between the two proteins is below that threshold, and a list of sequences is constructed, containing for each the total number of edges (cardinality). The algorithm then proceeds as follows:

(1) Each node is sorted descendingly according to the number of edges it has (cardinality).

(2) The node with the highest cardinality is selected from the list. If it does not belong to a cluster, a new one is created with this node as its center. All proteins linked to that protein by an edge are then assigned to this new cluster.

(3) The clustering proceeds iteratively down the cardinality list until no more nodes can be assigned.

The clustering procedure was run with a discrete set of *e*-value thresholds ranging from 1×10^{-4} (most permissive threshold) to 1×10^{-100} (most stringent threshold). For biological validation of the clustering, the GOA-UniProt release 45.0 and the GO release 2007-02 were used. Only the *molecular function* aspect of GO was considered, since the goal was to characterize the clusters in functional terms, and the relation between the other GO aspects and functional similarity is unclear.

3 RESULTS AND DISCUSSION

Within the chosen range of *e*-value thresholds, the size of the protein space achieved with clustering varied non-linearly with the threshold, growing from 9% of the original size (22799 protein clusters) with the most permissive threshold to 25% (62254 protein clusters) with the most stringent (Figure 1). As would be expected, this variation was accompanied by a reduction in the average cluster size from 18 to 3 proteins, as well as a growing fraction of singletons from 42 to 62%, from the most permissive to most stringent threshold.

As the goal of this work was to evaluate the clustering process at the functional level, using *molecular function* GO terms, only non-singleton clusters with at least one annotated protein were considered. While the fraction of proteins annotated with *molecular function* GO terms in the dataset is 84%, the fraction of annotated clusters grew with threshold stringency (from 72 to 87%) although the absolute number of non-annotated clusters was approximately constant (~4000 clusters).

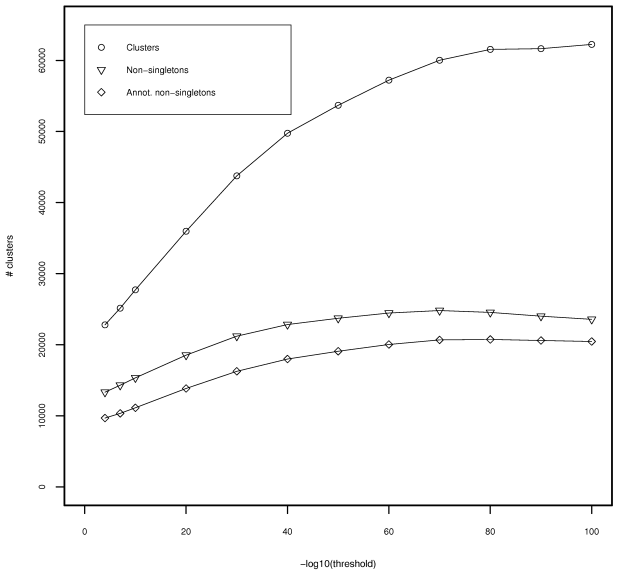


Fig. 1. Number of clusters, non-singleton clusters and annotated non-singleton clusters according to clustering *e*-value threshold.

In order to evaluate the clustering process, three GO term-based parameters were developed to measure different aspects of cluster quality: *GOoccurrence*, *GOscore* and *GO-center*. For a given cluster *C*, *GOoccurrence* is given by the average frequency of annotation within the cluster of each of the cluster's GO terms:

$$GOoccurrence(C) = \mathrm{AVG}_{term \in C}[\mathrm{freq}_C\,(term)]$$

and measures how coherent is the cluster functionally. The maximum for this parameter is achieved when all terms are annotated to all of the cluster's proteins (i.e. when all proteins are functionally identical). However, as annotations are considered at all levels of the GO graph (i.e. direct annotations and their ancestors) this parameter is slightly biased by the more general terms (which usually have greater frequency of annotation).

GOscore is given by the maximum of term information content (IC) times term annotation frequency:

$GOscore(C) = \mathrm{MAX}_{term \in C}[\mathrm{freq}_C(term) \times \mathrm{IC}(term)]$

and measures how well characterized is the cluster functionally, by capturing the most representative functional aspect: one that is simultaneously specific (with high IC) and frequent within the cluster. Note that the product of IC times frequency is actually proportional to the logarithm of the probability of the term occurring in the cluster with that frequency.

GOcenter is given by the fraction of the cluster's terms which are annotated to the cluster center protein, and provides a measure of how much of the cluster's functional aspects are captured by the center (*i.e.* how representative the center is of the cluster).

When analyzing the average variation of these three evaluation parameters with the threshold *e*-value used for clustering, they were found to increase non-linearly with threshold stringency (Figure 2), showing an apparent asymptotic behavior as they reach high stringency values.

The increases in *GOoccurence* (32% increment) and in *GOscore* (14% increment) agree with what would be expected of the clustering process: by increasing the level of sequence similarity required for a protein to be in a cluster, we obtain clusters that are functionally more coherent and better characterized.

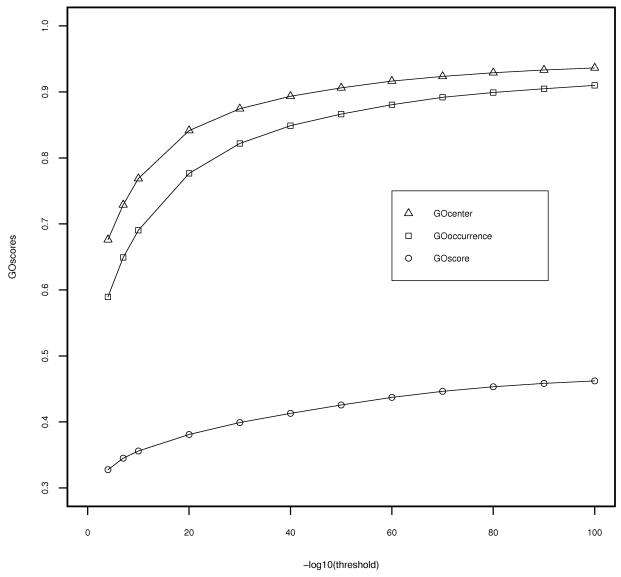


Fig. 2. Average GO term-based evaluation parameters as function of the threshold *e*-value used for clustering.

While the absolute values of *GOscore* may appear low, considering that their theoretical maximum value is 0.48 (average *IC* of the most specific term annotated to each protein in the dataset), their range corresponds to 59-96% of the maximum, which is similar to the ranges of the other parameters.

Thus, both *GOoccurrence* and *GOscore* support the quality of the clustering process, and provide a biological validation for the clustering algorithm used.

As for the *GOcenter* value, its increase (from 0.68 to 0.94) means that the representativity of the cluster centers is increasing with stringency, supporting the notion that the protein space can be successfully reduced through clustering without significant loss of functional information. This also shows that the algorithm is successful, since the cluster centers it selects capture most of the information in the clusters.

One issue that can be raised is that GO annotations inferred by sequence similarity lead to circularity of the results. However, the number of such annotations which are curated (evidence code ISS) amounts only to 1.5% of all annotations, and while it is likely that a portion of the electronic annotations (90.1% of all annotations) are also inferred by sequence similarity, that portion is difficult to estimate. Therefore, while the issue is acknowledged, there is no way to avoid it save ignoring all ISS and electronic annotations, which would imply losing the vast majority of the available information. Furthermore, the main consequence of circularity will likely be an increase in the absolute values of the evaluation parameters used, which is counterbalanced by the fact that proteins with no annotations cause a decrease in those parameters. It is unlikely that the patterns observed (Figure 2) are uniquely an artifact of data circularity, since that would mean that the majority of annotations are not only inferred by sequence similarity but also erroneous.

Interestingly, the approach developed and specifically the parameters *GOscore* and *GOoccurence* can be used to evaluate and improve the quality of annotations (by identifying protein clusters which have very incoherent annotations, and by providing a safer basis for inferring functional annotations).

4 CONCLUSION

A simple, greedy algorithm to cluster proteins based on sequence similarity was developed. A range of BLAST *e*-values was used as threshold for the clustering process so as to study the topology and functional quality of the cluster space as function of the sequence similarity.

Three parameters, based on GO *molecular function* annotations, were developed to evaluate different aspects of cluster quality: coherence, functional characterization and center representativeness.

Cluster quality in all these aspects was found to increase with threshold stringency, validating the biological significance of the clustering algorithm used and also the apparent ability of present method to reduce the protein space without significant loss of functional information.

Future work will focus on using additional biological validation methods to complement those based on GO, so as to overcome the issue of coverage (only 84% of the sequences in the dataset have GO annotations). The clustering algorithm will also be improved by using the proteins' level of annotation as an additional criterion to select the cluster centers. Furthermore, clusters will be characterized taxonomically and using GO-based functional semantic similarity.

ACKNOWLEDGEMENTS

This work was partially supported by the Peptides project of the LaSIGE research group, and also by grant ref. SFRH/BD/29797/2006 awarded by the Portuguese 'Fundação para a Ciência e a Tecnologia'.

REFERENCES

Abascal, F. and Valencia, A. (2003) Automatic annotation of protein function based on family identification. *Proteins*, **53**(3), 683–692.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet*, **25**(1), 25–29.

Charette, B., Macdonald, R., Wetzel, S., Berkowitz, D. and Waldmann, H. (2006) Protein structure similarity clustering: Dynamic treatment of pdb structures facilitates clustering. *Angew Chem Int Ed Engl*, **45,** 7666-7770.

Chen, Y., Reilly, K., Sprague, A., and Guan, Z. (2006) SEQOPTICS: a protein sequence clustering system. *BMC Bioinformatics*, **7** (Suppl 4), S10.

Consortium, The Uniprot (2007) The universal protein resource. *Nucl Acids Res*, **35** (Suppl 1), D193–197.

Duan, Z., Hughes, B., Reichel, L., Perez, D. and Shi, T. (2006) The relationship between protein sequences and their gene ontology functions. *BMC Bioinformatics*, **7** (Suppl 4), S11.

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**(10), 1275–1283.

Ma, Q., Chirn, G.-W., Cai, R., Szustakowski, J.D. and Nirmala, N. (2005) Clustering protein sequences with a novel metric transformed from sequence similarity scores and sequence alignments with neural networks. *BMC Bioinformatics*, **6**, (242-255).

Mohseni-Zadeh, S., Brezellec, P. and Risler, J.-L. (2004) ClusterC, an algorithm for the large-scale clustering of protein sequences based on the extraction of maximal cliques. *Computational Biology and Chemistry*, **28**(3), 211–218.

Nikolski, M. and Sherman, D. (2007) Family relationships: should consensus reign?--consensus clustering for protein families. *Bioinformatics*, **23**(2), e71–76.

Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D. and Yeates, T. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA*, **96**(8), 4285–4288.

Petryszak, R., Kretschmann, E., Wieser, D. and Apweiler, R. (2005) The predictive power of the CluSTr database. *Bioinformatics*, **21**(18), 3604–3609.

Sasson, O., Vaaknin, A., Fleischer, H., Portugaly, E., Yonatan, B., Linial, N. and Linial, M. (2003) Protonet: hierarchical classification of protein space. *Nucleic Acids Research*, **31**(1), 348–352.

Shen, H.-B., Yang, J., Liu, X.-J. and Chou, K.-C. (2005) Using supervised fuzzy clustering to predict protein structural classes. *Biochemical and Biophysical Research Communications*, **334**(2), 577–581.

Wetlaufer, D. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci USA*, **70**(3), 697–701.

Yan Y. and Moult, J. (2005) Protein family clustering for structural genomics. Journal of Molecular Biology, **353**(3), 744–759.

Yona, G., Linial, N. and Linial, M. (1999) Protomap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. PROTEINS: Structure, Function, and Genetics, **37**(3), 360–378.