2005

Toward Developing a Provenance Ontology for Biological Images

Jun Liu

Sudha Ram

The Eighth Annual Bio-Ontologies Meeting

Robert Stevens

Phillip Lord

University of Manchester

Robin McEntire

James A. Butler

GlaxoSmithKline

With thanks to David Benton

June 24, 2005
At ISBM 2005
Detroit, Michigan, USA

Talks Programme

Start	End	Speaker	Title
08:50	09:00		Introduction
09:00	10:00	Mark Musen	KEYNOTE:
10:00	10:20	Christohper Baker	The FungalWeb Ontology: Application Scenarios
10:20	10:40	Mary Shimoyama	Using Multiple Ontologies to Integrate Complex Biological Data
10:40	11:00	Nigam Shah	A case study in knowledgebase verification
11:00	11:20		Coffee
11:20	11:40	Aditya Vailaya	Ontology-based Statistical Analysis of Microarray Data
11:40	12:00	Olivier Bodenreider	Ontology-driven similarity approaches to supporting gene functional assessment
12:00	13:30		Lunch and Poster Session
13:30	13:50	Pankaj Jaiswal	Plant Structure and Growth Stage Ontologies to describe phenotypes and gene expression in angiosperms
13:50	14:10	Daniel McShan	Toward a Biochemical Ontology for Functional Classification of Arbitrary Compounds
14:10	14:30	Jun Liu	Toward Developing a Provenance Ontology for Biological Images
14:30	15:00		Coffee
15:00	16:20	Mark Musen	Panel Session
		Larry Hunter	
		Judith Blake	
		Eric Neumann	
16:20	16:40	Olivier Bodenreider	Linking the Gene Ontology to other biological Ontologies
16:40	17:00	Naren Ramakrishnan	Reconstructing Formal Temporal Logic Models of Cellular Events using the GO process Ontology
17:00	17:20	Barry Smith	On the proper treatment of pathologies in biomedical ontologies
17:20	17:40	Katherine Wolstencroft	Using Ontology Reasoning to Classify Protein Phosphatases
17:40	18:30		Poster Session

Poster Programme

Presenter	Title
Doina Caragea	INDUS: an ontology-based information integration system
Michael P. Cary Joanne S. Luciano	News & Views: The BioPAX Pathway Data Exchange Format
Yizong Cheng	Co-Citation of Genes in Ontology Groups
C. Forbes Dewey	ExperiBase - An Object Model Implementation for Biology
Midori Harris	Why GO there? Ensuring that the Gene Ontology Meets Biologists' Needs
Pankaj Jaiswal	Design and Implementation of the Environment and Trait Ontologies for Plants
Cliff Joslyn	Automating Ontological Function Annotation: Towards a Common Methodological Framework
Peter Karp	BioWarehouse: A Bioinformatics Database Warehouse Toolkit
Susie Stephens	A Novel Ontology Development Environment for the Life Sciences
Trish Whetzel	FuGO: Development of a Functional Genomics Ontology (FuGO)
W. Jim Zheng	Integration of the Gene Ontology into an object-oriented architecture

Acknowledgements

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

Thanks to David Benton for reviewing papers at short notice.

Thanks to Yeliz Yesilada for once again providing technical assistance with this booklet.

Panel Session

This years panel session will be passed around a discussion on a set of three statements. We are including the questions here to give you an opportunity to think about them before hand, in the hope that this will stimulate discussion at the meeting.

In the current context of rapidly growing activity in the bio-ontology sector and possible new horizons in the Semantic Web, there are many contentious issues facing our community:

Statement 1)

We should be more authoritarian and less liberal in the building of Bio-Medical Ontologies.

Background:

As the development of bio-medical ontologies has become more widespread, the development of multiple ontologies with overlapping terms is inevitable (and is, to some extent, already happening).

Currently, a very "free market" approach is being followed. Is this a strength? Or should it be replaced with something more centralised, similar to, for example, the Human Gene Nomenclature Committee.

Statement 2)

We are better at developing Bio-Medical Ontologies that we are at using them as key components of critical applications in academic research and commercial systems.

Background:

BioMedical ontologies now have a broad spread over the subject area. The main use of these ontologies is to annotate records which we then retrieve by query or navigation. Should we be doing more with these ontologies? Even this narrow use lacks good use facing and reusable tooling.

Statement 3)

The future for BioMedical ontologies is to be the semantic infrastructure for the computationally enabled systems biology view of life.

Background:

In the Semantic Web vision, data and services are described semantically, such that they become computationally amenable. Bioinformatics is in a strong position to realise this vision, which might affect both the way the world uses data, and the way biologists view life. Is this either or both possible or desirable.

The FungalWeb Ontology: Application Scenarios

** Christopher J. O. Baker, René Witte, Arash Shaban-Nejad, Greg Butler and Volker Haarslev*

Concordia University, Montreal, Quebec, Canada

ABSTRACT

Motivation: The FungalWeb Ontology aims to support the data integration needs of enzyme biotechnology from inception to product roll out. Serving as a knowledge base for decision support, the conceptualization seeks to link fungal species with enzymes, enzyme substrates, enzyme classifications, enzyme modifications, enzyme retail and applications. We demonstrate how the FungalWeb Ontology supports this remit by presenting application scenarios, conceptualizations of the ontological frame able to support these scenarios and semantic queries typical of a Biotech Manager. Queries to the knowledge base are answered with description logic (DL) and automated reasoning tools.

1 INTRODUCTION

Fungi are microorganisms well know for the range of novel enzymes they produce and enzymes of fungal origin are now used in industrial processes which amount to billions of dollars of revenue annually. The path to product development, namely gene discovery, enzyme characterization, mutational improvement and industrial application is long and fraught with numerous hurdles, both with respect to the domain knowledge and technical challenges. In an RnD environment many decisions are frequently made on incomplete knowledge. The current need is to have an integrated framework for discovery and decision support. This must integrate data from laboratory research, data accessed from distributed database, web and textual resources as well as the results of bioinformatics computation. To provide a reliable semantic resource in a contemporary RnD environment the scientific and technical span of ontology must encapsulate a more inter disciplinary range of concepts. The full range of conceptualizations required for commercial enzymologists includes taxonomy, gene discovery, protein family classification, enzyme characterization, enzyme improvement, enzyme production, enzyme substrates, enzyme performance benchmarking, and market niche. Inclusion of such concepts and instance data in ontology is within the scope of the FungalWeb data integration initiative.

2 ONTOLOGY DEVELOPMENT

The Fungal Web Ontology [1] is the result of integrating numerous biological database schema, web accessible textual resources and interviews with domain experts. The ontology includes both hierarchical structures supporting full-subsumption taxonomies and a broader conceptual frame with novel relationships for specific domain knowledge. The major resources for fungal terminologies and concepts come from the following sources: NCBI taxonomy and literature databases [2] , NEWT: is the taxonomy database [3], BRENDA enzyme database [4], *Saccharomyces* Genome Database [5], *Neurospora crassa* Genome Database [6], Commercial Enzyme Vendor Web Resources and the Enzyme Nomenclature Database [7].

The FungalWeb Ontology (FWOnt) reuses and integrates existing bio-ontologies and knowledgebases by merging, mapping and sharing common concepts using logics. Our ontology is an integrated ontology which used components of Gene ontology (GO) [8], TAMBIS [9] to establish the basic frame upon which biotechnology specific concepts have been added. The Ontology is a formal ontology written in OWL-DL, a sublanguage of Ontology Web language (OWL) with correspondence to description logics (DL). This provides maximum expressiveness, without losing computational completeness and decidability of reasoning systems. Protégé 2000 [10] was used (with Owl plug-in) as a knowledge representation editor. Aptness (considering completeness, consistency and conciseness) of the ontology for its intended application and the scientific integrity was evaluated by posing DL queries. RACER [11] was used as a description logic reasoning system with support for T-Box (concepts) and A-Box (instances).

3 APPLICATION SCENARIOS

We demonstrate the scope of our ontological conceptualization and the range of cross disciplinary queries that can be posed. We describe *junction* scenarios where a biotechnologist would ask support from the ontology and illustrate the how the diverse needs of the fungal biotechnology manager can be accommodated. The scientific context of these semantic queries and the conceptual frames designed to support them are outlined. nRQL syntax of DL queries to the ontology using Racer are presented for each scenario.

1.1 Enzymes acting on substrates

The ontology includes a concept representing the semantic stem of the systematic chemical names of enzyme substrates. This concept is instantiated with an NLP derived word stem of the most common term found in the enzyme descriptions of enzyme reaction classification scheme of the International Union of Biochemistry (IUB). By instantiating the semantically rich descriptions of the IUB into the conceptualization of the ontology we are able to query for mul-

* baker@cs.concordia.ca

tiple enzymes families known to degrade / modify a chemical substrate. A use case example querying for enzymes that act on the glucuronic acid polymer 'pectin' is described.

1.2 Enzyme provenance

A deep fungal taxonomy and enzyme reaction hierarchy are included in the ontology. The establishment of the relationship 'has been reported to be found in' between the concepts enzyme and fungus, reflecting information in scientific papers, facilitates the query of provenance of fungal enzymes (which enzyme is found in which fungal species). Such a query is further complemented by queries able to identify the common taxonomic lineage of all enzymes with a particular function and is of value to the biotech manager interested in the gene discovery and biodiversity.

1.3 Enzyme benchmark testing

An industrial specification is an important component of the FungalWeb ontology, representing concepts of value to the commercially oriented enzymologist. Access to information on commercial enzymes, product names, product parameters and vendors assists in the benchmarking the performance of newly discovered or mutationaly improved enzymes. Typically such information is distributed on diversely formatted and company websites and promotional literature.

Fig. 1. Instance data generated by Mutation Miner.

```
<Protein>
      <Name>xylanase</Name></Protein>
<Organisms>
      <Name>Bacillus circulans</Name></Organisms>
<label>PMID: 9930661: GI: 17942986</label>
<Mark>D37N</Mark>
<Context>The upward shift of the optimum pH of the D37N
mutant was predictable from the results of structural and
amino acid sequence comparison.</Context>
```

1.4 Enzyme Improvement

An additional need of the commercial enzymologist is access to information on mutational studies resulting in better enzymes. We discuss the inclusion of ontological concepts to support the instance data produced by the NLP tools [12] designed specifically to extract information on experimentally introduced mutations and their impact on protein performance. The ultimate goal being to interrogate the ontology regarding mutations resulting in improved enzyme performance under defined environmental conditions. Instances generated by the NLP tool are shown in Figure 1.

CONCLUSION

We have used semantic web technology to create ontology and a large knowledgebase in the domain of fungal biotechnology and genomics from trusted biological sources to provide unified semantic access to heterogeneous resources. We have demonstrated the capacity of the ontological conceptualization through a series of queries. Since our target audience is the decision making industry manager, not necessarily skilled in data mining technologies, we strive to facilitate answers without requiring advanced knowledge of query methodologies. We reason that sizeable time saving is made by and justifies the conceptual development of the ontology and its instantiation. Our semantic interrogations of the knowledge base provide us with further insight into structures of queries that the bio-scientific domain demands, thereby showing us the limits of the DL query technologies so that we can enhance the capabilities of Racer and nRQL.

ACKNOWLEDGEMENTS

This work was financed in part through the Genome Quebec project *Ontologies, the semantic web and intelligent systems for genomics* (V. Haarslev and G. Butler).

REFERENCES

[1] Sheban-Nejad A., Baker C. J. O., Butler G. Haarslev V. (2004) *The FungalWeb Ontology: The core of a Semantic Web Application for Fungal Genomics, 1st Canadian Semantic Web Interest Group Meeting (SWIG'04) Montreal, Quebec, Canada*

[2] Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA (2000). *Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2000 Jan 1;28(1):10-4*

[3] Phan, I. Q. H., Pilbout S. F., Fleischmann1 W. and Bairoch A. (2003) *NEWT, a new taxonomy portal, Nucleic Acids Research, Vol. 31, No. 13 3822-3823*

[4] Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. (2004) *BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res. Jan 1;32(Database issue):D431-3.*

[5] Saccharomyces Genome Database *http://www.yeastgenome.org/)*

[6] *Neurospora crassa* Database *(http://www.broad.mit.edu/annotation/fungi/neurospora/)*

[7] Bairoch A. *The ENZYME database in 2000* (2000) *Nucleic Acids Res 28:304-305*

[8] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000) *Gene ontology: tool for the unification of biology. Nat Genet, 25(1):25-9*

[9] Baker PG, Brass A, Bechhofer S, Goble C, Paton N, Stevens R. (1998) *TAMBIS--Transparent Access to Multiple Bioinformatics Information Sources. Proc Int Conf Intell Syst Mol Biol. 1998;6:25-34*

[10] Noy N. F., Sintek M., Decker S., Crubezy M., Fergerson R. W., & Musen M. A.. (2000) *Creating Semantic Web Contents with Protege-2000 IEEE Intelligent Systems 16(2):60-71,*

[11] Haarslev V, Möller R. (2001) *Description of the RACER System and its Applications. Proceedings of the International Workshop on Description Logics (DL-2001). Stanford, USA*

[12] Witte R and Baker C.J.O. (2005) *Combining Biological Databases and Text Mining to support New Bioinformatics Applicaions. A.Montoyo etal. (Eds.) NLDB 2005, LNCS 3513,310–321.*

Using Multiple Ontologies to Integrate Complex Biological Data

Mary Shimoyama, Victoria Petri, Dean Pasko, Susan Bromberg, Wenhua Wu, Jiali Chen, Nataliya Nenasheva, Simon Twigger, Howard Jacob*

Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, Wisconsin

ABSTRACT

Motivation: The strength of the rat as a model organism lies in its utility in pharmacology, biochemistry, and physiology research. Data resulting from such studies is difficult to represent in databases and creation of user-friendly data mining tools has proven difficult. The Rat Genome Database has developed a comprehensive ontology-based data structure and annotation system to integrate physiological data along with environmental and experimental factors, as well as genetic and genomic information. RGD uses multiple ontologies to integrate complex biological information from the molecular level to the whole organism, and to develop data mining and presentation tools. This comprehensive research platform will allow users to investigate the conditions under which biological processes are altered and to elucidate the mechanisms of disease.

1 ONTOLOGIES AT RGD

Initially, RGD used ontologies to provide a simple framework for classifying, representing and navigating across gene, phenotype, and disease information to link genomic data to function and disease (1, 2) and as a means to view biological information in the context of the genome. RGD implemented four ontologies: *G*ene *O*ntology (GO), *M*ammalian *P*henotype Ontology(MP), *D*isease *O*ntology(DO) and a *P*ath*W*ay ontology (PW). The MP was initially developed at Mouse Genome Informatics (3) and is now being developed in a collaborative effort between RGD and MGI. The disease ontology was adapted from the Medical Subject Headings (MeSH, 4) and the pathway ontology was developed at RGD in order to integrate data from existing pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (5), REACTOME (6), GenMapDB (7) and the Biomolecular Interaction Database (8), as well as pathway data found in the literature. It also includes "altered pathway" terms to allow for the representation of pathways whose events or interactions are altered by genetic or environmental factors.

2 ONTOLOGY BASED TOOLS

Current ontology based tools at RGD include the GViewer (Fig 1) which provides a genome-wide view of the genes and QTLs related to a single or multiple ontology query and GBrowse, which provides ontology tracks showing gene function, pathway, disease and phenotype information in the context of the genome.

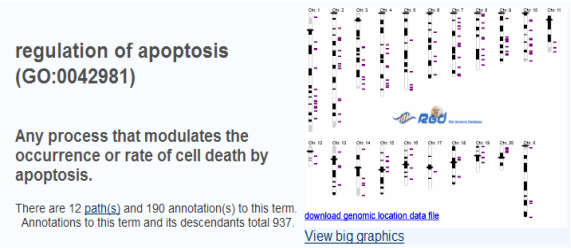
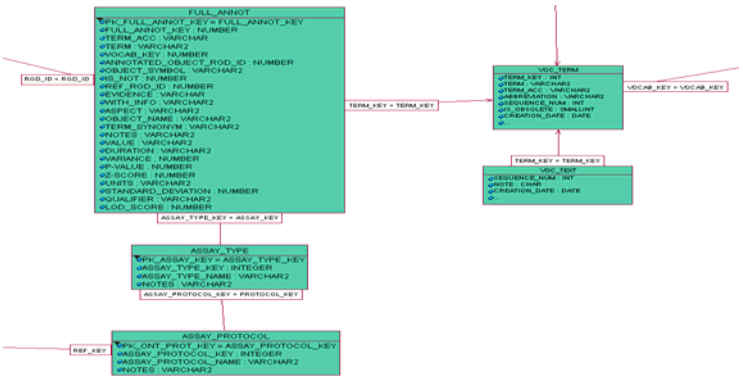


Fig 1. GViewer display of related genes across the genome

3 INTEGRATING PHYSIOLOGICAL DATA VIA MULTIPLE ONTOLOGIES

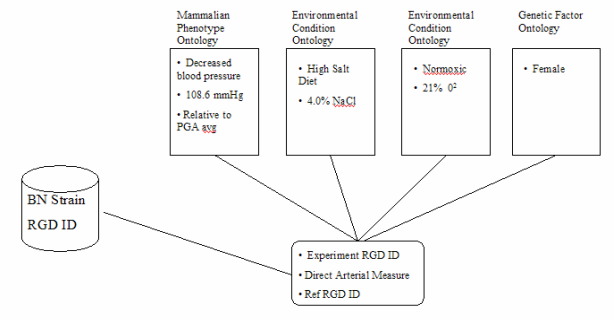
For the more complex data generated by much of the rat research community, the simple annotations provided by single ontologies are insufficient. They don't answer questions about the conditions under which the biological phenomena take place or what factors could inhibit or modify them. They also don't provide a mechanism for relating disparate types of biological information to allow researchers to elucidate patterns or mechanisms involved in disease. RGD developed a structure that would allow the integration of multiple ontology annotations as well as qualifiers and actual values into a single record. The relationships among multiple ontologies and values for a single annotation are achieved through an Experiment/Assay Table shown below.



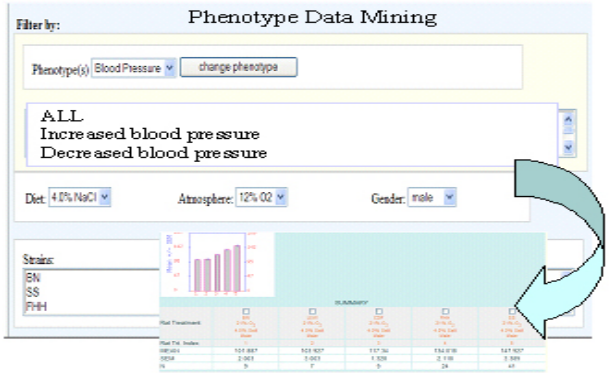
This compound annotation can be associated with the genomic elements in RGD such as genes, QTLs and strains, but it also can stand alone as an experimental record related

* To whom correspondence should be addressed.

to phenotypes, drugs, diseases, pathways and other physiological phenomena. Ontologies for rat anatomy, cell types, developmental stages, drugs, genetic factors, and environmental conditions as well as qualifiers are being added to the system for integration and representation of complex phenotype, disease, expression, pathway and pharmacological data. Phenotype annotations will thus include not only the phenotype ontology term, but also the actual value and the experimental and genetic factors involved. The use of the Experiment/Assay Data Object allows RGD to include pharmacological and physiological data that is not tied to a specific genetic or genomic object such as a gene or QTL.



Because the rat is used by a diverse community involved in physiological and disease research, investigators often are unsure of the best model to use to study particular phenotypes. By integrating environmental and genetic factors into our model, as well as the inclusion of actual values, RGD can provide phenotype analysis tools to aid the researcher in choosing appropriate models based on the phenotype and conditions of interest.



The multiple ontology data structure and annotation system supplies the user with an instant view of the processes, phenotypes, pathway(s) and environmental and genetic factors pertinent to a given disease. The design and implementation of additional sophisticated data mining tools for experimental data will allow investigators to more easily search for the answer to questions such as these:

Under what conditions is an increase in the severity of a phenotype or a change in the expression of a gene or mutant gene observed? Are diseases caused by associated with the same pathway different in their manifestations because of the differences in the nature of the alterations? Is, for instance, Notch signaling pathway compromised because the promoters of target gene are mutated, the receptors are not properly modified or because mutations in either receptor or ligand interfere with the normal activity? Are the manifold malformations (heart, eye, column) of the Alagille syndrome the result of the various instances of Jag1 ligand mutations, scattered across the entire gene? Are individuals affected with CADASIL condition more sensitive to environmental stress because the mutations within the Notch3 receptor irrevocably compromise a three-disulfide bond pattern and for this matter its structural integrity? It is precisely the ability to navigate between and link instances of expression, genetic and environmental attributes, where ontology annotations could help researchers unthread the interplay between genes, mutations and environment that underlie complex human diseases.

ACKNOWLEDGEMENTS

We are indebted to the Gene Ontology Consortium for their contribution to the use of ontologies for biological data and the staff at the Mouse Genome Database for the initiation of the Mammalian Phenotype Ontology. Special thanks to the curation and bioinformatics staff at the Rat Genome Database.

REFERENCES

1. Ashburner, M. and Lewis, S. (2002) On ontologies for biologists: the Gene Ontology--untangling the web. *Novartis Found Symp*, **247**, 66-80; discussion 80-3, 84-90, 244-52.
2. Stevens, R., Goble, C.A., and Bechhofer, S. (2000) Ontology-based knowledge representation for bioinformatics. *Brief Bioinform*, **1**, 398-414.
3. Smith, C.L., Goldsmith C.A., Eppig, J.T. (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.* **6** (1):R7.
4. Nelson, S.J., Johnston, D. and Humphreys, B.L., (2001) Relationships in Medical Subject Headings. in *Relationships in the organization of knowledge*, C.A. Bean, R. Green, and editors, Editors. Kluwer Academic Publishers: New York. p. 171-184.
5. Kanehisa, M. (2002) The KEGG database. *Novartis Found Symp*, **247**, 91-101; discussion 101-3, 119-28, 244-52.
6. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. 2005, Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. Jan 1:33.
7. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. (2002) Nat Genet May; 31(1):19-20.
8. Bader, G.D., Betel, D., and Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, **31**(1), 248-50.

A case study in knowledgebase verification

Stephen Racunas, Nigam Shah[*] and Nina V. Fedoroff

Huck Institutes of the Life Sciences, Penn State University, USA.

ABSTRACT

Background: Biological databases and knowledgebases are proliferating rapidly. In order to support human and computer-aided information integration and inference, a knowledgebase must be trustworthy. Further, it should structure information using an ontology that is expressive and well-structured enough to support computer-aided reasoning. Before using a pathway knowledgebase as a data source, it is therefore desirable to "proofread" it for trustworthiness and expressiveness.

Results: In this work we check the pathways stored in the Reactome knowledgebase to verify its trustworthiness and evaluate its usefulness for computer-aided reasoning. We make explicit the event language implicit in the Reactome ontology, and specify a complementary logic for this language. We use this logic to formulate a set of tests to verify desirable pathway properties in Reactome. We then perform such verification upon the latest two Reactome releases (10 and 11) and compare the results. We also discuss the expressiveness of the Reactome ontology and its potential for supporting computer-aided inference tools.

1 INTRODUCTION

The advent of high-throughput technologies has contributed to revolutionary increases in the volume and variety of data available to biologists, and many biological databases have been developed for storing and querying the rapidly accumulating data. However, locating, retrieving and integrating data have become increasingly burdensome tasks, and there is a growing need for tools that facilitate the interpretation of biological data. As a partial solution to this problem, some biological process databases represent information at a high enough level of abstraction to be designated "pathway knowledgebases." The pathway resource list (PRL) at www.cbio.mskcc.org/prl currently lists 167 pathway resources.

The structured information stored in pathway knowledgebases has the potential to be more immediately useful for computer-aided information integration than the raw data stored in more conventional databases because the information stored in such knowledgebases follows an explicit ontology. The usefulness of a pathway knowledgebase depends on characteristics we designate as trustworthiness and expressiveness. The trustworthiness of a knowledgebase is a measure of its quality and completeness, while the expressiveness of a knowledgebase is reflected in such properties as the complexity and sophistication of the queries it will support and its ability to represent biological systems at multiple scales. To be trustworthy, a pathway knowledgebase should be free of internal conflicts, explain as many steps as possible in each pathway, and provide the most complete set of pathway descriptions possible. Omissions, inconsistencies, errors in the order of steps in a pathway, missing steps, and extra steps all limit the utility of a knowledgebase. We believe that a trustworthy knowledgebase should minimally be complete, consistent, direct, gap-free, well-formed and acircular.

2 RESULTS

In this work, we give precise mathematical definitions for each of the above properties and present our logical framework for assessing a knowledgebase's trustworthiness using its ontology as a basis. We apply our tests to the latest releases of a deployed pathway knowledgebase (the Reactome knowledgebase) and thereby show how these tests can be used to proofread and fine-tune a knowledgebase while it is being developed.

The Reactome project, a collaborative effort involving the Cold Spring Harbor Laboratory, the European Bioinformatics Institute, and the Gene Ontology Consortium, is developing a knowledgebase comprising the core pathways and reactions in human biology. The information in the Reactome knowledgebase is authored by expert biological researchers and maintained by the Reactome editorial staff. The basic unit of Reactome is the reaction, defined as any biological event that converts inputs to outputs. We show that Reactome's event based ontology implicitly generates a formal language. We make this language explicit and then show how to define relationships between events. We then provide a definition of a model as expressed in the Reactome language and specify a logic under which we can perform model checking. We use this logical machinery to demonstrate how to check pathways by testing for certain

[*] To whom correspondence should be addressed.

desirable properties, the aggregate of which we defined as the trustworthiness of the knowledgebase.

We present the results of performing these tests on the two most recent releases of Reactome. We also discuss the concept of expressiveness and suggest modifications that can increase the expressiveness of the Reactome pathway knowledgebase. For the latest release of Reactome, we found that 14 pathways were incomplete (36 events) and 9 pathways had inconsistencies (24 events), corresponding to 3.7% incompleteness and 2.5% inconsistency. There were 5 pathways with gaps (6 events), 21 verbose pathways (57 events) and 3 terse pathways (3 events), corresponding to 0.6% gaps, 5.9% verbosity and 0.3% terseness. Of the 65 concrete human pathways in release 11, 30 were 80% well-formed and 43 were more than 50% well-formed.

We believe that evaluating knowledgebases for trustworthiness and expressiveness is an important step toward the overall goal of using knowledgebases like Reactome as effective resources for information integration and computer-aided reasoning about biological processes. This research facilitates our parallel efforts to develop a set of tools that will allow ontology based querying of existing information.

Ontology-based Statistical Analysis of Microarray Data

Aditya Vailaya[1], Allan Kuchinsky[1], Robert Kincaid[1], Annette Adler[1], Raymond Tabibiazar[2], Roger Wagner[2], Tom Quertermous[2]*

[1]Agilent Laboratories, 3500 Deer Creek Road, Palo Alto, CA-94304; [2]Donald W. Reynolds Cardiovascular Research Center, Division of Cardiovascular Medicine, Falk CVRC, Stanford University, 300 Pasteur Drive, Stanford, CA 94305

*aditya_vailaya@agilent.com

ABSTRACT

Motivation: Analyzing micorarray data in the context of biological processes can be a daunting task. Current algorithms are useful only in identifying statistically significant gene expression changes. However, a list of genes does not provide much insight to the biologist studying the underlying processes. In fact, identifying discriminatory pathways/networks of gene interactions from a set of significantly expressed genes can provide crucial information for understanding complex processes and identifying therapeutic targets. We present a method to analyze high throughput (HT) gene expression data based on three-way statistical analysis of ontological information for identifying interesting pathways.

1 METHOD

Given a list of interesting genes from microarray study (say output of SAM algorithm [1]) and a knowledge base of ontologies, our goal is to discover significantly over- and under-represented ontology terms. The ontology knowledge base was created from three different ontologies, namely Gene Ontology (GO, http://www.geneontology.org), curated pathways, and literature-based gene association networks. GO annotations were obtained using Biomolecule Naming Service (BNS) [2]. The curated pathway database contains 360 curated pathways collected at Stanford from various sources such as KEGG (http://www.genome.jp/kegg), BioCarta (http://www.biocarta.com/genes/allpathways.asp), and SPAD (http://www.grt.kyushu-u.ac.jp/spad/menu.html). A large association network (an association represents a relation among genes or proteins extracted from a sentence of text) was automatically constructed using BioFerret and ALFA [3] from over 350,000 PubMed abstracts. For each gene, "g", in the large network, a subnetwork was extracted consisting of "g" and its first neighbors, yielding 5,200 association networks, one for each gene.

1.1 Statistical Significance Score

Given a subset of differentially expressed ("interesting") genes, *l*, and a list of ontology (pathways/networks/GO) terms (each represented in terms of its genes), a statistical Z-score was computed for each term under a hypergeometric distribution assumption, as defined in [4].

$$Z(p,l) = \frac{(k - K\frac{n}{N})}{\sqrt{K\left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right)\left(1-\frac{K-1}{N-1}\right)}}$$

In this case, *K* is the total number of entries in the *N* microarray genes mapping to specific ontology term *p*, and *k* is the number of entries in *n* mapping to the same term *p*. The Z-score represents a surprise in finding "k" terms when we were expecting "*nK/N*" terms. A high positive value (signifying statistical over-abundance) or high negative value (signifying statistical under-abundance) of Z-score for an ontology term implies a significant surprise level, and hence, interestingness of the ontology term based on the experimental data. Ontology terms with |Z-score| >3 were considered significant.

2 RESULTS

We applied the statistical analyses to two microarray datasets.

2.1 Validation Study

We conducted a validation study on a mouse myocardial development dataset consisting of 16 samples, for ~20,300 genes per microarray, and comparing embryo tissue (9) to post-birth (7) samples. Our expectation was to test if the three different ontology-based analyses yield consistent results in terms of identifying significant biological processes in the two tissue types.

In the mouse myocardial development dataset, the three-way analysis validated that the embryo stage was extensively characterized with cell cycle and division processes (ontology terms corresponding to these processes were scored significantly higher in all the three analyses), whereas the post-birth stages over-abundantly represented cell respiration and metabolism processes (see Table 1.). We are encouraged by these results, especially for the literature-based ontology, since automatically extracted literature-based associations were not manually curated.

2.2 Mouse Heart Chamber Study

Encouraged by the results of the mouse developmental dataset, we analyzed a new microarray dataset consisting of mouse heart chamber samples. The data consisted of 12 samples, three samples per chamber (left and right atria and ventricle) for ~23,600 genes per microarray. We conducted three-way ontology-based statistical analyses on atria vs. ventricle samples to identify significant biological differences among these two tissue types.

In the mouse chambers data we discovered an over-abundant presence of ventricle up-regulated genes in metabolic pathways, such as oxidative phosphorylation and carbon fixation (see Table 2). We observed an over-abundant presence (albeit slightly weaker signal) of atrial up-regulated genes in signaling pathways, such as Ras signaling and IL-10 anti-inflammatory signaling pathway. We believe that these fundamental differences in gene expression in the two tissues may be explained by the inherent differences in the tissue functions. The higher energy expended in ventricular tissue for pumping blood probably explains why genes participating in energy metabolism are more active in ventricular tissue. Further, over-representation of signaling pathways in the atria tissue may suggest that atria are more susceptible to respond to external

GO Term	#Symbols	#Embryo-500	Z-Score
ribosome	176	51	10.47
cytosolic	32	16	8.98
replication	69	24	8.39
mitotic	41	16	7.50
ribonucleoprotein	39	12	5.35
mitosis	34	11	5.35
spindle	9	4	4.13
cycle	300	43	4.03

GO Term	#Symbols	#Post-Birth-500	Z-Score
nadh	27	20	13.74
ubiquinone	18	14	11.84
mitochondrion	235	62	11.54
dehydrogenase	104	36	11.00
chain	32	15	8.85
respiratory	17	10	8.41
tricarboxylic	14	8	7.38
malate	9	6	7.05

Pathway Term	#Symbols	#Embryo-500	Z-Score
Cell cycle - Homo sapiens	58	19	18.31
Cell Cycle: G1/S Check Point	14	6	11.90
Cycling of Ran in nucleocytoplasmic transpo	3	2	8.70
Selenoamino acid metabolism - Mus muscu	5	2	6.62
E2F1 Destruction Pathway	5	2	6.62
Cyclins and Cell Cycle Regulation	13	3	5.96

Pathway Term	#Symbols	#Post-Birth-500	Z-Score
Oxidative phosphorylation - Mus musculus	41	15	18.21
Carbon fixation - Mus musculus	13	6	13.04
ATP synthesis - Mus musculus	24	8	12.63
Pyruvate metabolism - Mus musculus	14	6	12.54
Citrate cycle (TCA cycle) - Mus musculus	10	5	12.42
Phenylalanine, tyrosine and tryptophan biosynthesis	5	3	10.60

Association	Mapped pathway	#Symbols	# Embryo	Z-Score
bub1	Cell cycle - Homo sapiens	10	7	16.70
cks1	Cell cycle - Homo sapiens	8	6	16.03
mcm7	Cell cycle - Homo sapiens	8	5	13.29
ccne2	Cell cycle - Homo sapiens	3	3	13.16
smarcb1	MAPKinase Signaling Pathway	3	3	13.16
mcm2	Cell cycle - Homo sapiens	12	6	12.93
rps19	Arginine and proline metabolism -	12	6	12.93

(a)

Association	Mapped Pathway	#Symbols	# Post-Birth	Z-Score
cox7c	Oxidative phosphorylation - Mus musculus	12	5	11.27
acadvl	Fatty acid metabolism - Mus musculus	5	3	10.60
acadm	Valine, leucine and isoleucine degradation	5	3	10.60
ndufs4	Oxidative phosphorylation - Mus musculus	5	3	10.60
ndufs2	Oxidative phosphorylation - Mus musculus	5	3	10.60
bgn	Butanoate metabolism - Mus musculus	6	3	9.62
cox6b	Oxidative phosphorylation - Mus musculus	3	2	9.14

(b)

Table 1. Ontology-based statistical analysis of mouse heart development data for GO, curated pathway, and literature-based association ontology terms; (a) and (b) represent statistically significant terms for 500 most discriminating genes in embryo (post-birth) stage (up-regulated in embryo (post-birth) and down-regulated post-birth (embryo)), respectively; curated pathway containing the largest subset of genes associated with a literature association term are also displayed with the literature association term.

Pathway Term	#Symbols	#VENTRICLE	Z-Score
Oxidative phosphorylation - Mus musculus	57	37	9.82
Citrate cycle (TCA cycle) - Mus musculus	13	12	7.33
Glycolysis - Gluconeogenesis - Mus musculu	38	23	7.28
Pyruvate metabolism - Mus musculus	24	15	6.04
Ubiquinone biosynthesis - Mus musculus	13	10	5.84
Proteasome - Homo sapiens	30	16	5.41
Reductive carboxylate cycle (CO2 fixation) -	5	5	5.01
Valine, leucine and isoleucine degradation -	24	13	4.94

Association	Mapped Pathway	#Symbols	#VENTRICLE	Z-Score
ndufs2	Oxidative phosphorylation - Mus musculus	8	8	6.34
eno1	Glycolysis - Gluconeogenesis - Mus musculus	14	11	6.23
psmb5	Proteasome - Homo sapiens	12	10	6.21
psmb2	Proteasome - Homo sapiens	12	10	6.21
ndufs8	Oxidative phosphorylation - Mus musculus	21	14	6.16
pfkl	Fructose and mannose metabolism - Mus musculus	7	7	5.93
ndufs7	Oxidative phosphorylation - Mus musculus	7	7	5.93
ndufv1	Oxidative phosphorylation - Mus musculus	7	7	5.93
ndufa6	Oxidative phosphorylation - Mus musculus	7	7	5.93

(a)

Pathway Term	#Symbols	#ATRIAL	Z-Score
Activation of PKC through G protein coupled receptor	3	3	4.01
Ras Signaling Pathway	16	8	3.77
Alternative Complement Pathway	2	2	3.27
Overview of telomerase protein component gene hTert Transcriptional	4	3	3.25
IL-10 Anti-inflammatory Signaling Pathway	7	4	3.01
Thrombin signaling and protease-activated receptors	7	4	3.01
Signaling Pathway from G-Protein Families	10	5	2.98
beta-Alanine metabolism - Mus musculus	10	5	2.98

Associati	Mapped Pathway	#Symbols	#ATRIAL	Z-Score
tnf	NF-kB Signaling Pathway	138	54	7.60
arf1	Rac 1 cell motility signaling pathway	24	17	7.42
cd48	IL 18 Signaling Pathway	21	14	6.42
wnt	WNT Signaling Pathway	132	46	6.07
col3a1	Multi-step Regulation of Transcription by Pitx2	13	10	6.06
lbp	Inactivation of Gsk3 by AKT causes accumulation of b-catenin in	9	8	6.03
cd9	Cell cycle - Homo sapiens	37	19	5.96
emb	Corticosteroids and cardioprotection	16	11	5.83

(b)

Table 2. Ontology-based statistical analysis of mouse heart development data; (a) and (b) represent significant ontological terms for most discriminating 1,715 ventricle and 1,015 atria genes, respectively; only pathway and literature-based ontology results are shown.

stimuli (say therapeutic) than ventricles. We also note that while the curated pathway analysis yields much lower Z-scores (only 6 curated pathways having score greater than 3), literature-based associations yield a number of sets of genes with higher Z-scores. We believe that although literature-based associations can be erroneous, they bring in important current information not yet manually curated into pathways or GO categories. Finally, combining results from all the three analyses can yield significant clues regarding biological processes that may be playing a major role in the disease under study. New experiments can be designed to validate these biological processes, identify biomarkers for disease prognosis, and identify therapeutic targets for cure.

3. CONCLUSIONS

We have developed statistical methods for ontology-based analysis of microarray data. We present a three-way analysis method for identifying biological processes that may play a role in the experimental condition under study. Application to two microarray datasets demonstrates the significance of these analyses in better understanding of HT gene expression data.

REFERENCES

1. Tusher et al. (2001), "Significance analysis of microarrays applied to the ionizing radiation response", PNAS 2001, 98: 5116-5121.
2. Kincaid et al. (2002), "BNS: An LDAP-based Biomolecule Naming Service", OiBC2002, Washington DC, http://openbns.sourceforge.net/.
3. Vailaya et al. (2005), "An architecture for biological information extraction and representation", Bioinformatics 2005, 21(4):430-438.
4. Doniger et al. (2003), "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data", Genome Biology, 4(I), Article R7, 2003.

Ontology-driven similarity approaches to supporting gene functional assessment

Francisco Azuaje[1,], Haiying Wang[1] and Olivier Bodenreider[2]*

[1]*University of Ulster, Northern Ireland,, UK*
[2]*National Library of Medicine, Bethesda, USA*

ABSTRACT

Motivation: Bio-ontologies, such as the Gene Ontology, represent important sources of prior knowledge that may be automatically integrated to support predictive data analysis tasks. The assessment of similarity of gene products provides the basis for the implementation of classification tools and the automated validation of functional associations. This study discusses alternative techniques for measuring ontology-driven similarity of gene products. Relationships between these types of similarity information and key functional properties, such as gene co-expression, are discussed.

1 INTRODUCTION

Bio-ontologies represent important knowledge bases, which have traditionally been applied to enhance database annotation and interoperation as well as cross-database information retrieval tasks. The *Gene Ontology™* (GO) (The Gene Ontology Consortium, 2001) is one such resource that is becoming the *de facto* standard for annotating gene products.

The relevance of the GO goes beyond annotation and information retrieval applications. It has been shown that GO may facilitate large-scale predictive applications in functional genomics. The analysis of GO annotations in gene expression analysis may help to explain why a particular group of genes share similar expression patterns. Several tools have been proposed to identify functionally-enriched clusters of genes. *FatiGO* (Al-Shahrour *et al.*, 2004), for example, extracts GO terms that are significantly over- or under-represented in clusters of genes. GO-based annotations have been incorporated to construct functional predictors that in combination with other information resources have shown to improve functional association prediction (e.g. protein-protein interactions) (Jansen *et al.*, 2003). Hvidsten *et al.* (2003) combined gene expression data with annotations originating from the GO biological process taxonomy. They applied *rough set theory* to assign biological process terms to genes represented by expression patterns. King *et al.* (2003) implemented *decision trees* and *Bayesian networks* to predict new GO terms-gene associations based on existing annotations from the SGD and FlyBase. Although these functional prediction tools process GO annotations they do not fully exploit the knowledge that can be extracted from analyzing relations of GO terms and their information content in different annotation databases. For instance, traditional functional prediction support or cluster analysis tools mainly process information about the frequency of individual annotation terms associated with a list of genes. Furthermore, such applications may be improved by explicitly considering similarity relationships between the genes, which may be estimated by analyzing both the information content and structure of the GO. It has been suggested that by ignoring such *semantic similarity* between closely related GO terms (e.g., between a parent and a child), traditional methods may fail to identify the functional similarity between genes annotated with these closely related yet distinct terms.

Thus, the GO has been proposed as a tool for measuring similarity between genes. Previous research showed significant relationships between semantic similarity of pairs of genes and their sequence-based similarity (Lord *et al.*, 2003). Also we have evaluated relevant quantitative relationships between GO-driven similarity and gene expression correlation (Wang *et al.*, 2004). GO-driven clustering algorithms based on such approaches have been recently reported (Wang *et al.*, 2005, Speers *et al.*, 2004). Moreover, they have provided the basis for developing tools that may facilitate the identification of relevant partitions from clustering, using, for example, GO-driven cluster validity indices (Bolshakova *et al.*, 2005)

This paper discusses our current research on the design of GO-driven similarity assessment techniques. It aims to compare two approaches to estimating between-gene similarity, which may be implemented using different schemes for measuring between-term similarity. Relationships between semantic similarity and gene co-expression are further investigated taking into account both approaches.

2 SEMANTIC SIMILARITY APPROACHES TO ASSESSING GENE SIMILARITY

Given a pair of terms, c_1 and c_2, a traditional method for measuring their similarity consists of calculating the distance between the nodes associated with these terms in the

* To whom correspondence should be addressed.

ontology, whose limitations have been discussed elsewhere (Zhong *et al.*, 2002). *Information-theoretic* models have been studied as alternative approaches to measuring similarity in an ontology. Let C be the set of terms in the GO. Information-theoretic approaches to measuring similarity between terms, $c \in C$, may be based on the *amount of information* associated with them or shared by them in common. Several techniques may be implemented using this principle, such as those proposed by Lin, Resnik and Jiang (Lord *et al.*, 2003, Wang *et al.*, 2004). Similarity (or distance) values for a pair of gene products described by GO terms may be calculated based on such techniques (Lord *et al.*, 2003, Wang *et al.*, 2004). Given a pair of gene products, g_i and g_j, which are annotated by a set of terms A_i and A_j respectively, where A_i and A_j comprise m and n terms respectively, the semantic similarity, $SIM(g_i, g_j)$, may be defined as the average inter-set similarity between terms from A_i and A_j:

$$SIM(g_i, g_j) = \frac{1}{m \times n} \times \sum_{c_k \in A_i, c_p \in A_j} sim(c_k, c_p) \qquad (1)$$

where $sim(c_k,c_p)$ represent the similarity between terms. This approach does not always meaningfully estimate similarity. For example, similarity is expected to be equal to 1 when the gene pair has the same set of annotation terms. However, this is not true when several annotations within a hierarchy are assigned to the genes. In order to address such a limitation we are currently evaluating an alternative approach that selectively aggregates maximum inter-set similarity values as follows:

$$SIM(g_i, g_j) = \frac{1}{m+n} \times \left(\sum_k \max_p (sim(c_k, c_p)) + \sum_p \max_k (sim(c_k, c_p))\right) \qquad (2)$$

2.1 Linking semantic similarity and other functional properties

The analysis of quantitative relationships between semantic similarity and other functional information resources is important to allow the identification of novel integrative prediction strategies. Such relationships may indicate whether semantic similarity may be combined with other large-scale predictive resources (e.g. gene expression correlation, sequence binding patterns, etc.) to improve key functional prediction factors, such as accuracy and coverage. Based on (1) previous research has confirmed that GO-driven similarity and expression correlation of pairs of gene products in *S. cerevisiae* are significantly interrelated (Wang *et al.*, 2004). This property has shown to be consistently valid for similarity information originating from all of the GO hierarchies. We are currently analyzing these relationships using (1) and (2) on the latest GO annotation release for *S. cerevisiae.*

We are assessing relationships between semantic similarity and other functional properties such as gene co-regulation and protein-protein interactions in *S. cerevisiae* and *C. elegans*. One of our hypotheses is that the GO-driven similarity of a pair of genes may be used as an indicator of regulatory and protein-protein interactions.

Furthermore, we are investigating how GO-driven semantic similarity may be applied to support the detection of spurious (co-regulation or protein-protein) interaction predictions. After studying this, one could in principle justify the design of prediction support tools for co-regulation and protein-protein interactions, which in combination with other resources, e.g. co-expression, may support a more accurate and biologically meaningful identification of functional networks.

REFERENCES

The Gene Ontology Consortium (2001) Creating the gene ontology resource: Design and implementation. *Genome Research*, **11**, 1425-1433.

Al-Shahrour,F., Diaz-Uriarte,R., and Dopazo,J. (2004) Fatigo: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578-580.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302** (17), 449-453.

Hvidsten, T., Lægreid, A. and Komorowski, J. (2003) Learning rule-based models of biological process from gene expression time profiles using Gene Ontology. *Bioinformatics*, **19**, 1116-1123.

King, O. D., Foulger, R. E., Dwight, S. S., White, J. V., and Roth, F. P. (2003) Predicting gene function from patterns of annotation. *Genome Research*, **13**, 896--904.

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

Wang,H., Azuaje,F., Bodenreider,O., and Dopazo,J. (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Proc. of IEEE 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, La Jolla, CA, USA, 25-31.

Wang,H., Azuaje,F., and Bodenreider,O. (2005) An ontology-driven clustering method for supporting gene expression analysis. In *Proc. of the 18th IEEE International Symposium on Computer-Based Medical Systems*, in press.

Speer, N., Spieth, C. and Zell, A. (2004) A memetic clustering algorithm for the functional partition of genes based on the gene ontology. In the *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, San Diego, USA, 252-259.

Bolshakova,N., Azuaje, F., and Cunningham,P. (2005) A knowledge-driven approach to cluster validity assessment. *Bioinformatics*, Advance Access published on February 15, 2005.

Zhong, J., Zhu, H., Li, Y. and Yu, Y. (2002) Conceptual graph matching for semantic search. In *Proc. of Conceptual Structures: Integration and Interfaces*, 92-106.

Plant structure and growth stage ontologies to describe phenotypes and gene expression in angiosperms

***Pankaj Jaiswal (A)**, Shulamit Avraham (B), Katica Ilic (C), Elizabeth A Kellogg (D), Susan R McCouch (A), Mary Polacco (G, E), Anuradha Pujar (A), Leonore Reiser (C), Seung Y Rhee (C), Marty Sachs (F), Lincoln Stein (B), Peter Stevens (D), Leszek Vincent, Doreen Ware (B, E), Felipe Zapata (D)*

(A) Cornell University, Ithaca, NY 14853; (B) Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; (C) Carnegie Institution of Washington, Stanford, CA 94305; (D) University of Missouri at St. Louis, St. Louis MO 63121;(E) University of Missouri - Columbia, Columbia, MO 65211; (F) Maize Genetic Cooperation - Stock Center, Department of Crop Sciences - University of Illinois, Urbana, IL 61801.(G) USDA-ARS

ABSTRACT

Plant Ontology Consortium (POC) (www.plantontology.org) is a collaborative effort of several plant databases and experts in plant systematics, botany and genomics. A primary goal of the POC is to develop simple, yet robust and extensible controlled vocabularies that accurately reflect the biology of plant structures and developmental stages. These provide a network of vocabularies linked by relationships to facilitate meaningful cross-species queries across datasets from various species or from plant databases. The current ontology release integrates the diverse vocabularies in use to describe Arabidopsis, maize, rice and Triticeae anatomy, morphology and growth stages. This integration spans two major flowering plant taxonomic divisions namely, the monocots and eudicots. Using the ontology browser, over 3500 gene annotations from three species-specific databases, TAIR, Gramene and MaizeGDB, can now be queried and retrieved. In the presentation, with the help of examples from member databases, we will demonstrate how the PO supports gene discovery, phenotype prediction, and gene product functions, as well as the organizing principles and rules followed in developing the Plant Ontologies. The project is supported by National Science Foundation grant No. 0321666.

* To whom correspondence should be addressed.

Towards a Biochemical Ontology for Functional Classification of Arbitrary Compounds

Dan McShan

University of Colorado Schoool of Medicine, Aurora, CO

ABSTRACT

Motivation: In order to predict the metabolic fate of an arbitrary compound based solely on structure, it is useful to be able to identify substructural "functional groups" which are biochemically reactive. These functional groups are the substructural elements that can be removed and replaced to transform one compound into another. This problem of identifying functional groups is related to the problem of classifying compounds. The research presented here discusses the state of the art in biochemical databases and how these sources may be applied to the problem of classifying compounds based solely on structure.

1 INTRODUCTION AND BACKGROUND

This research is mainly focused on developing novel methods for predicting the metabolic fate of compounds in an organism. While networks of reactions do exist in sources like KEGG and MetaCyc, it is not obvious how one might predict the metabolic fate of a xenobiotic from these networks. This project discusses our research in identification of functional groups in arbitrary compounds, and using these groups to classify the compounds. These groups and classes can then be used to help define substrate specificities for many enzymes. For instance, "alcohol dehydrogenase" should be a plausible catalyst for an alcohol. The question is whether or not a compound is an alcohol or at least has the alcohol functional group.

2 BACKGROUND

Previously, we have shown that these databases of endogenous metabolism can be used to aid in the prediction of xenobiotic metabolism. This is because the KEGG database curates "abstract reactions" which can be converted into rules and applied to specific compounds. For instance, the reaction that converts "alcohol" into "aldehyde" can be applied to a xenobiotic alcohol (in our example, furfurol) to produce the aldehyde form (furfural). However, specificity was found to be quite a problem. For instance, not all –OH moieties are functionally active as alcohols. Nevertheless, the substructural search/replace algorithm seems to be sufficient to capture many transformations of interest.

Another aspect of the specificity problem is that of classifying compounds. Fundamentally, the fact we are looking to capture are of the form "ethanol is an alcohol." These relationships are not curated in KEGG. The MetaCyc ontology also lacks this fact, though it is ontology based. MetaCyc has "ethanol is an unclassified-compound" . MetaCyc does have a class "an alcohol", which is also an unclassified-compound.

The KEGG database curates 835 compounds which have "abstract" structures – i.e. the chemical formula contains a Markush ("R") group. KEGG compound C00226 has formula CH3OR. The MetaCyc database, by comparison, lacks formula for the alcohol abstraction.

It is worth noting that in the MeSH classification, ethanol is a child of "alcohols". In MeSH, however, the parent-child relationship is not obviously an "isA" type relation. Consider that ethanolamines is a child of ethanol, and the statement "epinephrine is a ethanol" is not entirely true. Clearly there is some relationship, but it is more of a substructural "hasA" type relation. MeSH, of course, has no chemical structure nor links to other datasources. However, NCBI has a database of compounds called PubChem which does have links to the MeSH tree for specific compounds like ethanol, but not for classes like alcohol.

EBI has recently adopted Ashburner's chemical ontology as ChEBI. The ontology here is also a directed acyclic graph, and we can see, for instance that ethanol is a great-grandchild of "alcohols". It is interesting to note that ethanol is a child of ethanols, and is sibling to chloroethanols and (1S)-1-phehylethanol. Phenylethanol is a child of ethanol in MeSH. ChEBI does contain links to KEGG for specific instances like ethanol, but not for abstract classes like alcohol (or "ethanols" for that matter).

To summarize, we would like to identify functional groups in arbitrary structures. This capability should allow us to generate the ontological relationships expressed in resources like MeSH and ChEBI. Existing ontologies tend to be DAGs, and are quite limited in their expressive power. Since they only allow a single "parent-child" relationship, this gets uses in semantically sloppy ways. This lack of precision in defining the relationship makes using these existing ontologies challenging.

* To whom correspondence should be addressed.

3 APPROACH

Our approach is to integrate the above databases in a single ontology. Using the compound structures from KEGG and PubChem, we intend to improve on the unspecified relationships in the MeSH and ChEBI ontologies. For instance, phenylethanol hasSubstructure ethanol by virtue of the fact that the structure of phenol contains the structure of ethanol (give or take a few hydrogens). Using advanced cheminformatics tools, we intend to define the substructural relationship between related nodes in MeSH and ChEBI.

A preliminary attempt at this problem simply attempted to find substructural matches of the 835 abstract compounds in KEGG within the other 10,000+ compound in the database. This resulted in 120,455 substructural relationships. What we found, however, was that substructure alone is not sufficient to identify functional groups. This was expected – consider that the –OH in a carboxy does not behave as an alcohol. Furthermore, substructures are not sufficient, nor even necessary in some cases for compound classification. The classic example of the latter is the fact that proline is classified as an “amino acid” by ChEBI (“proline is child of glutamine family amino acids is child of amino acids”). Proline, however, is NOT an amino acid. It is an imino acid. This fact is actually encoded by MeSH – proline is a child of imino acid. Unfortunately, imino acid is also a child of amino acid, confusing the issue. Interestingly, imino acid is not even in ChEBI, although “imino group” is present, but unclassified.

Currently, we are exploring biochemical descriptors of the immediate milieu around substructural matches to see if they can help to distinguish reactive groups from nonreactive ones. This should aid in determining substrate specificities for promiscuous enzymes.

The classification problem is a bit trickier, since the classes are human defined, and not always rational. Proline is NOT an amino acid in the strict biochemical sense. However, it does behave similar to the other amino acids, and participates in similar reactions. We think that perhaps the function of the compound – e.g. what types reactions it participates in – may be a distinguishing characteristic. For instance, the fact that Proline participates in a transferase reaction with tRNA is indicative that it’s a protein building block. Similarly, the presence of the tRNA(Pro) compound also indicates the same function.

We are presently exploring these other avenues in our attempt to characterize and classify biochemical compounds.

ACKNOWLEDGEMENTS

The author thanks Dr. Imran Shah for vision and support in early phases of this research.

REFERENCES

Brooksbank, C.; Cameron, G. & Thornton, J. (2005), 'The European Bioinformatics Institute's data resources: towards systems biology.', Nucleic Acids Res 33(Database issue), D46-53.

Goto, S.; Nishioka, T. & Kanehisa, M. (1998), 'LIGAND: chemical database for enzyme reactions.', Bioinformatics 14(7), 591-599.

Karp, P.D.; Riley, M.; Paley, S.M. & Pellegrini-Toole, A. (2002), 'The MetaCyc Database.', Nucleic Acids Res 30(1), 59-61.

McShan, D.; Updadhayaya, M. & Shah, I. (2004), 'Symbolic inference of xenobiotic metabolism.', Pac Symp Biocomput, 545-56.

Toward Developing a Provenance Ontology for Biological Images

Sudha Ram and Jun Liu
Department of Management Information Systems, University of Arizona, USA

Nirav Merchant, Terrill Yuhas, and Patty Jansma
Arizona Research Laboratories, University of Arizona, USA

Abstract
Recording and maintaining metadata of biological images has been a challenging issue in life sciences. Our research focuses on managing provenance, an important type of metadata for biological images. We develop an ontology that captures the semantics of provenance for biological images. We also describe a software system that helps record and maintain provenance in a convenient way.

1 INTRODUCTION

A major challenge in life sciences is devising ways to manage the vast amounts of biological images generated using high throughput imaging devices and techniques such as confocal and electron microscopy, and Magnetic Resonance Imaging (MRI) to name a few. To ensure timely analysis, biological images need to be easily accessed and interpreted, necessitating that metadata (background description about the data) associated with images is accurately and efficiently captured, recorded, and represented. The metadata of a biological image should include information about the history and pedigree of the data [1] including facts such as who or what processes created the image, what initial sources were used, what instrument recorded it along with machine specific settings and parameters, when it was created and/or used, etc. We refer to all of this background as "provenance". Knowing the provenance of biological images is extremely important for scientific purposes because it helps assess the quality and usefulness of the images, and it also enables scientists to analyze the images in context. Therefore, to ensure that images, created by different techniques, are ready for scientific use, it is imperative that the provenance of the images be captured and easily accessible to their users.
Our research aims to develop a provenance ontology that captures the semantics of provenance for biological images. Our ontology represents different elements of provenance (e.g. instrument, processing procedures, people involved, storage location, etc.) and their relationships to each other. Working in collaboration with researchers from the Arizona Research Laboratories at the University of Arizona, we are developing a software system that records and maintains provenance of biological images in a convenient manner.

2 RELATED WORK

Current efforts to capture and record metadata on biological images include efforts such as the OME framework and the BioImage Database Project. The latter project is developing a collection of bio images recorded by various microscopic techniques relevant to life sciences [4]. In this project, a BioImage ontology has been designed to optimize the submission and retrieval of biological images. While this ontology focuses on recording the content of bio images, it also describes multi-media objects as well as scientific experiments [2]. Our work is complementary to this effort; our ontology focuses on the provenance rather then the content of bio images. Besides documenting experimental details involved in the generation and/or acquisition of images, our ontology records various provenance events in the image life cycle including creation, usage, and transformation of the images and documents different elements related to these provenance events.
Our research is also influenced by research on provenance of biological data. In the last decade, significant research has been carried out on describing the provenance of data in biology and genetics database such as SWISSPROT and OMIM. As an example, the ^{my}Grid project captures provenance of bioinformatics data by depicting the workflow of *in silico* experiments. The derivation path of a workflow records the process used to transform input data [3]. Our ontology goes beyond workflow provenance and captures a wide range of elements of provenance including details of who or what process created or processed the image, where, when, with what instrument, for what purpose.
Adoption of open standards for microscopy by equipment manufacturers and ability to store relevant meta data for further analysis have given rise to frameworks such as the open microscopy environment (OME) [5]. Efforts will be directed towards integrating our tools with community resources like the OME framework, supplementing its attributes and semantic types.

3 OUR PROVENANCE ONTOLOGY

We develop an ontology to represent the provenance of biological images. The structure of our ontology is shown in Figure 1. In developing this ontology, the domain of provenance has been conceptualized as a combination of seven interconnected elements. The element "what" captures provenance events such as creation, usage and transformation that can happen to a biological image. The element "why" describes the causes that can be attributed to a provenance event. The element "how" describes the processes that led/lead to a provenance event. The element of "with what" refers to the instruments that recorded or processed the data. The element "who" records information about people who create, use, access, and/or process an image. Finally, the element "when" defines the occurrence time of a provenance event, and the element "where" describe the storage location or source of an image. Each element of our provenance may be further classified into component elements. As shown in Figure 1, a process that

creates or transforms the data (represented by the node labeled "how" in the ontology) may include filtering, recording, synthesizing, etc. The instrument used in image creation may be a confocal microscope that has components such as a specific type of lens and/or a filter. With this design, our ontology provides answers to questions such as "who created the image", "what processes were used to create the image with what instruments", "why and when was the bio image created", "where has the image file been stored".

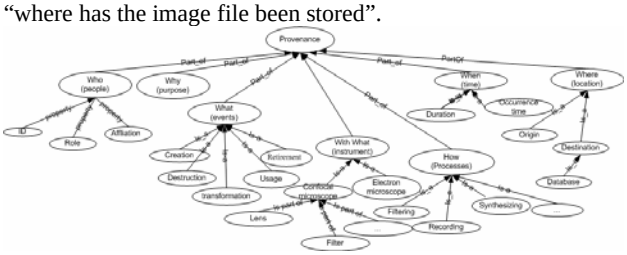


Fig. 1. A part of the provenance ontology

4 A SOFTWARE SYSTEM

We are in the process of developing a **Pro**venance **M**anagement **S**yst**e**m for Biological **I**mages (PROMISE) that utilizes the ontology to capture the provenance of biological images. The architecture of PROMISE is shown in Figure 2. Based on the ontology, the Provenance Capture Module in our system will generate a provenance template for different types of images to semi-automatically capture user-provided provenance. The captured provenance elements will be automatically recorded in XML/RDF format and stored in the data provenance knowledge base. A unique feature of PROMISE is that it will provide both image files and provenance to the user simultaneously. The user may retrieve various types of images stored in the database. Along with the images, the relevant provenance will be extracted from the provenance knowledge base and provided to the user upon request. Via a web-based graphical user interface, the Provenance Navigation Module in PROMISE will parse the provenance data, display the provenance graphically and enable the user to easily navigate and also modify/enhance the provenance if necessary. We also propose to develop a mechanism to capture the provenance automatically. For instance, every time a set of images are used to create other composite images, the provenance of all the images involved in the process will be updated. Anytime an image is accessed, its provenance will be update to reflect the access. We believe it is especially important to create this two way link between the data (bioimages) and the provenance.

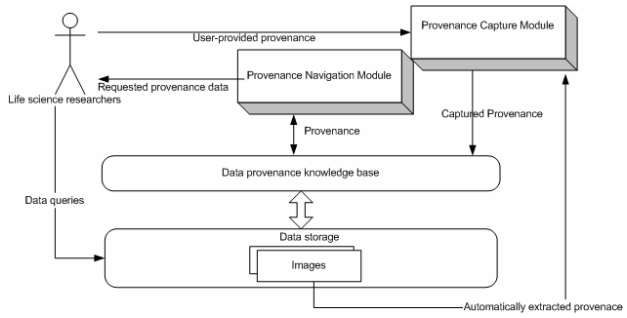


Fig. 2. Architecture for the provenance management system (PROMISE).

5 CONCLUSION

In summary, we are developing an ontology for capturing and representing the provenance of biological images. Our ontology extends existing approaches to recording biological image metadata that focus on describing the content of images. Our ontology captures various elements of provenance including the "who", "what", "why", "where", "when" and "how" of biological images. The recorded provenance will permit scientists to interpret the images in context and also to replicate and validate the procedure that created or processed the images. We are also developing a software system based on the ontology for automatically or semi-automatically capturing and maintaining provenance of biological images, which is essential to life scientists who are seeking a convenient way of recording a large amount of metadata. We believe that development of an automatic or semi-automatic provenance capture and deployment system will significantly reduce human effort in creating and managing provenance. It will also help to improve the quality and accuracy of the metadata by eliminating possible human mistakes.

REFERENCES:

[1]Buneman, P., Khanna, S., Tajima, K. and Wang, C.T.(2002) Archiving Scientific Data, *Proceedings of ACM SIGMOD International Conferences on Management of Data*, San Diego, California.

[2]Catton, C., Sparks, S., Shotten, D. (2004) Using ontologies to provide semantic richness in biological image databases *Abstract for Seventh Annual Bio-Ontologies Meeting.* Glasgow, UK.

[3]Greenwood, M., Goble, C., Stevens, R. Zhao, J., Addis, M., Marvin, D., Moreau, L. and Oinn, T. (2003) Provenance of e-Science Experiments – experience from Bioinformatics. *Proceedings UK e-Science All Hands Meeting.* Nottingham, UK.

[4]Linbek,S., Fritsch, R., Machtynger, J., de Alarcon, P., and Chagoyen, M. (1999) Design and realization of an on-line database for multidimensional microscopic images of biological specimens. *Journal of structural biology*, 125, pp. 103-111.

[5]Swedlow JR, Goldberg I, Brauner E, Sorger PK. (2003) Image informatics and quantitative analysis of biological images. Science 300:100-102.

Linking the Gene Ontology to other biological ontologies

Olivier Bodenreider[1,] and Anita Burgun[2]*

[1] National Library of Medicine, Bethesda, USA
[2]EA 3888, IFR 140, Université Rennes 1, France

ABSTRACT

The entities described in the Gene Ontology, (i.e., molecular functions, cellular components and biological processes), often make reference (in their names) to other entities, either from GO or from other ontologies, such as ontologies of chemical entities, cell types and organisms. We developed a method for mapping terms from the Open Biomedical Ontology (OBO) family to GO. We show that 55% of the 17,250 GO terms include in their names the name of some chemical entity (ChEBI). Our findings are consistent with that of other studies. Additionally, our study provides a quantification of the relations between GO terms and terms from other ontologies.

1 INTRODUCTION

Several approaches have been used to identifying relations among terms form the Gene Ontology (GO[1]) [1]. The lexical approach developed by Ogren et al. exploits the compositional properties of GO terms, i.e., GO terms nested within other GO terms [2]. They found that 65% of all GO terms contain another GO term as a proper substring. For example, the molecular function *electron transporter activity* includes in its name the biological process *electron transport*.

The goal of this study is slightly different: it is to investigate the degree to which GO terms are related to terms from ontologies external to GO. In particular, we are interested to make explicit the relations existing between GO terms and terms from other ontologies of the Open Biomedical Ontology (OBO) family[2]. OBO includes ontologies such as ChEBI (Chemical entities of biological interest), InterPro (protein families, domains and functional sites) and Plant ontology (plant structures and growth/developmental stages).

Related to this study is Obol [3], a language created for representing relations embedded in the names of GO entities, with the objective of facilitating the maintenance of the ontology. The work most closely related to ours is the GONG project [4], an attempt to convert GO into a description logics formalism. In addition to GO terms themselves, GONG also used entities from the KEGG database as a reference for the enzymes referenced in GO. The objective of this study is to generalize such cross-references (i.e., between GO and other ontologies) to all entities represented in OBO ontologies.

As suggested by Smith & al. [5], GO entities must be linked to entities in external ontologies such as cell types (e.g., *alpha-beta T-cell activation*) and organisms (e.g., *light-harvesting complex (sensu Viridiplantae)*). In a previous study [6], we investigated the relations between GO and ChEBI. This paper proposes to generalize the method developed for ChEBI to other members of the OBO family.

2 LINKING GO TO CHEBI

The first phase of this project consisted to link GO terms to chemical entities from the Chemical Entities of Biological Interest (ChEBI). ChEBI is "a freely available dictionary of 'small molecular entities' (i.e., atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc.); ChEBI entities are either products of nature or synthetic products used to intervene in the processes of living organisms." ChEBI is developed at the European Bioinformatics Institute (EBI). ChEBI names were extracted from the OBO file dated December 22, 2004. Both preferred names (*name* field) and synonyms (*synonym* field) are used in this study. A total of 27,097 names were extracted from the file (13,709 synonyms in addition to one preferred name for each of the 10,516 entities). For example, names for the ChEBI entity identified by CHEBI:26216 include the preferred name *potassium* and two synonyms: *kalium* and *K*.

2.1 Methods

Every ChEBI name is searched for in every GO name (Figure 1). ChEBI names of less than three characters are ignored. These names often correspond to chemical symbols (e.g., *K*, symbol of potassium) and may be ambiguous with words in English (e.g., *As* – symbol of arsenic – and the preposition *as*). As the names of ChEBI entities may be capitalized, the comparison between ChEBI and GO strings is rendered cased-insensitive. In order to avoid infelicitous matches, the name of a ChEBI entity is required to be not simply a substring, but a lexical item. In practice, the characters surrounding the name of the ChEBI entity in a GO name must be word boundaries (i.e., space, hyphen, punc-

* To whom correspondence should be addressed.

[1] http://www.geneontology.org/

[2] http://obo.sourceforge.net/

tuation, etc.). For example, the ChEBI entity *carbon* is identified in the GO name *carbon-oxygen lyase activity*, but not in *carbonic anhydrase activity*. Finally, we performed a limited normalization of the ChEBI names, principally to allow the names of classes of entities – often in plural form (e.g., cations, acids, esters, nitrates, etc.) to match names of entities derived from these classes, often present in singular form as in GO names. In practice, we complemented the list of synonyms provided by ChEBI by adding, if necessary, the singular form for the name of a plural class (e.g., *ester* for *esters*). 2,872 such synonyms were added to ChEBI[3].

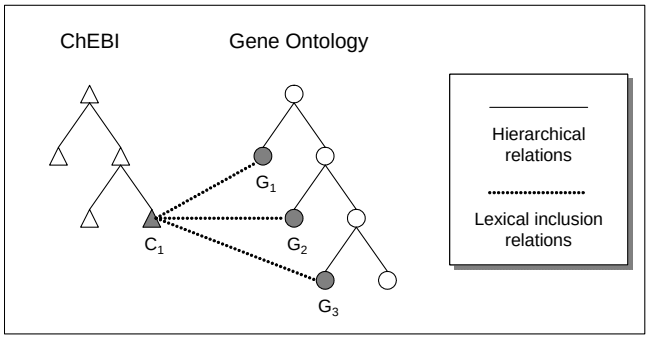


Figure 1 – Lexical inclusion relations between ChEBI terms and GO terms.

2.2 Results

Of the 10,516 entities in ChEBI, 2,700 (26%) were identified in the names of 9,431 GO terms. In other words, 55% of the 17,250 GO terms include in their names the name of some ChEBI entity. These name inclusion relations resulted in 20,497 associations between a ChEBI entity and a GO term.

3 GENERALIZATION TO OTHER BIOLOGICAL ONTOLOGIES

In addition to updating the results of an earlier mapping between GO and ChEBI, this study proposes to apply the method developed for ChEBI to the other members of the OBO family. All terms from the 23 other ontologies (apart from GO and ChEBI) will be mapped to GO and quantitative results will be reported.

These results will contribute to quantifying the relations existing between entities in GO and in the other OBO ontologies. This work can be understood as a first step towards the generalization of Obol to the other OBO ontologies [4]. In addition, as shown in [6], such relations can be used to suggest dependence relations among GO terms.

[3] As we simply removed the trailing *s* from ChEBI names, some inaccurate names were generated (e.g., *phosphoru* and *mustard ga*). Such incomplete names will not mach any lexical items in GO names and, beside slowing down slightly the matching process, this overgeneration has no detrimental consequences on the identification of ChEBI entities in GO names.

REFERENCES

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.
2. Ogren P.V., Cohen K.B., Acquaah-Mensah G.K., Eberlein J,. Hunter L. The compositional structure of Gene Ontology terms. Pac Symp Biocomput 2004, 214-25.
3. Wroe CJ, Stevens R, Goble CA, Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. Pac Symp Biocomput 2003:624-35.
4. Mungall, C. Obol: integrating language and meaning in bio-ontologies. Comparative and Functional Genomics. 2004;5:6-7, 509-520.
5. Smith B, Williams J, Schulze-Kremer S. The ontology of the gene ontology. AMIA Annu Symp Proc. 2003;:609-13.
6. Burgun A, Bodenreider O. An ontology of chemical entities helps identify dependence relations among Gene Ontology terms. First Symposium on Semantic Mining in Biomedicine 2005:(in press).

Reconstructing Formal Temporal Logic Models of Cellular Events using the GO Process Ontology

Naren Ramakrishnan, Marco Antoniotti, and Bud Mishra
Department of Computer Science, Virginia Tech, Blacksburg, VA 24061
Courant Institute of Mathematical Sciences, New York University, NY 10012
NYU School of Medicine, New York, NY 10016

ABSTRACT

We present an approach (GOALIE) to use the GO process ontology to reconstruct formal temporal logic models of cellular systems. The reconstructed models are expressed as Kripke structures and support various query, inference, and reasoning operations. This application highlights how the use of an ontology can help describe complex cellular dynamics in the vernacular of propositional temporal logic.

Introduction

The GO process ontology spans a wide range of biological events, from intra-nuclear processes such as DNA transcription, to organism-wide processes such as aging. The traditional use of such a vocabulary is in functional enrichment analysis of gene sets, as a driver for automated annotation of hypothetical proteins, or for model management in biological databases. Such applications essentially exploit only the taxonomical properties (e.g., membership, set containment) of the ontology but do not otherwise use its process-oriented nature to present dynamical perspectives on biological systems. In this paper, we present an approach (GOALIE; Gene Ontology Algorithmic Logic for Invariant Extraction) that uses the GO process ontology to reconstuct formal temporal logic models of cellular events.

The models reconstructed by GOALIE are formally referred to as *Kripke models* in the model checking literature [2]. For our purposes, a Kripke model is simply a directed graph whose nodes encode possible transcriptional states, edges indicate state transitions, and where the nodes are labeled by propositions that hold true in that state. By choosing these propositions from the set of 8517 possible GO process ontology terms, we ensure that any inferences made (e.g., a temporal invariant) on the resulting Kripke structures are interpretable as biologically relevant patterns and hypotheses. For instance, from Fig. 1 we see that *all* state transitions from a state where *q* is true to a state where *r* is true *must* pass through a state where *p* is true. This shows that cell size serves as an effective checkpoint in the transition into the DNA synthesis phase. The biologist can similarly pose other interesting queries about the satisfaction (or refutability) of temporal logic formulae, in the reconstructed model, under given conditions, obtaining affirmative or impossibility answers. Needless to say, a Kripke model is a powerful mechanism to reason about process happenings in a biological context.

GOALIE

We recover Kripke structures by utilizing the GO process ontology in conjunction with time course microarray datasets. We define the states of the Kripke structures as clusters obtained by partitioning (e.g., by a k-means algorithm) overlapping time windows of the time course dataset. These clusters are then *labeled* with the GO process ontology term using an empirical Bayes approach. Labeled clusters are then "chased" to yield transitions to clusters in neighboring time windows. The basis for relating clusters across time windows is the commonality of labelings as revealed by the previous step. The above stages are then repeated, as necessary, in an iterative fashion to refine the initial clusterings (in response to the identified state transitions) or to adjust the transitions (to reflect new cluster assignments). Since the propositions are taken from a controlled vocabulary, we can combine these propositions to create formulate in a propositional temporal logic (CTL), useful in describing complex cellular dynamics. For more details, see [1].

EXPERIMENTAL RESULTS

Fig. 2 depicts a screenshot of the GOALIE software for use in reconstructing a temporal logic model of cell cycle regulation in *S. cerevisiae* (dataset of [3]). GOALIE allows the user to interatively explore chains of GO labelings across time windows and track the validity of temporal formulae, to see if they change state. The system provides hyperlinks to external websites (e.g., related to definitions of GO categories, public repositories of experimental datasets) as well as visualization and query interfaces. Fig. 1 (right) shows the Kripke structure itself; the correspondence with the idealized diagram of Fig. 1 (left) is readily seen.

GOALIE is now being employed in many different case studies, including studying host-pathogen interactions, the dynamics of cancer progression, and the life cycle of the malaria parasite. We are building fast inference algorithms to answer interesting biological queries over large Kripke structures. Our aim is to develop GOALIE into a general framework for reasoning in any suitable vocabulary, not just of temporal processes as done here, but also other multimodal logics that can encompass richer abstractions of space, control, and variation.

References

[1] M. Antoniotti, N. Ramakrishnan, D. Kumar, M. Spivak, and B. Mishra, Remembrance of Experiments Past: Analyzing Time Course Datasets to Discover Complex Temporal Invariants, *Technical Report TR2005-858*, New York University, Feb 2005.

[2] E.M. Clarke Jr., O. Grumberg, and D.A. Peled, Model Checking, MIT Press, Jan 2000.

[3] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, Comprehensive Identification of Cell Cycle Regulated Genes of the Yeast *Saccharomyces Cerevisiae* by Microarray Hybridization, *Molecular Biology of the Cell*, 9(12), pages 3273-3297, Dec 1998.

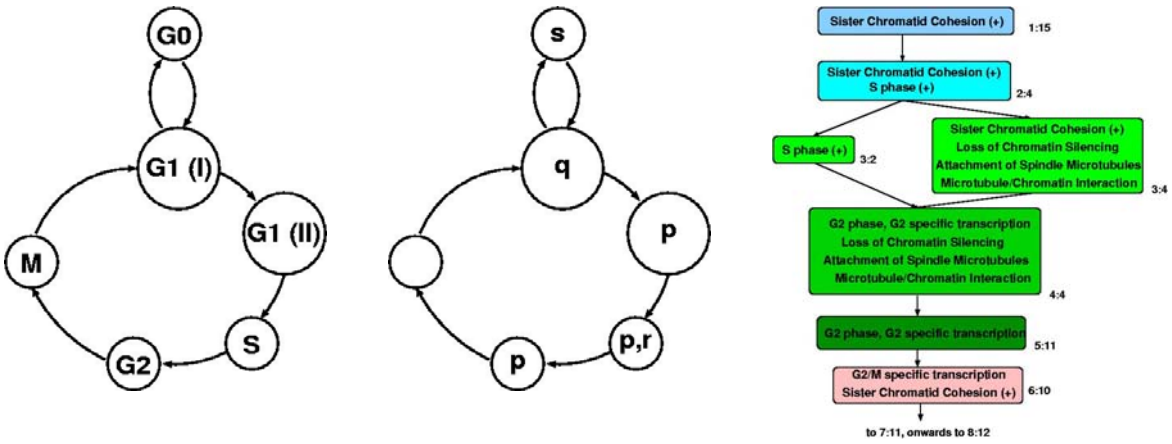


Fig. 1(left) State transition diagram depicting key stages of cell cycle regulation in *S. cerevisiae*. The nodes are labeled with the names of the stages – M(itosis), Gaps, and S(ynthesis). (middle) Kripke diagram of cell cycle regulation obtained manually. States in a Kripke diagram are labeled by the propositions that hold in them. Here, the propositions *p*, *q*, *r* and *s* denote "cell size large enough for division," "cytokinesis takes place," "DNA replication takes place," and "cell is in quiescence." For ease of illustration, not all states are labeled. (right) Kripke diagram of cell cycle regulation, obtained automatically by GOALIE. The nodes are identified by cluster numbers (arbitrary) in given time course windows and labeled by GO process ontology terms.

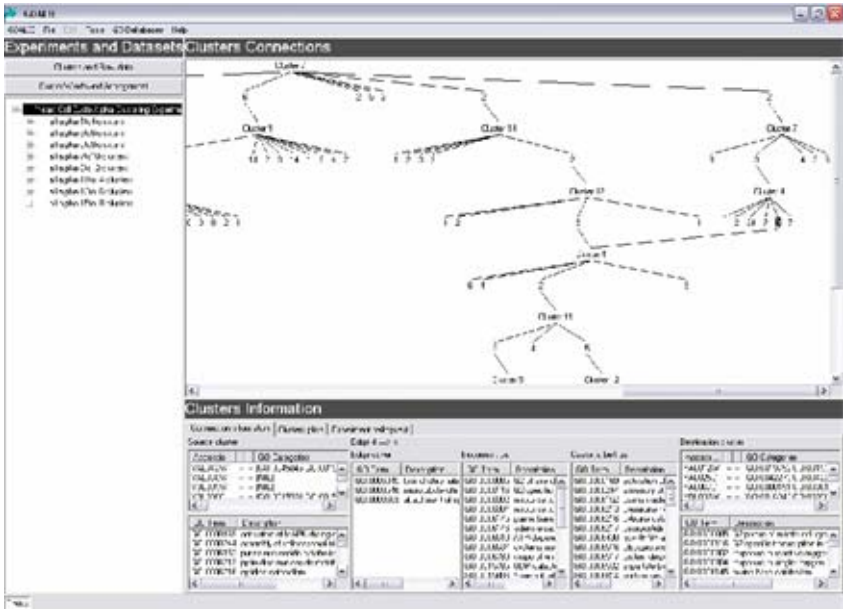


Fig. 2: A screesnshot of GOALIE.The left part shows the various time slices utilized in this study. The top right displays a snapshot of interactive exploration and chasing of clusters. The bottom right part identifies propositions that remain true when going from a source cluster to a destination cluster as well as propositions that become true and those that cease to be true. Notice that cluster 7 in the first time window has been "chased" to yield a chain through successive time windows (clusters 7, 4, 4, 11, and 12 respectively). The links between clusters are labeled with the cardinality of GO terms in common. For instance, the first edge in this chain involves 2 common GO terms, the second involves 3 common GO terms, and so on.

On the proper treatment of pathologies in biomedical ontologies

Barry Smith[1,2*] *and Anand Kumar*[1]

[1]*IFOMIS, University of Saarbrücken, Germany;* [2]*Department of Philosophy, University at Buffalo, USA*

ABSTRACT

Motivation: In previous work on biomedical ontologies we showed how the provision of formal definitions for relations such as *is_a* and *part_of* can support new types of automated reasoning about biomedical phenomena. We here extend this approach to the *transformation_of* characteristic of pathologies.

1 INTRODUCTION

Pathological entities exist in the biological domain at various levels of granularity, from cell components to whole populations. Such entities involve in every case processes – for example, the patho-physiological pathways involved in the etiology of a range of different types of diseases – which cross granular levels. The proper treatment of pathologies within a formal-ontological framework thus demands a facility for simultaneous representation of entities existing at different levels of granularity. We here present the outlines of such a framework, against the background of our previous work on the treatment of relations, time, change and granularity in biomedical ontologies especially within the context of the OBO Relation Ontology [1].

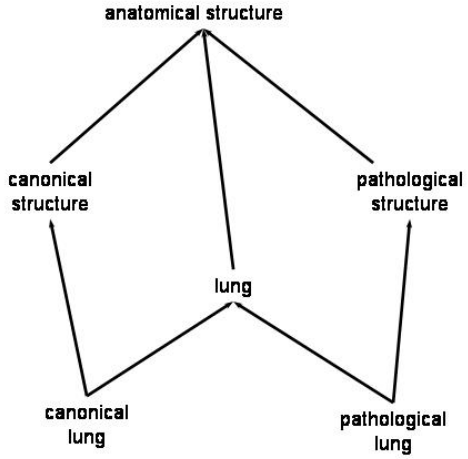
We shall confine our focus here on pathological continuant entities, which include neoplasms, blisters, punctures, fractures, portions of pus, portions of amyloid but also the bearers of such entities (a wounded knee, a carcinomatous colon, a fractured tibia, and so on).

To say that such entities are *continuants* is to say that they endure as self-identical through time even while undergoing a variety of different sorts of changes. This is in contrast to occurrents (events, processes, happenings), which unfold themselves through time in successive temporal phases or stages which share no parts in common.

Some pathological continuants are subclasses of the class *anatomical structure*, a designation we take over from the Foundational Model of Anatomy (FMA) [2]. Such pathological structures are classified as pathological relative to others classified as 'normal' or 'canonical'. We here leave open the question of what it is in virtue of which a given instance is to be classified as normal. We shall assume that both are subclasses of the same parent *anatomical structure*, whose instances include both pathological and normal structures, along the lines indicated in Figure 1.

* Corresponding author

Figure 1. A simple classification-schema for pathological structures



We then have, for example:

pathological colonic mucosal cell is_a pathological structure

canonical colonic mucosal cell is_a canonical structure

colon with pathological features is_a pathological structure

normal colonic mucosal cell is_a colonic mucosal cell

pathological colonic mucosal cell part_of colon

Here *is_a* and *part_of* are defined as follows, using variables We use *A*, *B*, *C* ... to range over classes (universals, types) of pathological continuants, *c*, *c′*, ... to range over the instances of such classes, and *t*, *t′*, ... to range of instants of time:

A is_a B =def. for all *c*, *t*, if *c* **instance_of** *A* **at** *t* then *c* **instance_of** *B* **at** *t*.

A part_of B =def. for all *c*, *t*, if *c* **instance_of** *A* **at** *t* then there is some *c′* such that: *c′* **instance_of** *A* **at** *t* and *c* **part_of** *c′* **at** *t*,

where '**part_of**' is the instance-level part relation (between, for example, this particular cell and this particular lung). The reference to *times* in these definitions is designed to do justice to the fact that one and the same entity can instantiate different classes and gain and lose parts in the course of time. Note also the all-some structure of the definition of *part_of*, which is characteristic of almost all relations between classes of the sort treated by biomedical ontologies.

Some (but not all) kinds of pathological structures are such that their instances are in every case transformations of ca-

nonical structures of entities of a given kind existing earlier. We define *transformation_of* as a relation between continuant classes:

A transformation_of B =def. for all *t* and all *c*, if *c* is an instance of *A* at *t*, then there is an earlier time *t′* at which *c* is an instance of *B*, and for no *t*, *c* is *c* an instance of both *A* and *B* at *t*.

The pathological colon mucosal cell can be a transformation either of the canonical colon mucosal cell or of its precursor, depending on whether the pathology is hereditary or acquired. Because *transformation_of* is transitive, however, we can assert quite generally:

pathological colon mucosal cell transformation_of canonical colon mucosal cell precursor

given that in every case:

colon mucosal cell transformation_of colon mucosal cell precursor

The temporal relationships between canonical entities and entities with pathological features have not been sufficiently addressed in ontologies thus far, and even developmental ontologies utilizing the methodology of stages have preferred not to incorporate a formal machinery for dealing explicitly with times [3]. Thus most of the instances of colonic mucosal cells with pathological features are a transformation of instances of normal colonic mucosal cell. In some cases, the former is a transformation of a precursor entity of the latter. Transtemporal relations of this sort are not recorded in the National Cancer Institute Thesaurus, where for example no relations are asserted between the two classes *abnormal cell* and *normal cell*, not even that they have a common parent: *cell* [4]. Transformation relations are also absent in the SNOMED CT terminology [5]. A relation which we do find in SNOMED CT, however, is that of location, for example in:

lung cyst finding_site lung structure

Better, however, would be to eliminate the epistemological connotations of '*finding_site*' by using a location relation such as GALEN's *locus* [6] or OBO's *located_in* [1]:

A located_in B =def. for all *c*, *t*, if *c* **instance of** *A* **at** *t* then there is some c_1 such that: C_1c_1t and *c* **located_in** c_1 **at** *t*.

PathBase [7] provides a subsumption hierarchy for various pathological processes. It has

endoplastic reticulum defect is-a subcellular defect

This relation can thus be used with the colon cell assertions above to generate for example:

pathological colon mucosalcell with endoplastic reticulum defect is-a pathological colon mucosal cell with subcellular defect

Further implications which can be drawn are:

endoplastic reticulum defect located-in endoplastic reticulum

fpathological colon mucosalcell has_level_of_granularity cell.

where levels of granularity can be inferred from an *is_a* hierarchy such as that of the FMA, e.g. from the fact that the colon is an organ we can infer:

colon has_level_of_granularity organ

and thus also that both *canonical and pathological colon have this same level of granularity.*

2 CANCER STAGING

We can use the framework to capture some of the information contained in systems for cancer staging such as the TNM (for: Tumour, Node, Metastasis) system, which is used to classify pathological states into specific categories important for carcinoma management. The T2 stage, for example, is defined as: *carcinoma has invaded the muscularis mucosa of the colon wall* and the T1 stage as: *carcinoma has invaded the mucosa*. N1 designates a stage with one to four lymph nodes, M1 a stage where a metastasis is present in a non-contiguous part of the body. We can then assert that a pathological entity of the type *carcinoma in colon* at stage T2N1M1 must be a transformation of either a T1N1M1 or a T2N0M1 structure. If a carcinoma is a transformation from T1N1M1 to T2N1M1 then there has occurred a process of the type *muscularis mucosa invasion*. If there is a transformation from T2N0M1 to T2N1M1 then this implies that the last process to take place was one of lymph node metastasis. If there is a transformation from T2N1M0 to T2N1M1, then this implies that the last process to take place was one of metastasis to a non-contiguous body region.

3 CONCLUSION

We have sketched a formal approach to class-class relations in the realm of pathologies that is designed to support new types of cross-granular reasoning and also reasoning about entities which exist at different points in time, for example in the domain of cancer staging.

REFERENCES

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

[2] Rosse C, Mejino JLV Jr: A reference ontology for bioinformatics: The Foundational Model of Anatomy. *J Biomed Informatics*, 2003; 36:478-500.

[3] Aitken S: Formalising concepts of species, sex and developmental stage in anatomical ontologies, *Bioinformatics*, in press.

[4]Ceusters W, Smith B, Goldberg L: A terminological and ontological analysis of the NCI Thesaurus, *Meth Inform Medicine*, in press.

[5] http://nciterms.nci.nih.gov/

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

[7] http://eulep.anat.cam.ac.uk/Pathology_Ontology/MPATH-dynamic.php

Using Ontology Reasoning to Classify Protein Phosphatases

*Katherine Wolstencroft, Phillip Lord, Lydia Tabernero, Andy Brass and Robert Stevens**

Faculty of Life Sciences and School of Computer Science, University of Manchester, Oxford Road, UK

ABSTRACT

The need for automation of protein classification is motivated by the growing number of genome sequencing projects and the resulting stock-pile of data requiring annotation. Classification plays a central role in the annotation process and is the first step in understanding the molecular biology of an organism. However, classification and annotation are now rate-limiting steps.

We present a method for the automated classification of a protein family from the protein complement of a genome using ontological reasoning.

Introduction

Classification of proteins by human experts is regarded as the gold-standard in biological data annotation. Human expertise is able to recognise the properties that are necessary and sufficient to place an individual gene product into a specific class. These differences are often subtle and automated annotation often fails to achieve the same classification at a fine-grained, subfamily level.

Many proteins are assemblies of domains. Each domain might have a separate function within the protein, such as binding or catalysis, but it is the composition of the different domains that gives each protein its specific function. There are many tools dedicated to discovering functional domains, for example, SMART (Letunic *et al*, 2004) and InterproScan (Mulder et al, 2005) but whilst they can report the presence of functional domains, bioinformaticians are required to perform the analysis that places a protein with a particular set of domains into a particular protein family or subfamily. To reach human expert standards, automated methods must also perform this analysis step. The ontology system presented here does just that. By capturing the necessary and sufficient conditions for membership of each protein family or subfamily, in an OWL ontology, we formalise the rules for class membership. This enables the use of ontology reasoners to perform the human analysis step of comparing individual proteins to the defined protein family classes and assign them to a place in the classification.

In this study, we present the results of analysing the protein phosphatase complement of the human and *Aspergillus fumigatus* genomes. Phosphatases were a suitable case-study because they are a large protein family involved in almost all cellular processes, making classification at a detailed level vital for understanding the specificity of individual proteins and for comparative genomic studies. Several phosphatase proteins have also been implicated in diseases, such as, diabetes, cancer and neurodegenerative conditions (Schonthal, 2001, Zhang, 2001 and Tian & Wang, 2002), making them important targets for medical and pharmaceutical research.

Methods

The automated classification system we present combines description logics (DL) reasoning (Baader et al, 2003) with service oriented architecture (oinn et al, 2004) to extract and classify the protein phosphatase complement of an organism. The foundation step was to produce an ontology in OWL (Web Ontology Language), describing the domain architecture of each protein phosphatase subfamily, derived from peer-reviewed literature by protein phosphatase experts. These class descriptions were then used to compare with the domain architecture of individual proteins using the Instance Store (IS) (Horrocks et al 2004). The IS combines a description logic reasoner with a relational database and allows reasoning over large numbers of individuals.

The domain architecture of individual proteins was determined by performing InterproScans of the raw sequence data and translating the results into abstract OWL format.

The combination of ontology, Instance Store, bioinformatics domain analysis and ontological reasoning provided the technology to facilitate the automated extraction and classification of any number of proteins from raw sequence data. Figure 1 shows the architecture of the ontology system.

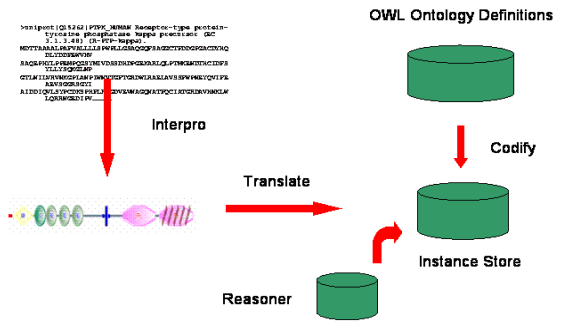


Figure 1: The ontology System Architecture

Results

* To whom correspondence should be addressed.

Human protein phosphatases have been isolated and characterised in previous studies (Alonso *et al*, 2004, cohen, 1997, kennelly, 2001), so comparing the results of the classification from the ontology system with the classification produced by domain experts provides a way of measuring the success of the ontology system. The results showed that all of the 118 phosphatase proteins identified and classified in previous studies were classified in the same place in the protein family hierarchy using the ontology method.

An interesting result from the analysis was that, using the ontology, we were able to identify additional functional domains in two dual specificity phosphatases that present the opportunity to refine the classification of the subfamily into further subtypes.

The results from the *A.fumigatus* investigation revealed large differences between the protein phosphatases the two species. Not only were there far fewer phosphatases in *A.fumigatus*, for example, 1 myotubularin and 1 MAP kinase phosphatase, as opposed to 16 and 11 respectively in human, but there were whole subfamilies not represented. Some of these missing subfamilies may reveal differences in phosphorylation pathways and are targets for further investigation.

The *A.fumigatus* results also identified a novel type of calcineurin phosphatase with an extra homeobox domain. Further investigation showed that this extra domain was present in closely related pathogenic fungi, but we were unable to identify it in any other species, making it as a potentially interesting drug target for pharmaceutical investigation

Discussion

The scale of data production in post-genomic bioinformatics presents new problems for the bioinformatician. The pace at which new data is produced is outstripping the pace at which it can be analysed and annotated. Often, compromises on the quality of annotation have to be made in order to interpret large data sets quickly, providing annotation at a more generic level, which results in the loss of information. The method we present here addresses part of this problem by encoding human expert knowledge as an ontology. The differences between protein classes can be captured at a detailed level to discriminate between closely related protein subfamilies.

The human phosphatase study demonstrated that this system equaled the performance of manual human expert classification. It was also discovered that the ontology system was efficient at uncovering novel, unexpected functional domains, revealing anomalies that did not fit the community knowledge.

The use of ontological technology in the bio-ontologies domain has been largely restricted to enhancing browsing and querying over existing data, or to statistical investigation. In this paper, we have described the application of ontological reasoning to enhance community-developed knowledge.

By encoding pre-existing community knowledge in this form we have gained the advantage of automation and additionally, the ability to systematically analyse large volumes of biological data.

ACKNOWLEDGEMENTS

This work was funded by an MRC PhD studentship and myGrid e-science project, University of Manchester with the UK e-science programme EPSRC grant GR/R67743. Preliminary sequence data was obtained from The Institute for Genomic Research website at http://www.tigr.org from Dr Jane Mabey-Gilsenan, University of Manchester. Sequencing of *Aspergillus fumigatus* was funded by the National Institute of Allergy and Infectious Disease U01 AI 48830 to David Denning and William Nierman, the Wellcome Trust, and Fondo de Investicagiones Sanitarias

REFERENCES

Alonso A, Sasin J, Bottini N, Friedberg I, Friedberg I, Osterman A, Godzik A, Hunter T, Dixon J, Mustelin T. (2004) Protein tyrosine phosphatases in the human genome *Cell.* **117(6)**:699-711

Baader F, Calvanese D, McGuinness D, Nardi D, Pater-Schneider p (2003) The Description Logic Handbook: Theory, Implementation and Applications, Cambridge University Press

Cohen PT (1997) Novel protein serine/threonine phosphatases: variety is the spice of life. *Trends Biochem Sci.* **22(7)**:245-51. Review.

Horrocks, L. Li, D. Turi, and S. Bechhofer (2004) The Instance Store: DL reasoning with large numbers of individuals. In Proc. of the 2004 Description Logic Workshop, pages **31-40**, 2004

Kennelly PJ (2001) Protein phosphatases--a phylogenetic perspective. *Chem Rev*. **101(8)**:2291-312. Review.

Letunic et al. (2004) SMART 4.0: towards genomic data integration Nucleic Acids Res 32

Mabey JE, Anderson MJ, Giles PF, Miller CJ, Attwood TK, Paton NW, Bornberg-Bauer E, Robson GD, Oliver SG, Denning DW (2004) CADRE: the Central Aspergillus Data Repository *Nucleic Acids Res*. **1;32**:D401-5

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. (2005). InterPro, progress and status in 2005. *Nucleic Acids Res.* **33,** Database Issue:D201-5

Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **22**;20(17):3045-54

Schonthal AH. (2001) Role of serine/threonine protein phosphatase 2A in cancer. *Cancer Lett.* **170**(1):1-13

Tian Q, Wang J. (2002), Role of serine/threonine protein phosphatase in Alzheimer's disease. *Neurosignals***. 11(5)**:262-269

Zhang ZY (2001) Protein tyrosine phosphatases: prospects for therapeutics. *Curr Opin Chem Biol.* **5(4):**416-23

INDUS: an ontology-based information integration system

Doina Caragea', Jie Bao, Jyotishman Pathak and Vasant Honavar

Artificial Intelligence Research Laboratory, Department of Computer Science, Iowa State University, Ames, IA 50010, USA

ABSTRACT

INDUS (Intelligent Data Understanding System) is a *federated*, *query-centric* system for information integration and knowledge acquisition from distributed semantically heterogeneous data sources. INDUS employs ontologies (controlled vocabularies of domain specific terms, and relationships among terms) and inter-ontology mappings, to enable a user to view a collection of such data sources (regardless of location, internal structure and query interfaces) as though they were a collection of tables structured according to a user-supplied ontology.

1 INTRODUCTION

Ongoing transformation of biology from a data-poor science into an increasingly data-rich science has resulted in a large number of autonomous data sources (e.g., repositories of protein sequences, structures, expression patterns, interactions). This has led to unprecedented, and as yet, largely unrealized opportunities for large-scale collaborative discovery in a number of areas: characterization of macromolecular sequence-structure-function relationships, discovery of complex genetic regulatory networks, among others.

At present, there are hundreds of databases of interest to molecular biologists alone [Discala et al., 2000]. Because the data repositories are typically autonomous, and often focused on specific subfields of biology, ontological (and hence semantic) differences among them are simply unavoidable. However, in exploring specific scientific questions of interest, scientists often need to be able to retrieve and analyze data from multiple sources. Effective use of such data in a given context requires reconciliation of semantic differences among the relevant data sources from a user's point of view. Hence, there is an urgent need for tools to support rapid and flexible assembly and analysis of data from semantically heterogeneous data sources [Jagadish and Olken, 2003].

2 APPROACH

INDUS is a *federated*, *query-centric* system for data integration and knowledge acquisition from distributed, semantically heterogeneous data (See **Fig. 1**). INDUS makes explicit data source specific information, such as the data source schema and (the typically implicit) data source ontologies. The resulting ontology-extended data sources [Caragea et al., 2004] enable users to specify semantic correspondences between the user ontology and the data source ontologies by specifying inter-ontology mappings.

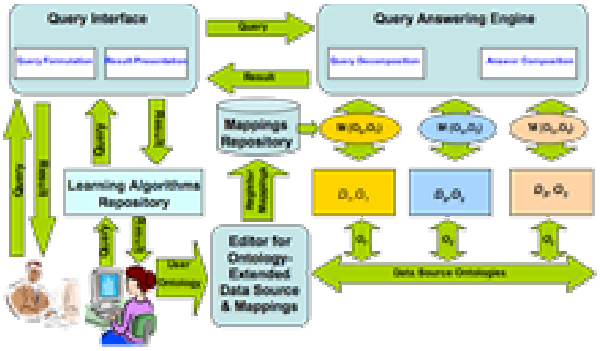


Fig. 1. INDUS: a system for data integration and knowledge acquisition from semantically heterogeneous distributed data.

This enables each user to view a collection of autonomous, semantically heterogeneous, distributed data as though they were a collection of inter-related tables structured according to an individual user's ontology. Thus, users can interact with and explore data sources of interest to them from multiple points of view simply by changing their perspective (i.e., user ontology and semantic correspondences between the user ontology and the data source ontologies). Queries posed using terms in the user ontology are transformed, using a sound query rewriting algorithm, into queries that can be answered by the individual data sources. The results are expressed in terms of the user's ontology [Caragea et al., 2004] (See **Fig. 1**).

3 INDUS PROTOTYPE

We have completed the implementation of a working prototype of the INDUS system to enable biologists with some familiarity with the relevant data sources to rapidly and flexibly assemble data sets from multiple data sources and to query these data sets. This can be done by specifying a user ontology, simple semantic mappings between data source specific ontologies and the user ontology and queries – all without having to write any code. An initial version of the INDUS software and documentation are available at **www.cild.iastate.edu/GM066387_homepage.htm**.

* To whom correspondence should be addressed.

The current implementation includes support for:

- Import and reuse of selected fragments of existing ontologies (e.g., Gene Ontology GO), editing of ontologies, specification of semantic relationships between ontologies using inter-ontology mappings [Bao and Honavar, 2004].
- Specification of semantic correspondences between a user ontology O_U and data source ontologies O_D [Caragea et al., 2004]. Semantic correspondences between ontologies can be defined at two levels: schema level (between attributes that define data source schemas) and attribute level (between values of attributes). INDUS allows the following types of semantic correspondences at both schema and attribute level: semantic equality (e.g., *AASequence:*O_D ≡ *ProteinSequence:*O_U), semantic subsumption (e.g., *MIPS:16.19.01:*O_D ≤ *GO: 0017076:*O_U) and procedural mappings (e.g., from *AASequence:*O_D to *AAComposition:*O_U). Consistency of semantic correspondences are verified by an efficient algorithm for reasoning about subsumption and equivalence relationships.
- Registration of a new data source using a data-source editor for defining the schema of the data source (the names of the attributes and their corresponding ontological types), location, type of the data source and access procedures that can be used to interact with a data source. In the current implementation several types of data sources can be defined including multiple relational databases (Oracle, MySQL, PostgreSQL), and files (e.g., ARFF files used in WEKA, a widely used open source machine learning software package).
- Specification and execution of queries across multiple large, semantically heterogeneous data sources with different interfaces, functionalities and access restrictions. Each user may choose relevant data sources from a list of data sources that have been previously registered with the system and specify a user ontology (by selecting an ontology from a list of available ontologies or by invoking the ontology editor and defining a new ontology).

Once the ontology-extended data sources and the user ontology have been specified, the user can select mappings between data source ontologies and user ontology from the available set of existent mappings (or invoke the mappings editor to define a new set of mappings). Once the necessary mappings are specified, the system can answer queries posed by the user. The data needed for answering a query is specified by selecting (and possibly restricting) attributes from the user ontology, through a user-friendly interface. Queries posed by the user are sent to a query-answering engine (QAE) that decomposes a user query into subqueries that can be answered by the individual data sources (using predefined or user-supplied mappings between the respective ontologies). The answer to the user query (expressed in terms of user ontology) is constructed and presented to the user by the QAE using results of queries to the distributed data sources. INDUS has been used to assemble several data sets used in the exploration of protein sequence-structure-function relationships [Caragea et al., 2005]. Examples of such data sets include: a data set used for building a classifier for automating functional annotation of protein sequences based on sequence composition [Andorf et al., 2004] and structural features of proteins and a comprehensive database of protein-protein interfaces **www.cild.iastate.edu/GM066387_homepage.htm**.

4 CONCLUSIONS

We have presented INDUS, a federated, query-centric approach to answering user queries from distributed, semantically heterogeneous data sources. INDUS assumes a clear separation between data and the semantics of the data (ontologies) and allows users to specify ontologies and mappings between data source ontologies and user ontology. INDUS enables users (or application programs e.g., learning algorithms) to retrieve results of queries from semantically heterogeneous data sources.

ACKNOWLEDGEMENTS

This work was funded in part by grants from the National Science Foundation (IIS 0219699) and the National Institutes of Health (GM 066387).

REFERENCES

Andorf, C., Silvescu, A., Dobbs, D. and Honavar, V. (2004). Learning Classifiers for Assigning Protein Sequences to Gene Ontology Functional Families. In: *Fifth International Conference on Knowledge Based Computer Systems (KBCS 2004)*, India.

Bao, J. and Honavar, V. (2004). Collaborative ontology building with wiki@nt - a multi-agent based ontology building environment. In: *Proc. of 3rd International Workshop on Evaluation of Ontology based Tools*, ISWC 2004, Japan.

Caragea, D., Pathak, J., and Honavar, V. (2004). Learning Classifiers from Semantically Heterogeneous Data. In: *Proceedings of the Third International Conference on Ontologies, DataBases and Applications of Semantics for Large Scale Information Systems* (ODBASE'04), October 25-29, 2004, Agia Napa, Cyprus.

Caragea, D., Silvescu, A., Pathak, J., Bao, J., Andorf., C., Dobbs, D. and Honavar, V. (2005). *Information Integration and Knowledge Acquisition from Semantically Heterogeneous Biological Data Sources*. In: Proc. of the 2nd Int. Workshop on Data Integration in Life Sciences (DILS'05), San Diego, CA.

Discala, C., Benigni, X. Barillot, E. and Vaysseix, G. (2000). DBcat: a catalog of 500 biological databases. Nucleic Acids Res. 2000 Jan 1;28(1):8-9.

Jagadish, H.V. and Olken, F. (2003). Data Management for the Biosciences. Report of the NSF/NLM Workshop of Data Management for Molecular and Cell Biology, Feb. 2-3, 2003.

News & Views: The BioPAX Pathway Data Exchange Format

BioPAX Workgroup

Presenters: Michael P. Cary[1] and Joanne S. Luciano[2]

[1] - Memorial Sloan-Kettering Cancer Center, New York, NY
[2] - Harvard Medical School, Boston, MA

ABSTRACT

Motivation: Gathering existing knowledge is the first step in modeling or analyzing a biological process, thus pathway data integration is vital for many applications of computational biology[1]. As the number of pathway databases increases, pathway data integration becomes more difficult. At the start of 2005, there were over 170 databases containing pathway information, widely varying in form and content (http://www.cbio.mskcc.org/prl). A standard exchange format for pathway data, supported by major pathway databases, will significantly reduce the amount of time and energy spent by computational biologists on data integration and lead to increased pathway data sharing.

1 INTRODUCTION

BioPAX (http://www.biopax.org) is a community-based effort to develop a biological pathway data exchange format. BioPAX Level 1, which focuses on metabolic pathway information, was released July 2004. BioPAX Level 2, which will add support for molecular interactions via inclusion of the PSI-MI data model, will be finalized mid-2005. Level 3, currently under development, will add support for molecular states and genetic regulatory networks. Future levels will be able to represent genetic interactions and generic molecules and processes.

BioPAX is being developed in a practical, leveled approach in which each level supports a greater variety of pathway data. BioPAX Level 1, implemented in OWL, is supported by BioCyc[2], WIT[3], KEGG[4] and others. BioPAX Level 2 is expected to be supported by aMAZE[5], Reactome[6] and others.

The BioPAX group is coordinating with other pathway related standards initiatives, such as SBML[7], CellML[8], and PSI-MI[9], to minimize duplication of work and to ensure compatibility with these standards in areas of overlapping coverage. Participation in BioPAX is voluntary and without fee. The BioPAX format and any associated software developed by the BioPAX group are open source and freely available to all under the GNU LGPL license (http://www.gnu.org/copyleft/lesser.html).

Address correspondence to: biopax-info@biopax.org

2 ONTOLOGY DETAILS

Four basic classes are defined in the BioPAX ontology to represent pathway information: the root level **entity** class and its three subclasses: **pathway, interaction** and **physicalEntity**.

Entity: Any concept referred to as **a discrete biological unit** when describing pathways.

Pathway: A **set of interactions**. A pathway is a collection of molecular interactions and reactions, often forming a network, which biologists have found useful to group together for organizational, historic, biophysical or other reasons.

Interaction: An entity that defines a single biochemical **relationship between two or more entities**.

PhysicalEntity: An entity that has a physical structure. This class serves as the super-class for all physical entities, although its current set of subclasses is limited to **molecules**. Physical entities are frequent building blocks of interactions.

BioPAX Level 1 defines **seven** main types of interaction and **four** types of physical entity. Interactions: conversion (and three subtypes: complex assembly, transport, biochemical reaction) and control (and two subtypes: catalysis and modulation); Physical entities: complex, protein, RNA, small molecule (DNA is available in BioPAX Level 2).

Pathway information in BioPAX is represented by creating instances of these classes. For example, defining a typical enzyme-catalyzed biochemical reaction requires physical entity instances for the substrates, products, and enzyme, a biochemical reaction instance to describe the conversion of the substrates to the products, and a catalysis instance to define the relationship between the enzyme and the reaction. For

more detail, see the full documentation at www.biopax.org.

3 EXAMPLE USE CASES

Pathway Data Warehouse BioPAX could make creation of pathway data warehouses easier if many databases provide access to their data in the BioPAX format. Similar to sequence data warehouses, these could be locally maintained to provide fast access to publicly available pathway data.

Pathway Analysis Software Example: Molecular profiling analysis Molecular profiling experiments, using such technologies as gene expression microarrays and mass spectrometers, are often compared across two or more conditions (e.g. normal tissue and cancerous tissue). The result of this comparison is often a large list of genes that are differentially present in the tissue of interest. It is interesting and useful to analyze these lists of genes in the context of pathways. For instance, one could look for pathways that are statistically over-represented in the list of differentially expressed genes. The result is a list of pathways that are active or inactive in the condition of interest compared to a control. The list of pathways is often much shorter than the list of input genes and easier to comprehend. BioPAX could facilitate this and other kinds of pathway-based analyses by giving tools easy access to a large body of pathway data in a common format.

Visualizing Pathway Diagrams Pathway diagrams are useful for examining pathway data. A number of formats are available for these images, but only a few available viewing tools link components in the image to underlying data. A mapping of BioPAX to a symbol library for pathway diagrams (such as Kohn maps - http://discover.nci.nih.gov/kohnk/symbols.html) could be the basis for a general pathway diagram generation tool.

Pathway Modeling Mathematical modeling to understand the dynamics of a pathway system is a frequent use of pathway information. Many of the tools available for pathway modeling support the SBML (http://sbml.org) and CellML (http://www.cellml.org) standards, which describe models in sufficient detail to allow model sharing between tools. While BioPAX is not designed to represent pathway models in as much detail as SBML and CellML, it contains a number of biological concepts not present in these standards. Use of BioPAX to annotate SBML and CellML models could allow linking models to pathway databases and the functional annotations contained therein.

ACKNOWLEDGEMENTS

The BioPAX project is a community effort involving individuals from a number of institutions. We would especially like to thank the following people, who have been instrumental to its development: Gary D. Bader, Emek Demir, Ken Fukuda, Peter Karp, Christian Lemer, Natalia Maltsev, Eric Neumann, Suzanne Paley, John Pick, Jonathan Rees, Andrey Rzhetsky, Chris Sander, Imran Shah, Andrea Splendiani, Mustafa Syed, and Jeremy Zucker.

REFERENCES

1. Cary, M.P., G.D. Bader, and C. Sander, *Pathway information for systems biology.* FEBS Lett, 2005. **579**(8): p. 1815-20.
2. Keseler, I.M., et al., *EcoCyc: a comprehensive database resource for Escherichia coli.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D334-7.
3. Overbeek, R., et al., *WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction.* Nucleic Acids Res, 2000. **28**(1): p. 123-5.
4. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.
5. Lemer, C., et al., *The aMAZE LightBench: a web interface to a relational database of cellular processes.* Nucleic Acids Res, 2004. **32 Database issue**: p. D443-8.
6. Joshi-Tope, G., et al., *Reactome: a knowledgebase of biological pathways.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D428-32.
7. Hucka, M., et al., *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.* Bioinformatics, 2003. **19**(4): p. 524-31.
8. Lloyd, C.M., M.D. Halstead, and P.F. Nielsen, *CellML: its future, present and past.* Prog Biophys Mol Biol, 2004. **85**(2-3): p. 433-50.
9. Hermjakob, H., et al., *The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data.* Nat Biotechnol, 2004. **22**(2): p. 177-83.

Co-Citation of Genes in Ontology Groups

Yizong Cheng[1], Jun Xu[2], Steven Zhao[2], and Andy Fulmer[2]*

[1]Department of ECECS, University of Cincinnati, Cincinnati OH 45221-0030, and [2]Corporate Biotechnology, MVIC, Proctor & Gamble, Cincinnati, OH 45239-8707, USA

[1]yizong.cheng@uc.edu

ABSTRACT

Motivation: Genes residing at different branches in the Gene Ontology (GO) DAG are often used as benchmark standards in validating data mining algorithms. It is desirable to reverse this validation procedure and investigate what kind of data features may characterize these GO terms in various data collections.

Method: Genes sharing the same biological process, according to the Gene Ontology consortium, are mapped to a co-citation gene networks. Connectivity and density of the resulting subgraphs are studied. Furthermore, interconnectivity between these subgraphs is scored and some biological processes with high interconnectivity are listed.

Results: GO groups may not always map into dense areas in the co-citation network. Interconnectivity between the mapping images of GO groups indicates their relationships.

1 INTRODUCTION

When new methodologies and algorithms are proposed to generate gene lists from data (for example, "signatures" from supervised learning or "modules" from unsupervised learning), the concentration of biological processes attributed to the genes is often a standard way to show the effectiveness of the methodologies and algorithms and the significance of the findings. Sometimes one may wonder why a gene list generated from one method, say, a particular graph clustering algorithm, fits more to a biological process than another list. To answer this question, an inverse investigation is desirable in discovering the natural structure of genes involved in a biological process, in terms of specific data. Study exists on global comparison of different gene networks, including those based on biological functions. There is the need to investigate further, for more detailed and localized analysis. Here we report some results from our investigation of the similarity between the grouping of genes based on biological processes, defined by the Gene Ontology consortium, and gene network defined by Medline co-citation.

* To whom correspondence should be addressed.

We have devised graph clustering algorithms to find "dense spots" in a network of genes and showed that many of these clusters contain genes of particular biological processes. Similar results on different gene networks have been reported elsewhere. In this investigation, we tried to answer three questions. First, how likely two genes within the same biological process are adjacent or closely related in a certain gene network. Secondly, to what varying degrees gene groups at specific levels of the gene ontology DAG (directed acyclic graph) are densely connected in those networks. Thirdly, what biological processes may be mapped to subgraphs with high interconnectivity.

2 GENE ONTOLOGY TERMS MAPPED TO THE CO-CITATION NETWORK

We collected mammalian genes indicated at the three lowest levels of the Gene Ontology DAG, as GO groups. These include 425 groups at the sink or leaf level of the DAG, 642 groups at the level above the sinks, and 207 at two levels above the sinks, all within the size range between 3 and 700.

We mapped these groups to a gene network of 12,727 genes and 106,142 edges connecting genes co-cited significantly in MedLine abstracts. We found that 29,024 (27.3%) of these edges were connecting genes within some GO group in our collection. On the other hand, there were 1,376,012 pairs of genes belonging to the same GO group, and thus the chances for genes within the same GO group to be co-cited were very low (2.1%).

When we allowed genes in the co-citation network to be considered in the mapping when they were connected with a path of length 2, the latter percentage increased to 32.5% (with 447,341 intra-GO-group pairs connected in the co-citation network directly or by paths of length 2). This percentage increased to 80.8% (1,112,029 pairs connected) when the maximum path length was 3.

Table 1 summarizes the proportions of intra-GO-group pairs that are also connected in the co-citation network. It is interesting to notice that GO groups at the sink level (level 0) gave significantly less improvement in the percentage for longer path lengths, compared with those at higher levels. This may have something to do with the relatively smaller sizes of the level-0 GO groups. (The average GO group

sizes for level 0, level 1, and level 2 are 9.3, 25.1, and 15.9, respectively.

Table 1. Proportions of intra-GO-group pairs that are also connected in the co-citation network. Group level indicates how close the GO term is to the sinks or leaves of the DAG. Path length indicates how apart a pair are in the co-citation network.

Group level	Path length = 1	Path length <= 2	Path length <= 3
0	2.5%	24.0%	67.5%
1	2.0%	32.3%	81.0%
2	2.7%	38.4%	85.9%
0-2	2.1%	32.5%	80.8%

A GO group, as a gene list, is mapped to the co-citation network as a subgraph. The percentages listed in Table 1 are also the average density of these subgraphs within the co-citation network. Here, the density of a subgraph is the proportion of the possible edges between its vertices that are actually present in the graph (network). Figure 1 below shows the distribution of density vs. GO group size for three different path lengths.

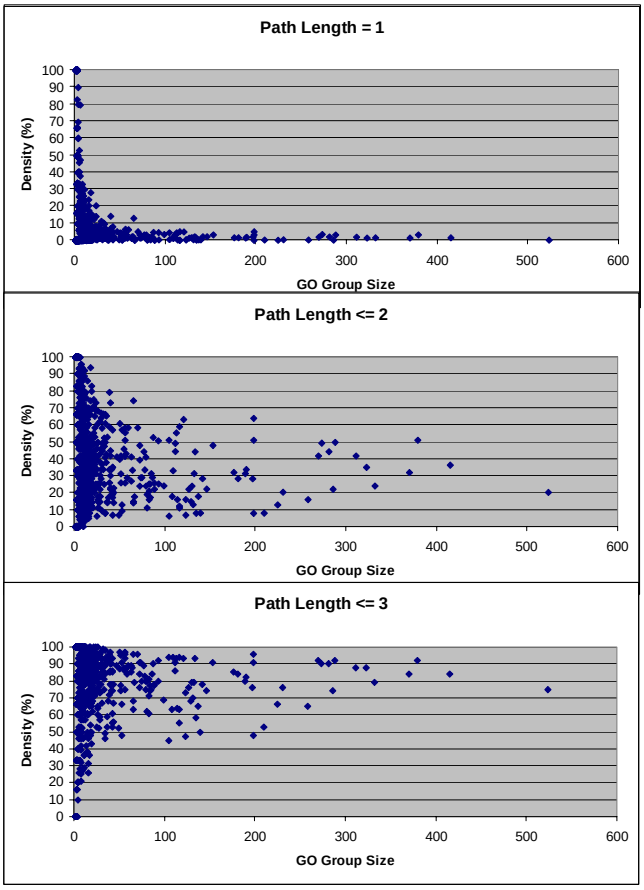


Figure 1. Density distribution of GO groups mapped to the co-citation network with three different ranges of path lengths to be considered as adjacent.

3 CO-CITATION ACROSS GO GROUPS

After mapping GO groups onto the co-citation network, they form (possibly overlapping subgraphs). Given the sizes of two GO groups, *S* and *T*, the size of the intersection between them, *U*, and the number of edges in the co-citation network between non-overlapping members from different groups, *W*, we define an *interconnectivity* score for the pair of groups as $W/((S-U)(T-U))$. This score is the proportion of the possible inter-group pairs that are adjacent in the co-citation network. Table 2 lists come of the GO terms with some of the highest interconnectivities.

Table 2. Some GO terms with the highest co-citation interconnectivity.

Group ID	Biological process	Group ID	Biological process	Inter-connectivity
6882	negative regulation of follicle-stimulating hormone secretion	50808	synapse organization and biogenesis	61.2%
6684	sphingomyelin metabolism	9798	axis specification	57.7%
7129	synapsis	45162	clustering of voltage-gated sodium channels	57.7%
6118	electron transport	15893	drug transport	36.3%
6512	ubiquitin cycle	46661	male sex differentiation	35.9%

4 CONCLUSION

From the study we can see that even though genes involved in the same biological process are expected to have been co-cited in literature, sometimes it is not the case. The disparity shown by different GO groups in density distribution when mapped to other gene networks encourages further study on their structural characteristics in those networks. On the other hand, the interconnectivity scores give rise to a network of biological processes, indicating their interrelationships.

REFERENCES

Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J., Davis,A., Dolinski,K., Dwight,S., Eppig,J. et al. (2000) Gene Ontology: tool for unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Balasubramanian,R., LaFramboise,T., Scholtens,D. and Gentleman,R. (2004) A graph-theoretic approach to testing associations between disparate sources of functional genomics data. *Bioinformatics*, **20**, 3353-3362.

ExperiBase – An Object Model Implementation for Biology

C. Forbes Dewey, Jr., Aidan Downes, Howard Chou, and Shixin Zhang[†]

Massachusetts Institute of Technology, Cambridge MA USA
[†]Presently with Oracle Corporation, Redwood Shores, CA

ABSTRACT

ExperiBase is the instantiation of an object model to handle an important selection of the experimental protocols that currently exist in biology. A consistent data definition strategy is demonstrated that handles gel electrophoresis, microarrays, fluorescence activated cell sorting, mass spectrometry, and microscopy within a single coherent set of information object definitions. Other experimental methods can be added with relative ease because the object model used to describe the data is easily extended. The object model supports previous work done with microarrays (MAGE OM and Miamexpress), HUPO definitions for mass spectrometry, and the object model for microscopy proposed by OME.

1 INTRODUCTION

Over the past several years, the explosive growth of biological data generated by new high-throughput instruments has literally begun to drown the biological community. There is no infrastructure to deal with all of these data in a consistent and successful fashion. Available methods generally store the data either in spreadsheet form or in flat files where, in advanced cases, the files contain metadata identifying some of the parameters of the experiment that produced the data. It is nearly universally true that the data are in general not queryable across important properties such as the sample origin and treatment, instrument settings, or the results of analytical methods applied to the data. It is also often the case that the scope of the annotating metadata is far less than is required to create self-describing information objects; file names, ad hoc context clues such as directories for files, and file dates substitute for defined queryable fields.

ExperiBase is an object model that supports the use of extensive metadata to describe a biological experiment, its results, and its relation to the outside world including access privileges and projects to which it is related. It is an object model that supports object models and ontologies proposed by others so that many different experimental methods can be treated in a consistent and uniform manner, and the data from different experimental methods can be stored and queried in a uniform way. Considerable economy of effort and an improved ability to support new and evolving requirements results.

Many of the paradigms used in ExperiBase follow logically from the object models developed for medical images by the American National Standards Institute and the National Electrical Manufacturers Association (ACR-NEMA). The resulting standard, called DICOM, was completed in 1993 and has been an enormous success, allowing all of the major medical image modalities (MR, CT, Ultrasound, X-Ray, ECG, Pathology images, Angiograms, and Nuclear images) to be treated in a uniform and consistent manner. The4 existence of active standing committees for the different specialties has allowed the standard to evolve over a number of years, adding new methods and evolutionary changes to the existing modalities. The standard has been actively used as the primary tool for handling digital medical images since its inception; every piece of medical imaging equipment produced in the world must support the standard if the manufacturer wishes to sell it.

2 METHODS

Several important experimental techniques in contemporary biology have been used to create a single composite schema. The results bear a striking relationship to the DICOM standard of 1993 that provides information object definitions of all of the major medical imaging modalities (MR, CT, US, XA, NM, VL, CR, and Waveforms). The *de novae* information object definitions developed for gel electrophoresis by the authors of this paper were found to be very similar to the existing MAGE-OM information model for microarrays. Further investigation revealed that similar object definitions characterized other experimental biology methods as well. These were generalized and a full object-relational data schema was developed. The appended references cite a number of the proposed standards that were used to develop the object model.

3 RESULTS

A first implementation of this work is called *ExperiBase*. It can store and query data generated by the leading experimental protocols used in biology within a single database. ExperiBase also has provisions to store derived data from analysis as a part of an expanded definition of the information object. Transport of the raw data and analytical results

* To whom correspondence should be addressed.

between ExperiBase and external analysis packages currently uses web-based network technologies and XML representation of the data itself. The information object model is used to define the form of the XML data document. Import and export of data in spreadsheet format is also supported. ExperiBase has been ported to three leading database platforms: Oracle, DB2 and Informix. There are no platform-specific dependencies other than the necessity to support object models and large binary data types in efficient native format. From an implementation point of view, the database in which ExperiBase is implemented should support sparse data matricies with no significant storage penalties.

Figure 1 is a high-level view of the organization of the data and metadata that, together, comprise a single experiment. Figure 2 is a condensed view of the expanded object model containing the elements used to describe each of the top-level concepts shown in Fig. 1.

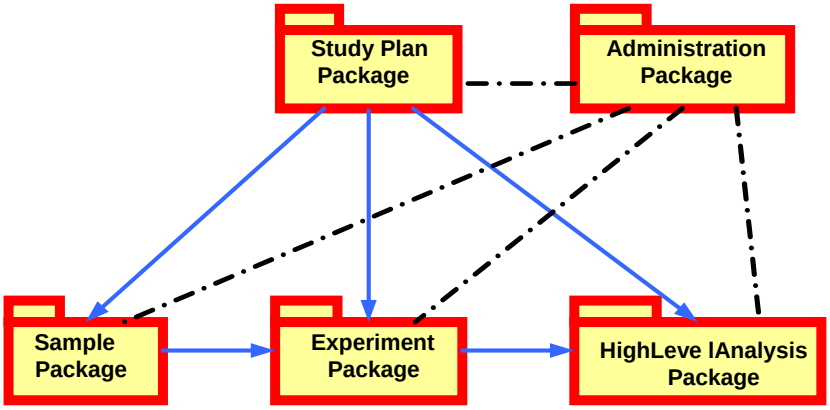


Figure 1 Biological experimental data can be grouped into five packages: study plan (also called as project), sample, experiment, high level analysis, and administration package.

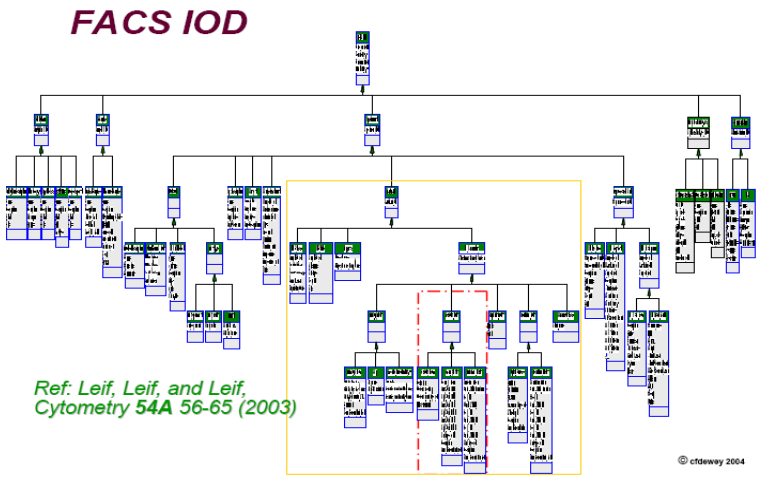


Figure 2 The Information Object Definition for fluorescent activated cell sorting following Leif et al.

4 CONCLUSIONS

The medical and biological communities are invited to participate in this effort to develop international standards to handle the massive data collections that are now being created in every pharmaceutical company and every academic biology laboratory. Having consistent formats for the information objects will greatly speed the development of analysis tools

ACKNOWLEDGMENTS

This research was supported by the Defence Advanced Research Projects Agency and the Pacific Northwest National Laboratories (Department of Energy). We are especially grateful for the active participation and critical contributions of Steven Wiley and Ron Taylor of PNNL.

REFERENCES

Taylor, CF et al. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnology*, **21**, 247-254.

Leif, RC. Leif, SH. Leif, SB (2003) CytometryML, An XML Format based on DICOM for Analytical Cytology Data. *Cytometry* **54** 56-65.

Swedlow JR, Goldberg I, Brauner E, Sorger PK (2003) Informatics and quantitative analysis in biological imaging. *Science* . **300**:100-102.

Brazma, A et al. (2001) Minimum information about a microarray experiment (MIAME) – towards standards for microarray data. *Nature Genetics* **29** 365-371.

kWhy GO there? Ensuring that the Gene Ontology Meets Biologists' Needs

Midori A. Harris

The Gene Ontology Consortium and EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, UK

ABSTRACT

The Gene Ontology (GO) project (http://www.geneontology.org) is a collaborative effort to construct and use ontologies to facilitate the biologically meaningful annotation of genes and their products in a wide variety of databases. Participating groups include the major model organism databases and other database groups.

The GO ontologies provide a systematic language for the description of attributes of gene products, in three key domains that are shared by all organisms: Molecular function describes elemental roles, such as catalytic or binding activities, at the molecular level. Biological process describes broad objectives, each accomplished by one or more ordered assemblies of molecular functions. Cellular component describes locations where a gene product may act, and includes both subcellular structures and macromolecular complexes.

From its inception, the GO project has developed its ontologies for the express purpose of gene product annotation. As biological research advances, and as the GO vocabularies are applied to more species, the ontology content must be continually refined so as to remain useful for the purpose of functional annotation. The model organism database curators who use GO terms for gene product annotation therefore play a key role in guiding the development of GO: The interaction between ontology editors and annotators lies at the core of the GO update procedure, ensuring that GO terms reflect usage in the biomedical literature to the greatest extent possible, and that the relationships between terms accurately capture biological knowledge.

When GO is applied to new organisms, many ontology changes are typically required to accommodate biological systems not previously represented. For example, when The Arabidopsis Information Resource (TAIR) joined the GO Consortium in 2000, the ontologies– previously applied only to animal and fungal gene products—were extended to cover plant biology. Many more such changes lie ahead, as the GO Consortium has begun an effort to actively encourage new groups to use GO for gene product annotation, and to make the resulting annotation data available to the public as part of the GO repository. The understanding that GO will adapt its ontologies as needed to permit accurate and complete annotation in any species is a key factor in GO's widespread uptake by the biological database community. (Improvements made at the request of annotators additionally benefit many other users, both within and outside the GO Consortium, who may use GO data for a wide variety of applications.)

To complement input from the annotators who use GO intensively, the GO Consortium strives to involve members of the research community in the ontology development process. Experts in various biomedical fields can provide thorough, detailed knowledge of their particular topics that complements GO curators' understanding of existing GO structures and conventions. The GO Consortium uses several mechanisms to promote communication among these various contributors and ensure consistency within the ontology even as its content is modified. All changes to the ontologies are centrally coordinated by the GO Editorial Office staff, who have backgrounds in biology. A small group of curators—also biologists—have write access to the CVS repository in which GO files are maintained. To supplement mailing lists and online documentation, we have adapted the online tracking system provided by SourceForge to manage suggestions for changes to the ontologies (see http://geneontology.sourceforge.net/). With the SourceForge system, any user can track the status of a suggestion, see changes currently under consideration, and comment on suggestions; all suggestions and discussions are archived. In 2002 the GO Consortium established Curator Interest Groups to focus on areas within the ontologies that are likely to require extensive additions or revisions. Interest Groups may include outside experts as well as curators from GO Consortium databases. Of the 28 Interest Groups now established, the four most active also have archived mailing lists associated with them. The latest addition to the range of ontology development approaches is the initiation of a series of meetings devoted to specific biological content areas in GO. At the first content meeting, held last year, three topics were selected, and members of the relevant research community joined GO curators to determine how GO should represent each area.

A different source of suggested changes is computational analysis of existing GO terms and relationships, which can identify missing relationships and missing or misplaced terms. These computational efforts, most notably the OBOL project (see http://www.fruitfly.org/~cjm/obol/), improve the logical consistency of GO, and will eventually enable GO to adopt more formal computational representations for its ontologies.

Similar principles, both sociological and technical, govern the development of other ontologies in the Open Biomedical Ontologies (OBO) collection The GO Consortium welcomes feedback from the biology and bioinformatics communities on any aspect of its ontologies or their use.

Automating Ontological Function Annotation: Towards a Common Methodological Framework

Cliff A Joslyn[*]*, Judith D Cohn, Karin M Verspoor, and Susan M Mniszewski*

Los Alamos National Laboratory, Los Alamos, NM, USA

ABSTRACT

Motivation: Our work in the use of ontology categorization for functional annotation is motivating our focus on an overall methodological framework for ontological function annotation (OFA). We draw on our experiences to discuss test set selection, annotation mappings, evaluation metrics, and structural ontology measures for general OFA.

1 INTRODUCTION

A new paradigm for functional protein annotation is the use of automated knowledge discovery algorithms mapping sequence, structure, literature, and/or pathway information about proteins whose functions are unknown into a functional ontology, typically (a portion of) the Gene Ontology (GO, GO Consortium 2000)[1]. For example, our own work (Verspoor *et al.* 2004, 2005) involves analyzing collections of GO nodes (e.g. annotations of protein BLAST neighborhood) using the POSet Ontology Categorizer (POSOC, Joslyn *et al.* 2004)[2] to produce new annotations. Both in executing this work and in examining similar efforts (e.g. Pal and Eisenberg 2005, Martin *et al* 2004), we have uncovered a variety of methodological issues which we believe could be valuable for the community to focus on. Here we first explicate our sense of a generic architecture for automated ontological functional annotation (OFA) into the GO, and then discuss specific methodological issues which are generic to OFA, illustrated by our own experience.

2 GENERIC AUTOMATED OFA

A simple formulation for protein function annotation into the GO assumes a collection of genes or proteins X and a set of GO nodes (perhaps for a particular branch) P. Then in the most general sense, annotation is a function $F : X \to 2^P$ assigning each protein $x \in X$ a collection of GO nodes $F(x) \subseteq P$. So while a known protein x may have a known set of annotations $F(x)$, a new protein y may not have any known annotations, and instead we wish to build some method G returning a predicted set of GO nodes $G(y) \subseteq P$. Typically, we have information about y such as sequence, structure, interactions, pathways, or literature citations, and to build G we exploit knowledge of the proteins "near" y in that space which have known functions. In a testing situation, we take a known protein x and compare its known annotations $F(x)$ against its predicted annotations $G(x)$. Thus to measure the accuracy of our prediction G, we need to compare two different sets of GO nodes, $F(x)$ and $G(x)$, against each other over the set of known proteins X.

3 METHODOLOGICAL ISSUES

We now briefly survey the methodological issues we will explicate completely in the presentation and full paper.

3.1 Protein Test Sets

First we select one or more "gold standard" test sets X of proteins with trusted annotations in the GO. While any such test set should be shared within the community, nonetheless requirements for a gold standard will vary among research groups. POSOC currently needs a test set containing both sequence and structure data, and so we use Swiss-Prot protein sequences with existing PDB structures[3]. Other groups have used a variety of test sets, for example Pal and Eisenberg (2005) use a set of protein sequences from the FSSP structure library[4] to evaluate their ProKnow system; Martin *et al* (2004) use sequence data from seven complete genomes to test GOtcha.

A further consideration is non-redundant test data which is sampled to avoid over-representation in any part of the test space. For example, the non-redundant Astral subsets of SCOP domains are designed to cover the variation in SCOP structure space while ensuring that no two SCOP domains in a particular subset have a sequence homology greater than a specified cutoff value (e.g. 95% or 40%) (Chandonia *et al.* 2004). We propose development of a non-redundant test set covering GO function space.

3.2 Annotation Mappings

The value of any gold standard is very much tied to the accuracy of their known annotations F. POSOC uses the GOA[5] UniProt[6] annotation set for protein sequences, and it

[*] To whom correspondence should be addressed: MS B265 LANL, Los Alamos, NM 87545 USA, joslyn@lanl.gov

[1] http://www.geneontology.org

[2] http://www.c3.lanl.gov/~joslyn/posoc.html

[3] http://www.rcsb.org/pdb

[4] http://www.chem.admu.edu.ph/~nina/rosby/fssp.htm

[5] http://www.ebi.ac.uk/GOA

could be useful for this set, or other annotations for other data types, to be regularized as a community standard to provide a means of comparing various studies, including studies attempting to create better annotation sets. Extension to include the source of annotations for a particular type of data and a common ranking for the evidence codes included in GO annotation files (e.g. IC = inferred by curator, IEA = inferred from electronic annotation), as implemented by Pal and Eisenberg (2005), could also be very helpful.

3.3 Evaluation Metrics

Given a gold standard *X*, annotation mapping *F*, and prediction function *G*, we next need evaluation metrics which compare "ground truth" annotations *F(x)* against predictions *G(x)*, typically comparing ratios of true and false positives and negatives. While precision, recall, and F-score are standard measures, some architectures force different choices. The results *G(x)* produced by POSOC do not form a simple set, but rather a ranked list of effectively indefinite length, requiring a non-standard measure of precision. Such alternative measures are available from the information retrieval literature, and include measuring precision at different recall levels, computing average precision at *n* correct results, and others such as mean average precision, *R*-precision (precision at rank equal to the total number of correct results for a given query), and reciprocal rank.

These alternatives need to be considered and evaluated for meaningfulness specifically in the context of annotating into a structured ontology. Kiritchenko *et al* (2005) propose an explicitly hierarchical extension of precision and recall with respect to the subgraph containing the predicted node and all of its ancestors (the "node subgraph") and the node subgraph of the correct node. Pal and Eisenberg (2005) consider precision at various ontology depths, hierarchically matching nodes in the node subgraph of the predicted node and nodes in the node subgraph of the correct node.

3.4 Ontology Distance Metrics

In the context of the GO, what we even mean by a "true positive" or a "near miss" must be questioned. For example, every annotation to a node $p \in P$ should also be considered an annotation to all its ancestors, and in many cases predicting a parent or grandparent of a correct annotation may be preferable to an "exact match". Currently POSOC measures performance first with respect to both direct hits, and then also "nuclear" (parent, child, sibling) and "extended" (grandparent, uncle, cousin, etc.) family relations between nodes.

Beyond such a simplistic sense of "family" relations, determining the amount of overlap between *F(x)* and *G(x)* generalizes from precision and recall to the determination of aggregated distance measures between all pairs of GO nodes in those sets. Such metrics are still in development either directly (Joslyn 2004, Joslyn and Bruno 2005) or indirectly (Kiritchenko *et al*. 2005). Different concepts of rank (depth) and location in the GO are available, but also still in development, and these all need to be appreciated and internalized by the bio-ontology community better.

3.5 Ontology Structural Statistics

Sets of GO nodes abound in OFA, including at least *F(x)*, *G(x)*, and *G(y)* for various known and unknown proteins *x* and *y*; for POSOC, annotation sets of BLAST neighborhoods around *x* or *y*; and even whole branches CC, MF, and BP of the GO. So central to all OFA methodology is the need for infrastructure to analyze such portions of the GO. We are currently developing mathematical methods and analytical software to measure such statistical properties as the average depth of a node set, the size of the region it circumscribes, the relative amount of back-branching or "multiple inheritance" present, and the extent to which nodes exist "comparably" in vertically connected chains or "noncomparably" in horizontally separated antichains. While partially exploratory, this work has been motivated by the direct need to analyze POSOC results in different GO branches with respect to different kinds of test protein sets, annotation mappings, and evaluation metrics described here.

[6] http://www.ebi.ac.uk/uniprot/index.html

REFERENCES

JM Chandonia, G Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE: (2004) "The ASTRAL Compendium in 2004", *Nucleic Acids Research* 32:D189-D192

Gene Ontology Consortium: (2000) "Gene Ontology: Tool For the Unification of Biology", *Nature Genetics*, v. 25:1, pp. 25-29

CA Joslyn: (2004) "Poset Ontologies and Concept Lattices as Semantic Hierarchies", in *Conceptual Structures at Work, LNAI*, v. 3127, ed. Wolff *et al.*, pp. 287-302, Springer-Verlag, Berlin

CA Joslyn and WJ Bruno: (2005) "Weighted Pseudo-Distances for Categorization in Semantic Hierarchies", *2005 Int. Conf. on Conceptual Structures*, to appear in *Lecture Notes in AI*

CA Joslyn, SM Mniszewski, AW Fulmer and GG Heaton: (2004) "The Gene Ontology Categorizer", *Bioinformatics*, v. 20:s1, pp. 169-177

S Kiritchenko, S Matwin, and AF Famili: (2005) "Functional Annotation of Genes Using Hierarchical Text Categorization", to appear in Proc. BioLINK SIG on Text Data Mining

D Martin, M Berriman, and G Barton: (2004) "GOtcha: A New Method for Prediction of Protein Function Assessed by the Annotation of Seven Genomes". *BMC Bioinformatics* 5:178

D Pal and D Eisenberg, David: (2005) ``Inference of Protein Function from Protein Structure", *Structure*, v. 13, pp. 121-130

KM Verspoor, JD Cohn, SM Mniszewski, and CA Joslyn: (2004) "Nearest Neighbor Categorization for Function Prediction" In CASP 06 abstract book.

KM Verspoor, JD Cohn, CA Joslyn, SM Mniszewski, A Rechtsteiner, LM Rocha, and T Simas: (2005) :"Protein Annotation as Term Categorization in the Gene Ontology Using Word Proximity Networks", *BMC Bioinformatics* 2005 vol 6(suppl 1)

uBioWarehouse: A Bioinformatics Database Warehouse Toolkit

Peter D. Karp

SRI International, 333 Ravenswood Ave. EK207, Menlo Park, CA 94025 USA pkarp@ai.sri.com

ABSTRACT

We introduce the ontology behind BioWarehouse. BioWarehouse is an open source toolkit for constructing bioinformatics database (DB) warehouses using the MySQL and Oracle relational database managers.

1 INTRODUCTION

BioWarehouse addresses the database integration problem in Bioinformatics by integrating its component DBs into a common representational framework within a single DB management system, thus enabling multi-DB queries using the Structured Query Language (SQL) but also facilitating a variety of DB integration tasks such as comparative analyses and data mining.

BioWarehouse currently supports the integration of UniProt (SWISS-PROT and TrEMBL), GenBank, ENZYME, KEGG, BioCyc, NCBI Taxonomy, and CMR. BioWarehouse supports the simultaneous storage of multiple versions of a given DB.

BioWarehouse loader tools, written in the C and JAVA languages, parse and load the preceding DBs into a relational DB schema. The loaders also apply a degree of semantic normalization to their respective source data, decreasing semantic heterogeneity. That is, the schema (ontology) behind BioWarehouse defines a common ontological framework for representing and querying bioinformatics data.

The BioWarehouse schema supports the following bioinformatics datatypes: chemical compounds, biochemical reactions, metabolic pathways, proteins, genes, nucleic acid sequences, features on protein and nucleic-acid sequences, organisms, organism taxonomies, and controlled vocabularies. That is, BioWarehouse can store controlled vocabularies such as Gene Ontology, and a GO loader for BioWarehouse is almost complete.

This presentation will provide an overview of the BioWarehouse architecture, and will present the design of the BioWarehouse schema in detail. We will also discuss the principles that have driven the design of the BioWarehouse schema.

As an application example, we applied BioWarehouse to determine the fraction of biochemically characterized enzyme activities for which no sequences exist in the public sequence DBs. The answer is that no sequence exists for 36% of enzyme activities for which EC numbers have been assigned. These gaps in sequence data significantly limit the accuracy of genome annotation and metabolic pathway prediction, and are a barrier for metabolic engineering. Complex queries of this type provide examples of the value of the data warehousing approach to bioinformatics research.

Availability: BioWarehouse is an open source project that is freely available under the Mozilla license. For information on obtaining BioWarehouse, see URL http://bioinformatics.ai.sri.com/biowarehouse/.

A Novel Ontology Development Environment for the Life Sciences

*Susie Stephens[*] and Mark Musen**

*Oracle, 10 Van de Graaff Drive, Burlington, MA 01803, * Stanford Medial Informatics, Stanford School of Medicine, Stanford University, Stanford, CA 94305-5479.*

ABSTRACT

Motivation: Ontology development environments need to take advantage of scalable, reliable and secure data repositories. This is becoming increasingly important as ontologies become larger in size and the number of simultaneous users grows. This paper describes the merits of integrating Protégé with the RDF Data Model in the Oracle Database.

1 INTRODUCTION

Ontologies provide the common vocabulary for the integration of the hundreds of different knowledge bases, meta-data formats, and database schemas that are used in the biomedical domain. An ontological framework enables researchers to access a knowledge base, appraise its content, determine if resources are relevant, and to integrate and aggregate the data with in-house resources and data.

Semantic Web technologies such as RDF and OWL are being increasingly used for providing the ontological framework as they provide a means to represent data, meta-data about resources, and for defining relations between components of the resources.

2 ARCHITECTURE

2.1 Protégé

Protégé is the most widely used freely available, platform-independent, open-source technology for managing and developing large terminologies, ontologies, and knowledge bases. Protégé has been used as the primary development environment for several projects in the biomedical domain and is supported by a strong community of developers and users. Examples of these projects include Cerner's Clinical Bioinformatics Ontology, MGED Ontology, the Foundational Model of Anatomy (Rosse & Mejino 2004), and verification and identification of errors and inconsistencies in the Gene Ontology (Yeh *et al.* 2003).

Protégé is based on Java, is extensible, and provides a platform for customized knowledge-based applications (Gennari *et al.* 2003). Protégé provides support for building Semantic Web applications through its knowledge model, which is based on the Open Knowledge Base Connectivity (OKBC) protocol (Chaudhri *et al.* 1998). This enables ontology editors to be built for different ontology languages including RDF and OWL.

2.2 Oracle Spatial RDF Data Model

The Oracle Database is the market leading relational database management system (RDBMS) in the biomedical domain. With Oracle Database 10*g* release 2 a new Oracle object type (SDO_RDF_TRIPLE_S) is introduced for storing RDF and OWL data (Alexander *et al*. 2004). This object type is built on top of the Oracle Spatial Network Data Model (NDM), which is the Oracle solution for managing graphs within the RDBMS (Stephens *et al*. 2004).

There are many advantages to storing RDF data as an object type, rather than in flat relational tables. Benefits include making it easier to model and maintain RDF applications, simplifying the integration of RDF data with other enterprise data, re-use of RDF objects, and no mapping is required between client RDF objects and database columns and tables that contain triples.

With the Oracle RDF Data Model triples are parsed and stored in the database as entries in the NDM node$ and link$ tables. Nodes in the RDF model are uniquely stored and reused when encountered in incoming triples. In user-defined application tables, only references are stored in the SDO_RDF_TRIPLE_S object to point to the triple stored in the central schema. The RDF Data Model also simplifies reification by utilizing an Oracle XML DB DBUri to directly reference the reified triple in the database, and thereby only requires one additional triple to be stored for each reification.

2.3 Integration of Protégé with the Oracle Spatial RDF Data Model

In preliminary performance testing the Oracle RDF Data Model is demonstrating comparable performance to that obtained with a relational-based storage implementation. It is therefore expected that one of the main benefits of this novel architecture is the ability to manage RDF applications more easily, and a more performant approach to data reification.

[*] To whom correspondence should be addressed.

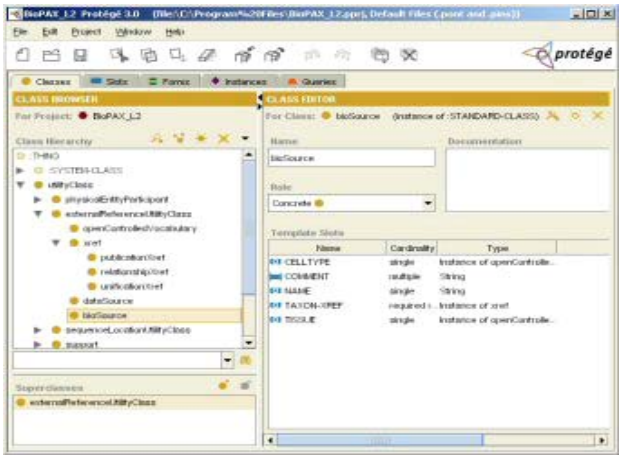


Fig. 1. BioPAX Ontology in Protégé

ACKNOWLEDGEMENTS

We would like to acknowledge the Oracle Spatial Development group for the implementation of the RDF Data Model.

REFERENCES

Alexander, N., Lopez, X., Ravada, S., Stephens, S. and Wang, J. (2004) RDF Data Model in Oracle. http://lists.w3.org/Archives/Public/public-swls-ws/2004Sep/att-0054/W3C-RDF_Data_Model_in_Oracle.doc

Chaudhri, V., Farquhar, A., Fikes, R. Karp, P. and Rice, J. (1998) *OKBC: A programmatic foundation for knowledge base interoperability*. In: Fifteenth National Conference on Artifical Intelligence (AAAI-98), 600-607. Madison, Wisconsin: AAAI Press/The MIT Press.

Gennari, J., Musen, M. A, Ferguson, R. W., Grosso, W. E., Crubezy, M., Eriksson, H., Noy, N. F., and Tu, W. W. (2003) *The evolution of Protégé: An environment for knowledge-based systems development*. International Journal of Human-Computer Interaction 58(1).

Rosse, C., and Mejino, J. L. V. (2004) A reference ontology for bioinformatics: The foundational model of anatomy. *J. Biomed. Informat.*

Stephens, S., Rung, J. and Lopez, X. (2004) *Graph Data Representation in Oracle Database 10g: Case Studies in Life Sciences.* IEEE Data Engineering Bulletin. http://sites.computer.org/debull/A04dec/stephens.ps

Yeh, I. Karp, P., Noy, N. and Altman, R. (2003) Knowledge acquisition, consistency checking and concurrency control for gene ontology (GO). *Bioinformatics* 19:241-248.

FuGO: Development of a Functional Genomics Ontology (FuGO)

Patricia L. Whetzel[1], Helen Parkinson[2], Assunta-Susanna Sansone[2] Chris Taylor[2], and Christian J. Stoeckert, Jr.[1]*

Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA[1], EMBL-European Bioinformatics Institute, Hinxton, Cambridge, UK[2]

ABSTRACT

Motivation: Data standards and object models are being developed for a variety of functional genomics domains. Many of these object models include a reference to an ontology concept in order to provide a rich set of terms for annotation. The MGED Ontology was developed to provide terms to be used with the MicroArray and Gene Expression Object Model and has been successfully implemented in production annotation applications. This work is being used as the foundation to develop a Functional Genomics Ontology, intended to model additional functional genomics domains such as Proteomics, Metabol/nomics Toxicogenomics, Environmental Genomics and Nutrigenomics as well as Transcriptomics.

1 INTRODUCTION

The MGED Ontology (MO) was developed as a collaborative effort by members of the MGED Ontology working group to provide descriptors required to interpret microarray experiments (Stoeckert and Parkinson, 2003). The concepts that these descriptors represent were derived from the MicroArray and Gene Expression Object Model (MAGE-OM), which is a framework to represent gene expression data and relevant annotations (Spellman *et al*., 2002). The MAGE-OM contains a mechanism to add annotations from an ontology by providing an association from a given class to the MAGE-OM class OntologyEntry. The MO provides descriptors for the concepts in the MAGE-OM that have associations to the class OntologyEntry. The MO also includes concepts from the MAGE-OM to indicate what object model class the terms are to be used for, but does not replicate the entirety of the object model. These policies for ontology development have resulted in a rich and expressive ontology that is fully supportive of the MAGE-OM and is commonly used in microarray annotation applications.

Since the development of MO, other functional genomics domains have planned to or are developing an ontology (Pedrioli *et al*., 2004). This wider functional genomics context will significantly affect the structure and content of ontologies. Core biological descriptors need to be shared, as well as descriptors relating to the experimental design, sample generation and treatments, therefore requiring extensive liaisons between communities. The MGED Ontology Working Group (http://mged.sourceforge.net/ontolgies/index.php), the MGED Reporting Structure for Biological Investigations (http://www.mged.org/Workgroups/rsbi/rsbi.html), the HUPO Proteomics Standards Initiative (http://psidev.sourceforge.net/) and the Standard Metabolic Reporting Structure (http://www.smrsgroup.org/) working groups can clearly draw in large numbers of experimentalists and developers and feed in the domain-specific knowledge of a wide range of biological and technical experts. This extensive collaboration aims to develop a Functional Genomics Ontology (FuGO) by expanding the scope of the MGED Ontology to model other functional genomics technologies, such as Proteomics, Metabol/nomics as well as Transcriptomics and biological domains, including Toxicogenomics, Nutrigenomics and Environmental Genomics. The resulting ontology will provide a consistent mechanism for annotating functional genomics experiments that encompasses different technological and biological domains and aid in cross-comparison of data.

2 METHOD: DESIGN PRINCIPLES

FuGO is designed to model the functional genomics domain. That is, all concepts required to model the domain are included in the Functional Genomics Ontology. This is in contrast to the design of the MGED Ontology, which was developed to provide terms for object model classes in the MAGE-OM. The decision to include all concepts within the functional genomics domain in the ontology is based on past experience regarding the use of the MGED Ontology. Although the goal of the MGED Ontology was to provide descriptors for concepts and to be used in conjunction with the MAGE-OM, other applications were developed that used the ontology itself as a model of the microarray domain. Development of the Functional Genomics Ontology will be done in parallel with efforts to develop a functional genomics object model and therefore the ontology will be designed to provide descriptors for concepts in these object models that require an ontology annotation as well. A middle layer, which provides the mapping of object model classes to those in FuGO, will be generated to aid the use of FuGO

* To whom correspondence should be addressed.

with functional genomics object models. Therefore, FuGO will be designed to take into account both of these possible uses of the ontology.

2.1 Scope of FuGO

The current scope of FuGO is the following domains: Transcriptomics, Proteomics, Nutrigenomics, Environmental Genomics and Toxicogenomics. The top node classes will include the classes Common and Bio, which will contain concepts that are shared across functional genomics domains. In addition, classes to hold concepts that are specific to a given functional genomics domain will also be included.

During the re-engineering of MO to model additional functional genomics domains, classes that represent concepts that exist in two or more functional genomics domains will be added as children to either Common or Bio as is appropriate. Additional branches will exist in the FuGO to hold technology specific concepts such as Transcriptomics, Proteomics, Metabol/nomics and others.

2.2 Ontology Development Process

The initial development of FuGO will involve re-engineering MO by moving classes from the MAGE-OM package structure into the classes Common and Bio. In addition, definitions will be modified to remove references to the MAGE-OM. Ontological changes will also be included in the development of FuGO such as moving individuals to classes and using properties to define classes based on reasoning over the ontology.

New concepts that are required to represent functional genomics domains will be added to FuGO in branches for these domains. If the functional genomics domain requires a new concept that is unique to the domain, the concept will be added to the domain specific branch of the ontology and community members that represent the domain will be responsible for approving the term and definition. If a domain requires a term that represents an existing concept in Common or Bio, the term will be added as subclass to the appropriate class and will be marked as being specific to the domain using properties. In this case, members of the domain will be responsible for approving the term and definition. Lastly, if the concept can be used by two or more functional genomics domains, the term will be added to either Common or Bio as appropriate and the term will be approved the all those involved in the ontology development.

2.3 Implementation Details

FuGO will be developed in Protégé using the OWL plugin (Noy *et al*., 2003). The current working version of FuGO is available as a downloadable OWL format file (http://mged.fuge.net/ontologies/FuGO.owl). Protégé was selected as it is an expressive system, covering frame based and description logics with a variety of export formats.

ACKNOWLEDGEMENTS

The authors would like to thank the members of the Ontology Working Group: Helen Causton, Liju Fan, Jennifer Fostel, Gilberto Fragoso, Laurence Game, Mervi Heiskanen, Norman Morrison, Philippe Rocca-Serra, and Joe White for their contribution to this work.

REFERENCES

HUPO Proteomics Standards Initiative http://psidev.sourceforge.net/

MGED Ontology Working Group: http://mged.sourceforge.net/ontolgies/index.php

MGED Reporting Structure for Biological Investigations http://www.mged.org/Workgroups/rsbi/rsbi.html

Noy N.F., Crubezy M., Fergerson R.W., Knublauch H., Tu S.W., Vendetti J., Musen M.A. (2003) Protege-2000: An Open-source Ontology-development and Knowledge-acquisition Environment. *AMIA Annu Symp Proc*., 953.

Pedrioli P.G., Eng J.K., Hubley R., Vogelzang M., Deutsch E.W., Raught B., Pratt B., Nilsson E., Angeletti R.H., Apweiler R., Cheung K., Costello C.E., Hermjakob H., Huang S., Julian R.K., Kapp E., McComb M.E., Oliver S.G., Omenn G., Paton N.W., Simpson R., Smith R., Taylor C.F., Zhu W., Aebersold R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol*., **22**, 1459-1466.

Standard Metabolic Reporting Structure: http://www.smrsgroup.org/

Spellman P.T., Miller M., Stewart J., Troup C., Sarkans U., Chervitz S., Bernhart D., Sherlock G., Ball C., Lepage M., Swiatek M., Marks W.L., Goncalves J., Markel S., Iordan D., Shojatalab M., Pizarro A., White J., Hubley R., Deutsch E., Senger M., Aronow B.J., Robinson A., Bassett D., Stoeckert C.J. Jr., Brazma A. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*., **3**, research0046.1-0046.9.

Stoeckert, C. J. Jr. and Parkinson H. (2003) The MGED ontology: a framework for describing functional genomics experiments. Comparative and Functional Genomics, **4**, 127-132.

Integration of the Gene Ontology into an object-oriented architecture

*Daniel Shegogue and W. Jim Zheng**

Dept. Biostat., Bioinformatics & Epidemiology, Med. Univ. of South Carolina, Charleston, SC 29425

ABSTRACT

Motivation: Gene Ontology (GO) has been categorized into biological processes, molecular functions, and cellular components. However, there is no single representation that integrates all the terms into one cohesive model. Furthermore, GO definitions have little information explaining the underlying architecture that forms these terms, such as the dynamic and static events occurring in a process. In contrast, object-oriented models have been developed to show dynamic and static events. A portion of the TGF-beta signaling pathway, which is involved in numerous cellular events including cancer, differentiation and development, was used to demonstrate the feasibility of integrating the Gene Ontology into an object-oriented model.

1 INTRODUCTION

Three independent ontologies, molecular function, biological process, and cellular component domains, have been developed to describe gene products. When applied to a gene, that gene is annotated with a concise description using these ontologies. It has been noted that there remains a need for a unifying architecture that integrates all three GO domains as part of a gene product's annotation. Furthermore, to enhance the Gene Ontology and facilitate its use as a cross-disciplinary tool, several additional issues need to be addressed. First, relationships between the biological processes, molecular functions and cellular components are not readily apparent [1-5]. Second, GO terms lack details. For instance, when one looks at molecular function there is no indication of what is inputted or outputted. Finally, existing tools such as GO-DEV [26] only contain software used for tool development and information retrieval, not software modeled directly after the three domains of the Gene Ontology. However, these issues can be resolved by integrating the Gene Ontology into an object-oriented system.

On a conceptual level, the Gene Ontology has features that support an object-oriented architecture. For example, the functions of gene products are captured in the molecular function domain of the Gene Ontology. These are analogous to the operations that an object can perform in an object-oriented paradigm. Attributes, which define key properties of a component that when changed may alter the function of that component, may be defined by the cellular component and molecular function sections. In addition, each biological process terms can be viewed as a use case in an object-oriented model. However, GO biological process terms do not contain descriptive information about the dynamics or static interactions defined by the terms. By translating a biological process into an object-oriented model the dynamic and static events occurring within a process can be represented. In addition, building a static and dynamic model of a biological process requires defining the components of the process as well as the functions and attributes contained within these components. These components are biological entities (bioentities) that may include individual gene products, whose processes, functions and cellular components are captured in the Gene Ontology, or other higher-level entities such as gene product complexes. As a result, a complete object-oriented model can integrate three domains of Gene Ontology.

The unified modeling language has been used to capture various aspects of biology [6-8]. These examples highlight the utility of the unified modeling language as a tool for biological data integration, and indicate that it can be applied to construct large, complex biological models. Therefore, to demonstrate the feasibility of integrating the Gene Ontology into an object-oriented model we have created unified modeling language (UML) representations of a GO biological process, "transforming growth factor beta (TGF-beta) receptor complex assembly" (GO:0007181).

2 RESULTS

The TGF-beta receptor pathway is involved in numerous cellular events including apoptosis, tumor development, differentiation, and development. These processes stem from the binding of TGF-beta to its cellular receptors (TGF-beta receptor complex assembly, GO:0007181). Object-oriented model was constructed using a linear, sequential software engineering process [8].

2.1 Sequence diagram generation

The GO biological process term, TGF-beta receptor complex assembly (GO:0007181), contains both static and dynamic features. The events of the TGF-beta receptor com-

* To whom correspondence should be addressed. zhengw@musc.edu

plex assembly (GO:0007181) process include TGF-beta binding (GO:0050431) to its receptors and SMAD binding (GO:0046332) and activation (GO:0042301). To capture the dynamic nature of these actions as an object-oriented software system, sequence diagrams were created. The events leading to Smad 2 activation are reflected chronologically in a high-level sequence diagram. The creation of the sequence diagram first entails identifying gene products and their functions by literature searches. Simple or complex bioentities are modeled as objects, which are represented by rectangles with vertical lifelines in the diagram. Ontology terms taken from the molecular function domain that best corresponded to these functions were incorporated as object functions, which represent the functions of these gene products. These functions are implemented by the methods contained within the objects. Furthermore, these methods allow an object to communicate and interact with other objects, thus capturing cellular activities. To capture interactions between objects, one object can call a method of another object by connecting object lifelines in the sequence diagram. This invocation of a function of one object by another is described as one object sending a message to another object. Alternatively, a message may be passed from an object to itself as in the case of self-checks or auto-activation signals. In this way, real world processes may be captured using an object-oriented approach. For instance, to capture the formation of the TGF-beta and TGF-beta RII complex a GOid that closely corresponds to this ability is chosen as the method name. In this way the method can be cross-referenced to a GO term.

2.2 Activity diagram generation

Biological processes are created from a series of complex events. While there may be one main event scenario that most frequently leads to a specific outcome often, alternative scenarios that lead to a process conclusion exist. This is exemplified by the sequence of events found in the TGF-beta receptor complex assembly (GO:0007181). For instance, TGF-beta may initially bind to TGF-beta RII or TGF-beta RIII. To capture these alternative events as part of the dynamic architecture, an activity diagram was created to reflect the initial stages of TGF-beta signaling (Figure 3). Unlike the sequence diagram, which captures main scenario events, the action sequence or flow of the activity diagram can portray alternative outcomes. Taking the example above, if TGF-beta binds to the type III receptor then an alternative flow of events occurs for a time that then returns to the main flow of events. Other possible divergences that were modeled included whether to internalize the TGF-beta receptors via clathrin-dependent or lipid raft-dependent mechanisms. These pathways lead to either complex degradation or signal promotion. Because complex degradation is not specified in our use case, for simplicity, this event is routed to the final state. However, the main success scenario, signal promotion, continues until SMAD2 is released and TGF-beta complex assembly is finished. Together, the dynamic events occurring during the biological process, TGF-beta receptor complex assembly (GO:0007181) are captured

2.3 Class diagram generation

The major components of a biological system are bioentities with functions and interactions. Likewise, the center of an object-oriented software system is objects. Complex bioentities formed from multiple gene products along with their relationships, are contained within the biological system encompassing the biological process term, TGF-beta receptor complex assembly (GO:0007181). To represent the components that execute the process, we captured these components as bioentities with functions, and their interactions. The events of the TGF-beta receptor complex assembly (GO:0007181) process include TGF-beta binding (GO:0050431) to its receptors, and SMAD binding (GO:0046332) and activation (GO:0042301). To capture this static architecture, class diagrams were generated that model the bioentities, operations, and interrelationships that occur between TGF-beta, its receptors, and Smad 2.

ACKNOWLEDGEMENTS

Daniel Shegogue is supported by NLM training grant 5-T15-LM007438-02. W. Jim Zheng is partly supported by a grant (DE-FG02-01ER63121) from the Department of Energy.

REFERENCES

1. Zhang S, Bodenreider O: Comparing Associative Relationships among Equivalent Concepts Across Ontologies. Medinfo 2004, 2004:459-466.
2. Smith B, Williams J, Schulze-Kremer S: The ontology of the gene ontology. AMIA Annu Symp Proc 2003:609-613.
3. Ogren PV, Cohen KB, Acquaah-Mensah GK, Eberlein J, Hunter L: The compositional structure of Gene Ontology terms. Pac Symp Biocomput 2004:214-225.
4. Smith B, Kumar A: Controlled vocabularies in bioinformatics: a case study in the gene ontology. DDT: BIOSILICO 2004, 2(6):246-252.
5. GO-DEV: http://www.godatabase.org/dev/index.html.
6. Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I et al: A systematic approach to modeling, capturing, and disseminating proteomics experimental data. Nat Biotechnol 2003, 21(3):247-254.
7. Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M et al: Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 2002, 3(9):RESEARCH0046.
8. Shegogue D, Zheng WJ: Object-oriented biological system integration: a SARS coronavirus example. Bioinformatics 2005.