Sixth Annual Bio-Ontologies Meeting
Brisbane, 2003

Chaired by Robert Stevens & Robin Mcentire

28th June 2003

Contents

1 Programme **5**

2 Indexing Medline with GO **7**

2.1 The Problem 7

2.1.1 Literature Search 8

2.1.2 Ontologies 8

2.2 The Solution 9

2.3 GO-KDS 9

3 Biological Natural Language Processing **13**

3.1 Introduction 13

3.2 GO as a Source of Lexical Data 15

3.3 Text as a Source of Ontological Data 16

3.4 Integration into the NLP System 16

4 Ontologies for Text Analysis **19**

5 Half Way Up The Ladder **25**

6 Large Scale Ontologies **29**

7 Bio-data Warehouse **33**

8 BioPAX **37**

8.1 Introduction 37

8.2 Objectives and Scope 38

8.3 Use Cases 38

8.4 Ontology Development 39

8.5 Summary of Progress to Date 39

8.6 History of the BioPAX Initiative 39
8.7 How to Contact BioPAX 40

9 Structural Classification 41

10 The Sequence Ontology 47

11 The Physiome Project 49

12 Disease Ontology 53

13 The Rat Genome Database 55

Chapter 1

Programme

08:00 – 08:45	**Registration**
08:45	**Introduction**
09:00	BioLink Text Mining SIG *Christian Blaschke*
09:30	All of MEDLINE Indexed to the Gene Ontology *Tony C. Smith and John G. Cleary*
10:00	Interactions Between the Gene Ontology and a Domain Corpus for a Biological Natural Language Processing Application *Cornelia M. Verspoor, Cliff Joslyn, and George Papcun*
10:30 – 11:00	Morning Coffee
11:00	Using Ontologies for Text Analysis *Lawrence Hunter and K. Bretonnel Cohen*
11:30	Half Way Up The Ladder *M.W. Wright, E.A. Bruford, M.J. Lush, R.C. Lovering, V.K. Khodiyar, C.C. Talbot Jr.2, H.M. Wain, and S. Povey*
12:00	Large Scale Ontologies for Information Retrieval *Nick Tilford*
12:30 – 13:30	Luncheon
13:30	Application of Gene Ontology in Bio-data Warehouse *Shunliang Cao, LeiQin, WeiWang, Yangyong Zhu, and YiXue Li*
14:00	BioPAX - Data Exchange Ontology for Biological Pathway Databases *BioPAX Group*
14:30	Structural Classification in the Gene Ontology *Cliff Joslyn, Susan Mniszewski, Andy Fulmer, and Gary Heaton*
15:00 – 15:30	Afternoon Tea
15:30	A Report on the Sequence Ontology *Suzanna E. Lewis, Karen Eilbeck, Michael Ashburner, Judith Blake, Michele Clamp, Richard Durbin, Lincoln Stein, Colin Wiel, Mark Yandell, and Christopher J. Mungall*
16:00	Ontologies for the Physiome Project *Poul Nielsen, Matt Halstead, and Peter Hunter*
16:30	Disease Ontology: Structuring Medical Billing Codes for Medical Record Mining and Disease Gene Association *Patricia Dyck and Rex Chisholm*
17:00	Integration of Ontology Data onto the Rat Genome Database (RGD) System *Norberto B. de la Cruz, Simon Twigger, Jedidiah Mathis, and Peter J. Tonellato*
17:30	**Finished**

Table 1.1: Programme for Sixth Annual bio-Ontologies Meeting: Brisbane 2003

Chapter 2

All of MEDLINE indexed to the Gene Ontology

Tony C. Smith[1] and John G. Cleary[2]
[1]Department of Computer Science
University of Waikato
Hamilton, New Zealand
tcs@cs.waikato.ac.nz
[2]Reel Two Ltd.
9 Hardley Street
Hamilton
New Zealand
jcleary@reeltwo.com

2.1 The Problem

Most of what is known about genes and proteins is contained in the biomedical literature. The problem for the biologist is how to connect novel sequence data to relevant published documents (3). One way is to BLAST the sequence and then follow the literature links established in genomic/proteomic databases for known sequences with similar structure. Another way is to find the closely-matching genes or proteins in the Gene Ontology, then retrieve documents associated with GO terms. The advantage of this approach is that it provides a conceptual context for discovering possible genetic roles or molecular functions for a new sequence. The problem with both search strategies, however, is that they return only a small portion of the available literature. We are solving this problem by amplifying the available documents associated with GO terms to cover the entirety of the MEDLINE corpus.

2.1.1 Literature Search

It is the literature review that presents the principal obstacle when trying to investigate a gene or protein. There are three fundamental problems: finding relevant documents, synthesising their contents, then organising and ranking the documents in terms of their significance to the present study. Sequence databases typically include a list of references to published work about a gene or protein. The list is often quite short, but provides a starting point. Following citations through the reference sections in each document can also create a fruitful trail to more and more articles, but at some point the researcher usually ends up turning to a literature-search service. Literature searches typically proceed as a keyphrase search over the fulltext index of a large document repository, such as the MEDLINE collection. However, such traditional search techniques frequently miss key documents, or bury them in a slew of irrelevant ones, making it hard for researchers to find what theyre looking for (1). Moreover, the ranking strategy employed by keyphrase search systems may prevent the most relevant articles from making it to the top of the search results where they can be spotted by the user. Such problems arise in keyphrase searching for any subject domain, but they are particularly acute when searching biomedical literature because the terminology is complex, the vocabulary is very large, and the vernacular employed for any given topic is frequently not uniform.

2.1.2 Ontologies

Ontologies provide another, more controlled and systematic way to look for documents. Indeed, one of their primary purposes is to overcome the problems inherent with keyphrase systems, facilitating uniform querying over large biological databases. Moreover, the inherent organisation of ontologies adds taxonomical context to search results, making it easier for the researcher to spot conceptual relationships in data. The Gene Ontology (GO), for instance, organises gene products into a hierarchy of functional categories standardized through the consensus of a consortium of molecular biologists.

The problem, however, has been that very little biomedical literature is expressly linked to terms of the Gene Ontology. For example, a significant effort in the first half of 2002 (made by the authors of this paper) to trawl most of the large publicly available genomic/ proteomic databases[1] turned up fewer than 27,000 MEDLINE documents either directly or indirectly associated to terms from the Gene Ontology. The situation has improved over the past year, and a second trawl done in April 2003 turned up over 110,000 MEDLINE documents with GO associations. The trouble is there are almost six million abstracts in MEDLINE[2], and at this rate it will take a very long time before they all get linked to GO terms. Moreover, the Gene Ontology itself is still in a state of flux with terms being added, deleted, renamed and

[1] The databases used were SwissProt, GenBank, FlyBase, GOA, Gramene Oryza, MGI, PomBase, RGD, SGD, TAIR, TIGR, WB, InterPro and AmiGO.

[2] MEDLINE is usually reported to have around 12 million abstracts, but about half of these are actually corrections and retractions.

moved from month to month, so that the whole process of assigning MEDLINE documents to GO categories has to be repeated often. It seems that the only hope for realising this goal is if the whole process can be automated.

2.2 The Solution

Machine learning technology makes it feasible to put all relevant MEDLINE abstracts into the Gene Ontology. The procedure is to obtain example documents for each GO term, use those examples to *train* classification models for each term (i.e. inferred mathematical characterisations of a semantic category), then use those models to assign the rest of the MEDLINE abstracts to appropriate GO terms.

Raychaudhuri *et al.* (2) demonstrated the feasibility of using machine learning to create document classification models for individual Gene Ontology terms. For their study, they created a training corpus of 16,000 documents for 21 GO terms by retrieving PubMed abstracts using keyphrase queries constructed from MeSH terms. Using maximum entropy models, they achieved classification accuracy of just over 72% on a test corpus of about 200 documents. Given such promising performance, it seems logical to extend the idea to more comprehensive coverage of the Gene Ontology.

2.3 GO-KDS

Like Raychaudhuri *et al.*, we also were looking at developing machine learning techniques to create a practical (though perhaps somewhat more ambitious) document classification system for the Gene Ontology. Our goal was to model as much of the Gene Ontology as possible and then classify all MEDLINE abstracts. The end-product is a publicly available web service called the Gene Ontology Knowledge Discovery System (GOKDS) that can be accessed at `www.go-kds.com`.

A number of significant logistical obstacles had to be overcome to create GO-KDS. For example, there are (as of April 2003) 13,584 terms in the Gene Ontology and about six million MEDLINE abstracts. Very few machine learning algorithms (including maximum entropy models) can scale to this size of problem without running out of system resources or taking forever to produce their results. To remedy this, we developed a new algorithm (loosely described as an optimised Naive Bayes) that can complete the training and classification tasks in about 30 hours on a 1.33 GHz machine using about 1 GByte of memory (space limitations preclude including details of the algorithm here). This level of performance is sufficient to keep up with any changes to the Gene Ontology or MEDLINE.

Leave-One-Out cross-validation experiments show that GO-KDS achieves a 76.7% class average (at the precision-recall breakeven point) in predicting GO terms for MEDLINE abstracts. In an effort to compare this performance against the methods used by Raychaudhuri *et al.* we tried to reproduce the corpus they used and predict documents to the same 21 categories. The result was 70.5% predictive accuracy (at the breakeven point). This is similar to the Raychaudhuri results albeit on a smaller training set and using different evaluation methodology.

To infer an accurate model for a GO term usually requires more than ten sample documents (although some categories have been characterised quite well with as few as four or five exemplars). Raychaudhuri *et al*. were able to obtain between 175 and 1200 examples for each of the 21 GO terms they modeled by using keyphrase queries to retrieve documents from NCBI. This is an effective approach for getting a lot of training data, but runs the risk of causing the learning algorithm to infer models of the query itself, rather than of the Gene Ontology term. To make sure the models characterise the semantics of a GO category *as understood and interpretted by biologists* we only used documents that experts had curated and manually annotated with a GO term and subsequently registered in an established genomic/proteomic database. Although the April 2003 web crawl mentioned earlier only turned up enough sample documents to model 3764 GO terms, the good news is that only another 50,000 or so documents need to be manually linked to GO terms before models of the entire ontology can be inferred[3]. This is considerably fewer than if all six million MEDLINE abstracts had to be annotated by hand!

[3] This estimate of 50,000 assumes needing at least ten exemplars for each of the 9820 unmodeled GO terms with each document being able to serve as training data for several models

References

[1] Chang, J. T. and Altman, R. (2002). Promises of text processing: Natural Language Processing meets AI. Technical report SMI-2002-0934, Stanford Medical Informatics, Stanford, California.

[2] Raychaudhuri, S., Chang, J. T., Sutphin, P. D., and Altman, R. B. (2002). Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research*, 1:203–214.

[3] Yandell, M. D. and Majoros, W. H. (2002). Genomics and natural language processing. *Nature Reviews Genetics*, 3(8):601–610.

Chapter 3

Interactions Between the Gene Ontology and a Domain Corpus for a Biological Natural Language Processing Application

Cornelia M. Verspoor, Cliff Joslyn, and George Papcun
Los Alamos National Laboratory
Computer & Computational Science Division
PO Box 1663, MS B256
Los Alamos, NM 87545
verspoor@lanl.gov|joslyn@lanl.gov|gjp@lanl.gov

3.1 Introduction

In any natural language processing (NLP) application, there is a critical need to manage lexical resources in a manner which supports representation of syntactic and semantic constraints on lexical use. In domains which contain much highly specific terminology, such as the biological domain, it is often a daunting task to construct such lexical resources. We turn, therefore, to existing terminological and ontological resources for the domain. However, while there is significant overlap in the requirements for an NLP system with those of ontological data representations, the requirements are not identical. It is important to consider with care the integration of a lexicon with an ontology into a single application.

Specifically, an NLP system is heavily focused on terminological management issues. Words which are synonymous from the perspective of a given

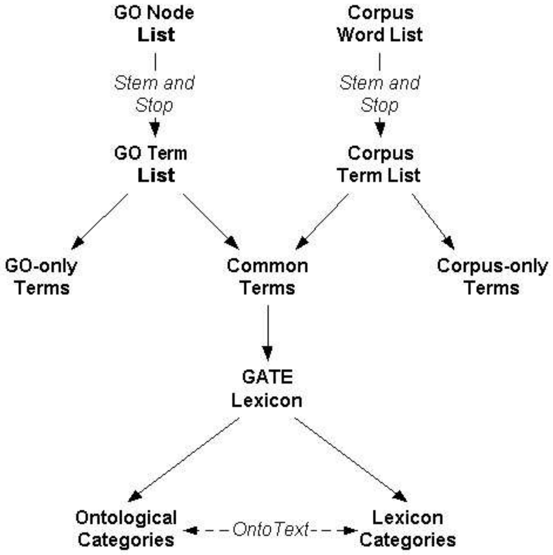

Figure 3.1: Biological natural language processing.

ontology may behave quite differently from a linguistic perspective. The internal structure of a multi-word term is largely irrelevant for ontological use, but may be critical in linguistic processing to support recognition of the term in text, where there may be intervening words or surface variations not captured in the ontology. However, the semantic grounding provided by an ontology can be extremely important for enabling precise analysis of the meaning conveyed in relevant text sources.

We discuss a prototype system, currently under development, that aims to extract regulatory relationships from biological text (3), and which depends on the existence of domain-specific lexical resources. While our customer has supplied some lists of terms that are associated with particular semantic types, these lists are invariably incomplete and exist independently of any domain ontology. We therefore turn to the Gene Ontology (GO, `http://www.geneontology.org`) (1) as a source of richer semantic data for lexical resources. The architecture we follow for construction of those resources is shown in Figure 3.1. Term lists are derived from the GO and a customer-supplied public text corpus respectively, and then stemmed in order to determine distinct term lists. We maintain multi-word terms as phrases in addition to breaking them down into terms consisting of individual words. Finally, certain terms considered to be uninteresting (stop words), including linguistic function words and extremely frequent words, are eliminated from the lists. Terms held in common to GO and the corpus are extracted as the lexicon for our system. The result is a lexicon in which terms can be directly associated with the semantic categories of the domain ontology.

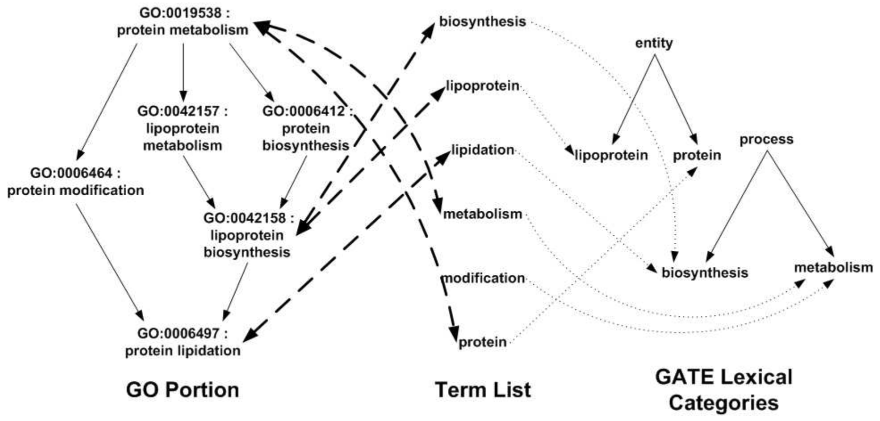
GO:0019538 : protein metabolism
GO:0042157 : lipoprotein metabolism
GO:0006412 : protein biosynthesis
GO:0006464 : protein modification
GO:0042158 : lipoprotein biosynthesis
GO:0006497 : protein lipidation
biosynthesis
lipoprotein
lipidation
metabolism
modification
protein
entity
lipoprotein
protein
process
biosynthesis
metabolism
GO Portion
Term List
GATE Lexical Categories

Figure 3.2: Biological natural language processing.

3.2 GO as a Source of Lexical Data

As a controlled vocabulary, the GO provides an important source of domain-specific terminology that can be used to inform lexicon development for an NLP system. It can be used in the following ways:

- Ontological relations represented in the GO can be reasoned upon in combination with linguistic analysis in order to establish ontological relations among individual terms. We see an example of this type of processing in Figure 3.2, in which relations between heads of phrases are inferred from the relation between the phrases as a whole, e.g. that lipidation is a kind of biosynthesis. We are exploring the extent to which relations in the GO can be exploited in establishing relations between individual terms in the lexicon.

- The hierarchical structure of the GO can be exploited to represent semantic constraints and generalizations in linguistic rules, since each term derived from the GO is associated with a node in the ontology. For instance, a rule may require that a particular argument be some type of protein metabolism. With reference to the GO, we can verify that this holds for a given phrase identified in the text. These types of constraints allow us to more accurately identify particular relationships.*

- Definitions of terms in the GO can be used to establish additional lexical relations; words which are used to define a given word can be assumed to have a contextual relationship with that word. This in turn can be used in the NLP system to support word sense disambiguation in the face of words with multiple meanings or in the case of overlapping multi-word units. This is in the spirit of word sense disambiguation work based on machine readable dictionaries (2).

- Multi-word phrases occurring as nodes in the GO may correspond to non-decomposable word sequences that can be recognized during linguistic parsing to improve structural analysis.

3.3 Text as a Source of Ontological Data

The corpus of domain texts can also be viewed as a source of ontological data that may or may not be represented in the reference ontology. To the extent that the corpus contains information not captured by the ontology, the ontology may be insufficient (depending on its intended purpose). We are exploring the use of NLP technologies to identify ontological relations expressed in the corpus. These relations would be proposed for integration with the ontology, such that it becomes congruent with the corpus. The implemented techniques would draw on the lexicon, so this represents a feedback loop between the ontology and the NLP system.

3.4 Integration into the NLP System

The lexicon resulting from intersecting the GO with the domain corpus is represented in terms of gazetteers (term lists) in the General Architecture for Text Engineering (GATE) framework (`http://gate.ac.uk`). GATE itself only supports the assignment of major and minor types to a given list of lexical items, as shown in Figure 3.2. This alone does not provide sufficient semantic granularity to enable precise relation extraction, and furthermore does not allow us to take advantage of the semantic structure provided by the grounding of the terms in the GO. We therefore incorporate extensions to GATE provided by OntoText Lab (`http://www.ontotext.com`) which allow us to define mappings of ontological categories from GO to lexical features in the GATE lexicon. With this in place, lexical items can be considered by the NLP system in the far richer semantic context provided by the GO.

Acknowledgements: This work was funded in part through a Los Alamos National Laboratory collaboration with Procter & Gamble Corporation.

References

[1] Ashburner, M., Ball, C., and Blake, J. e. (2000). Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25(1):25–29.

[2] Lesk, M. (1986). Automatic Sense Disambiguation using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the Fifth International Conference on Systems Documentation*, pages 24–26. ACM.

[3] Papcun, G., Sentz, K., Fulmer, A., Xu, J., Lubeck, O., and Wolinsky, M. (2003). A construction grammar approach to extracting regulatory relationships from biological literature. In *Pacific Symposium on Biocomputing 2003*.

Chapter 4

Using Ontologies for Text Analysis

Lawrence Hunter and K. Bretonnel Cohen
Center for Computational Pharmacology
School of Medicine
University of Colorado Health Sciences Center

Many approaches to molecular biology text processing have relied on either pattern-matching (e.g., Ono *et al*. 2001, Blaschke and Valencia 2002) or machine learning techniques (e.g., Craven and Kumlein 1999). More linguistically oriented systems often are either essentially purely syntactic (e.g., Park *et al*. 2001, Yakushiji *et al*. 2001) or take the traditional approach of syntactic analysis first, followed by some kind of semantic post-processing (e.g., Rindflesch *et al*. 2000), or combine syntactic and semantic knowledge via semantic grammars and selectional restrictions (e.g., Friedman *et al*. 2001). We hypothesize that an ontology can be the central component of not only semantic filtering, but also of syntactic processing of text. Direct Memory Access parsing (e.g., Riesbeck 1986, Martin (1990), and Fitzgerald (1995)) belongs to a family of "conceptual" approaches to parsing-approaches that involve mapping input texts to a conceptual representation. The basis of the conceptual representation in the DMA implementation that we are using, the Conceptual Memory Parser produced by I/NET, Inc. (`www.inetmi.com`), is a knowledge base consisting of inheritance-based *isA*/*hasKinds* relations, partonymy relations, and various additional attribute/value relations specific to particular portions of the inheritance hierarchy. The attributes of any concept in the ontology can include "phrasal patterns." Phrasal patterns consist of sequences of characters and pointers to objects in the conceptual memory. Architecturally a minor addition to the knowledge representation, they allow mapping between concepts in memory and input text, and are the mechanism for recognition in the input text of concepts that exist in memory. Knowledge-based parsing has sometimes been thought to be impractical due to the cost and difficulty of building the required knowledge resources.

However, the past few years have seen an explosion in publicly available data sources related to molecular biology. Preliminary results support the hypothesis that this publicly available data, unavailable in the early days of natural language processing, today makes knowledge-based parsing practical.

Here we outline our approach, use the Gene Ontology as an example of what can be accomplished with publicly available data sources, point out some problems that come up with using those data sources, and discuss our approaches to solving those problems through the combination of multiple knowledge sources, application of limited linguistic analysis, and empirically derived heuristics. We present results of a pilot project showing that the system can extract gene/disease relations from a corpus of Medline sentences.

A major advantage of the Direct Memory Access approach is the location and delimitation of multi-word terms-it uses the ontology's vocabulary itself to suggest syntactic constituents. The advantage of using lexical resources in this way is demonstrated by the GENIES system (Friedman *et al*. 2001). Consider, for example, the work involved in processing the phrase *regulation of cell migration* (which occurs in many Medline abstracts) through a shallow parser. In a FASTUS-like system, the parser will have to successfully match *cell migration* as a noun group, then match of *cell migration* as a prepositional group, and finally connect that prepositional phrase with *regulation*. In contrast, the DMAP approach recognizes *regulation of cell migration* directly as GO term 0030334, simultaneously positing it as a syntactic constituent and activating its inheritance structure in the ontology. Our system differs from GENIES in that the ontology is elevated from a lexical resource to being the central organizational structure of the system, and that its memory-based, non-procedural construction makes both the lexical resource and the results of the parse available for further inferencing tasks without an intervening database call.

The availability of a large volume of high-quality semantic data in the Gene Ontology makes possible a "semantics first" approach to parsing, but natural language input presents challenges that fall outside of the purview of bio-ontologies. These include entity identification, syntactic complexity, and contextualizing or "meta-science" statements. One sentence can provide examples of all three: *These findings suggest that FAK functions in the regulation of cell migration and cell proliferation*. (Gilmore and Romer 1996)

An entity identification problem is presented by the string FAK in (1). Neither the Gene Ontology, HUGO, nor LocusLink has an entry for this *ad hoc* abbreviation. However, these data sources do contain data sufficient to allow for identification of the concept to which it refers. A molecular biology acronym handler, such as those described in Chang *et al*. 2002, Liu and Friedman 2003, or Schwartz and Hearst 2003, can map the abbreviation FAK to the string *focal adhesion kinase* on the basis of evidence earlier in the document, where the abbreviation is first introduced. Even this name is not official and returns no results at either the LocusLink web site or the HUGO search utility. However, heuristics for mapping non-canonical forms of gene names described in Cohen *et al*. 2002 license the mapping of *focal adhesion kinase* to *focal adhesion kinase 1*, a synonym for the appropriate human entry in LocusLink.

The need for syntactic analysis is presented by the string *regulation of cell migration and cell proliferation*. Simple text matching with no syntactic analysis suggests mapping this string to *regulation of cell migration* (GO:0030334) and *cell proliferation* (GO:0008283). However, a better matching for the second string would be to *regulation of cell proliferation* (GO:0042127).

The substring *cell migration and cell proliferation* has a number of possible parses based on any reasonable grammar of noun phrases, but only the correct one, which posits *cell migration* (GO:0016477) and *cell proliferation* (GO:0008283) as conjuncts yields the proper parse. Furthermore, the constituents of this parse maps to the concepts that would be favored by a longest-match-first constraint, i.e. *regulation of cell migration* (GO:0030334) and *regulation of cell proliferation* (GO:0030334), such as the one used by MetaMap (Aronson (2001)).

The Gene Ontology and LocusLink provide resources that can be exploited by Direct Memory Access Parsing to achieve coverage of much of the content of sentences like (1). However, they leave uncovered the beginning of the sentence, *These findings suggest that*, which provides important context for the factual assertion. The importance of these sort of clauses in natural language processing has been known since at least Harris *et al.* 1989, who called them *meta-science*, and Jackson and Moulinier 2002 discussed handling them in a CYK parser in the context of an information extraction system. Some insight into how to handle them within an ontology-centric model of language processing can be gained from Tanabe 2003, which presents data suggesting that an ontology of these context-establishing statements can be built by leveraging a statistical language model to make use of what we know about the role of frequent constructions in sublanguages.

Previous work in linguistically-informed approaches to processing molecular biology texts with the ARBITER and GENIES systems, as well as the work of Park *et al.* and Yakushiji *et al.*, confirms the utility of various aspects of our approach to applying ontological resources in this domain. Our work explores the consequences of taking the next step: making the ontology the central element of the text analysis system.

References

[1] AkaneYakushiji, Tateisi, Y., and Miyao, Y. (2001). Event extraction from biomedical papers using a full parser. In *Pacific Symposium on Biocomputing 2001*, pages 408–419.

[2] Blaschke, C. and Valencia, A. (2002). The frame-based module of the Suiseki information extraction system. *IEEE Intelligent Systems*, 17:14–20.

[3] Chang, J. T., Schtze, H., and Altman, R. B. (2002). Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*, 9(6):612–620.

[4] Cohen, K. B., Dolbey, A. E., Acquaah-Mensah, G. K., and Hunter, L. (2002). Contrast and variability in gene names. In *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, pages 14–20. Association for Computational Linguistics.

[5] Craven, M. and Kumlein, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, pages 77–86. AAAI Press.

[6] Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(1):74–82.

[7] Gilmore, A. P. and Romer, L. H. (1996). Inhibition of focal adhesion kinase (FAK) signaling in focal adhesions decreases cell motility and proliferation. *Molecular Biology of the Cell*, 7(8):1209–24.

[8] Harris, Z., Gottfried, M., Ryckman, T., Mattick Jr., P., Daladier, A., Harris, T. N., and Harris, S. (1989). *The Form of Information in Science*. Kluwer Academic Publishers.

[9] Jackson, P. and Moulinier, I. (2002). *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*. John Benjamins Publishing Company.

[10] Liu, H. and Friedman, C. (2003). Mining terminological knowledge in large biomedical corpora. In *Pacific Symposium on Biocomputing 2003*, pages 415–426.

[11] Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161.

[12] Park, J. C., Sook Kim, H., and Kim, J. J. (2001). Bidirectional incremental parsing for automatic pathway identification with combinatory categorical grammar. In *Pacific Symposium on Biocomputing 2001*, pages 396–407.

[13] Riesbeck, C. K. (1986). From Conceptual Analyzer to Direct Memory Access Parsing: an overview. In Sharkey, N. E., editor, *Advances In Cognitive Science I*. Ellis Horwood Limited.

[14] Rindflesch, T. C., Rajan, J. V., and Hunter, L. (2000). Extracting molecular binding relationships from biomedical text. In *Proceedings of the ANLP-NAACL 2000*, pages 188–195. Association for Computational Linguistics.

[15] Schwartz, A. S. and Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing 2003*, pages 451–462.

[16] Tanabe, L. (2003). Text Mining The Biomedical Literature for Genetic Knowledge. unpublished doctoral dissertation.

Chapter 5

Half Way Up The Ladder

M.W. Wright, E.A. Bruford, M.J. Lush, R.C. Lovering, V.K. Khodiyar, C.C. Talbot Jr.[2], H.M. Wain, and S. Povey
HUGO Gene Nomenclature Committee (HGNC)
The Galton Laboratory
Department of Biology
University College London
Wolfson House
4, Stephenson Way
London
NW1 2HE.
[2]The Johns Hopkins School of Medicine
Institute of Genetic Medicine
The Johns Hopkins University
Baltimore
MD, 21287
USA.

The discovery of novel genes is still ongoing and with each new gene comes the important decision of what to name it. After many years of hard work attempting to isolate a gene, a researcher can become very attached to it and might have given it a pet name or symbol. However, whilst this symbol may be familiar to the researcher and their lab, it may not make sense to anyone in the scientific community unless it reflects the function of that gene or its relationship to other already known genes. Moreover, if this symbol has already been used for another gene then unnecessary confusion is created.

The HUGO Gene Nomenclature Committee (HGNC) strives to avoid this confusion when talking about genes by agreeing a common nomenclature, so that each gene has its own symbol that everyone can recognise. This ongoing project has to date provided unique gene symbols for over half of the estimated 30,000 human genes.

Promisingly, the need for standardisation is being recognised by the community, as more and more authors publishing in other journals contact us prior to publication because they want their work to be readily accessible via the use of approved symbols in the public databases. To this end we edit our own online nomenclature database, and NCBI's LocusLink, and this information proliferates to many other databases.

We have a system of data exchange with three other databases as well as LocusLink: SWISS-PROT, GDB and MGD. Our strong collaboration with these databases ensures that the data are accurate and can be identified using the approved gene symbol. Other online resources that use approved gene symbols include:

Ensembl `http://ensembl.ebi.ac.uk/`

GENATLAS `http://bisance.citi2.fr/GENATLAS/`

GeneCards `http://bioinformatics.weizmann.ac.il/cards/`

The Genome Database (GDB) `http://gdbwww.gdb.org/`

Genew `http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl`

GeneTestsGeneClinics `http://www.genetests.org/`

Human Gene Mutation Database `http://www.hgmd.org/`

LocusLink `http://www.ncbi.nlm.nih.gov/LocusLink/`

MGD, Mouse Genome Informatics `http://www.informatics.jax.org/`

Online Mendelian Inheritance in Man `http://www.ncbi.nlm.nih.gov/Omim/`

SWISS-PROT `http://www.expasy.ch/sprot/`

UCSC Human Genome Project Working Draft `http://genome.cse.ucsc.edu/`

The quest for a common nomenclature is also furthered by collaboration with the Mouse Genomic Nomenclature Committee (MGNC). Together we try to co-ordinate our efforts so that each human gene is named the same as its counterpart in the mouse. As well as gene symbols we assign approved names and we have been updating some of these to better synchronise with those found in the Mouse Genome Database (MGD); this includes adopting US spelling and the removal of unnecessary punctuation.

Last year HGNC and MGNC published, in the same issue of Genomics, our latest guidelines (Wain *et al.*, 2002, Maltais *et al.*, 2002). The new human gene nomenclature guidelines were the culmination of a long consultative process during which we canvassed the opinions of genetic scientists worldwide, both online via our web site (`http://www.gene.ucl.ac.uk/nomenclature/guidelines/draft_2001.html`) and face to face at side meetings at HGM2001 and ASHG (American Society of Human Genetics) conferences.

Any input on our guidelines, which are regularly updated would be particularly welcome.

We maintain a high profile in the scientific community via our publications (recent examples are listed below) and attendance at international genetics conferences. We normally attend both HGM and ASHG (American Society of Human Genetics), staffing a booth where anyone with a nomenclature query can come and talk to us face to face.

The HGNC team analyses gene data submitted to us by authors and from public databases, operating an unbiased and confidential service to name novel human genes. Contact us via email at `nome@galton.ucl.ac.uk`.

Chapter 6

Large Scale Ontologies for Information Retrieval

Nick Tilford
BioWisdom
Babraham Hall
Babraham
Cambridge
UK
Web site: http://www.biowisdom.com
Email: nick.tilford@biowisdom.com

For an ontology to be used effectively in information retrieval (IR) it needs to be highly granular, comprehensive, validated and accessible from a number of points of view. BioWisdom is producing a methodology to construct large scale evidence-based ontologies.

Ontologies can be utilised at several points in the information retrieval process from query term selection and construction to categorised presentation of search results to facilitate navigation, education and knowledge discovery. Furthermore, an ontology can be used as a framework to "index" or categorise database records or documents from an information source to allow rapid IR and this method can be extended to improve IR precision and recall by using concept context to disambiguate terms. For optimal performance in IR the ontologies must cover all the major concepts in the information source being queried e.g. in the biomedical area the ontologies should encompass domains such as gene products, disease, species, anatomy, biological processes and drugs. The use of ontologies in IR can be enhanced further by the use of linguistic methods such as natural language processing and faceted categorisation. IR can also be extended by using the rich set of relationships to conduct inference to highlight a new logical proposition.

There are two main approaches to ontology construction. A 'top-down' approach where the starting point is a high level model of the domain, followed by sub-classification directed by domain experts. This method relies mainly on manual construction and annotation. Alternatively, a "bottom-up" approach de-constructs databases to a set of assertions that are automatically parsed into a conceptual framework. BioWisdom has brought these two approaches together to construct a large repository of evidence-based assertions that are semantically organised with a high-level "upper" ontology. Once the system is established, it is possible to inferentially derive assertions, either by probabilistic, heuristic or rule-based methods, from data sources where the assertion is not explicitly stated.

BioWisdom focuses on providing ontologies to support the drug discovery process and has based its "upper" ontology on this process. The process is broken down into its essential elements and the inter-relationships between these elements. For example, "drug targets" are sub-classified into proteins and genes, and these can be directly associated with disease and tissues. The upper ontology consists of high-level concepts with a series of properties that can take the form of relationships between concepts. The domains e.g. disease, symptom, protein, drug identified in the upper ontology direct the collection of assertions. The relationships between concepts define the type of reasoned associations that can exist between domains e.g. drug *treats* disease.

Taxonomies are good source of assertions providing established "is-a" or "is-part-of" relationships between concepts e.g. MeSH, ICD10. However, taxonomies usually do not provide a wealth of other properties. When a database is identified as a potential source of assertions, it is analysed so that the principle components of each record are mapped to concept types in the upper ontology. For example, LocusLink is a database of genes and so the initial assertion that can be captured from each record is that the record unique identifier, LocusID, relates to gene. Each record has a source organism so LocusID can be related to species. Using this method, at least 8 logical assertions can be produced from one LocusLink record. Each assertion extracted from the record will have a logically named relation relating two elements of the record e.g. *has-source-organism*, *has-cytogenetic-location*. This is important to precisely represent the relationship between concepts which assists understanding and the application of inference. The database record acts as the evidence for each assertion. Using databases such as Locuslink (200,000 records) and Swissprot (100,000 records) the number of usable assertions in the repository will number over 3 million. Using a Genbank (24 million records) the size of the repository grows to over 120 million assertions. When several data sources are used several assertions can be linked to associate concepts through a network of relationships.

Some data sources e.g. Medline or annotation fields in databases need to be processed to extract meaningful assertions. When the assertion is embedded in a block of text, information extraction techniques need to be employed to identify concepts and relationships. This entails the identification of noun and verb phrases within the text. The noun phrases relate to concepts and can further discriminated as particular concept types e.g. protein, tissue etc. Verb phrases encapsulate the relationship between concepts e.g. "is expressed in". As this method relies on a series of assumptions, the validity of an assertion produced by this method is lower than a fact derived from a database record.

The validity of assertions can be scored, based on the method by which the assertion was retrieved, to assist their use in information retrieval and related processes including inference and categorisation.

Different data sources use different terms to denote the same entity and these terms need to be normalised to a single entity. This process of semantic normalisation of terms occurs once assertions from several data sources are collected. This process takes a term from an assertion and maps it to a core concept with an equivalent meaning. The terms linked to a concept in this way can be viewed as synonyms or aliases of the concept. Using normalisation several base assertions will be shown to be semantically indistinct i.e. an association between two concepts will be supported by several threads of evidence.

BioWisdom is developing a system that can manage a high volume of evidence-based assertions mapped to a single logical framework to produce a large scale ontology. This broad and granular ontology is highly valuable when applied to information retrieval applications.

Chapter 7

Application of Gene Ontology in Bio-data Warehouse

Shunliang Cao[1,2,*], LeiQin[3] , WeiWang[1], Yangyong Zhu[1], and YiXue Li[3]
[1]Department of Computing and Information Technology, Fudan University, Shanghai, 200433,China
[2]Department of Computing and Information Technology, Ningbo University, Ningbo, 315211,China
[3]Shanghai Center for Bioinformatics Technology, Shanghai, 201203, China
[*]Corresponding authors. E-mail: caoshunliang@163.com

The development of bioinformatics seriously depends on the accumulation of tremendous amounts of bio-data from various databases. The integration of biology databases is critically important because of the interconnectedness of biological research. The integration of genome, proteome research information in time and the maturation of bio-data warehouse compose the key factors of the bioinformatics technology. To technically interpret and physically migrate heterogeneous biological databases into a single consistent database and provide a united GUI (graphical user interfaces) is the way to achieve bio-information integration, intellectualized multiple, complex, intercross search and data sharing based on high performance computing platforms. It will support modern biology research greatly,such as functional genome research and drug discovery. As a main method of data integration, bio-data warehouse is a central repository, which extracts bio-information from various biological resources. It calls for data integration from multiple sources into a coherent form that is the base for further analysis (e.g., data mining). However, data sets that are interesting for computational biologists are often heterogeneous in structure, content, and semantics. Problems that might arise due to heterogeneity of the data have already been well known within the distributed database systems community: structure heterogeneity and semantic heterogeneity. Structural heterogeneity means that different information systems store their data in different structures. Semantic heterogeneity considers the content of an information item and its intended meaning. In order to achieve semantic heterogeneity in a heterogeneous information system, the meaning of the information that is interchanged has to be understood across the systems.

Using ontology for the explication of implicit and hidden knowledge is a possible approach to overcome the problem of semantic heterogeneity. Ontology provides a mechanism for capturing a community's view of a domain in a shareable form, which is both accessible by humans and computationally amenable. Ontology provides a set of vocabulary terms that label concepts in the domain. By capturing knowledge about a domain in a shareable and computationally accessible and computable semantics about the domain knowledge they describe, bio-ontology provides a model of biological concepts that can be used to form a semantic framework for many data storage, retrieval and analysis tasks. A lot of ontology systems have been developed in the field of bioinformatics. Currently the most important ontology in the bioinformatics community is the Gene Ontology. Gene Ontology, as a tool for the unification of information about gene products, aims at producing a dynamic controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells. It comprises three orthogonal taxonomies or aspects, which hold terms that describe the attributes of molecular function, biological process and cellular component for a gene product.

To address these problems we have developed a bio-data warehouse named BioDW, in which a central database is used to uniformly collect and store biology data from heterogeneous schemas/formats and scattered over numerous (public) repositories. Moreover, it aims at a semantic integration of annotations using Gene Ontology. Now we have integrated DNA databases (such as Genbank), protein databases (such as swissprot) and functional related databases(such as KEGG) into BioDW.

As illustrated in Figure 7, BioDW is build with GO. The advantages of integrating GO into BioDW can be described as follows:

- First, we download Gene Ontology tables and their data to construct a local GO system in our Bio-data warehouse. Based on this, the gene products listed under special GOid can be out linked to the corresponding entries stored in the data warehouse.

- Second, because entries in swissprot have already been well annotated with GO terms. Based on the dbxref table of swissprot, the corresponding DNA entries from Genbank can also be annotated with the same GO terms. In KEGG, KO database also integrated GO terms in its entries too. So, entries in the data warehouse are assigned with GO terms to the maximum extension.

- Third, terms and definitions in GO system have been well organized logically. This precise logical relationship can be used in the function of semantical search. The output of query would be entries with similarity to the query terms listed according to the homology while the unrelated search results have been excluded as much as possible, which has distinct advantage over simple text search method.

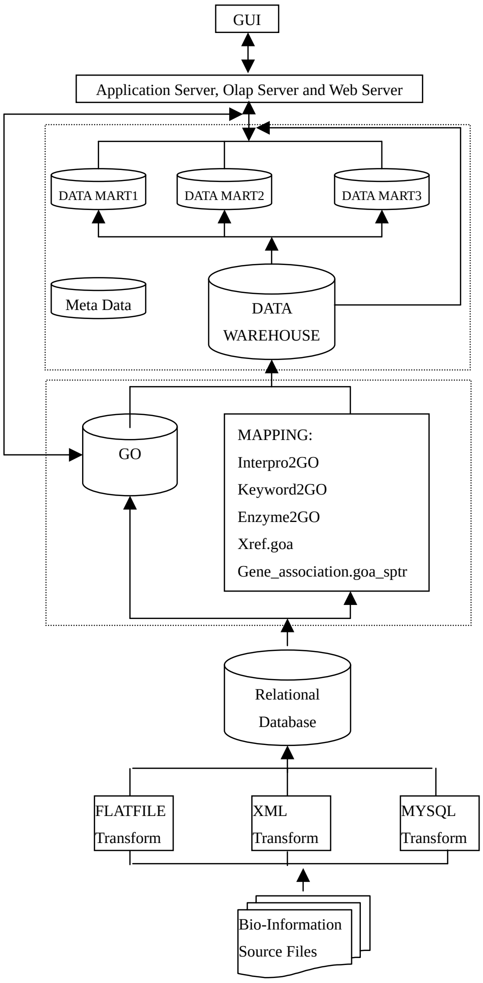

Figure 7.1: Architecture of BioDW.

- Fourth, Gene Ontology can also be a powerful tool on data marts constructing and going for data cleaning (eliminating data redundancy).

Data mart is a special database constructed to satisfy users with specified requirment. Entries in a data mart should also be highly related to each other in a specialized area, in special organisms, in special processes or with the similar function, etc. These relationships are also well represented in GO systems. With the help of GO, entries fitting the criteria of data mart can be easily extracted from data warehouse and arrayed in well order.

Eliminating data redundancy is hard work and costs a lot of time and money. With the help of GO and interrelationships among GO terms, possible redundant entries in the data warehouse will be clustered for further analysis. This process saving much analysis work used to be done by biologists.

Acknowledgements: This project is imbursed by the Hi-Tech Research and Development Program (“863” program) of China (under Grant No: 2002AA231011) and the Major Project of Shanghai Community of Science & Technology (under Grant No :02DJ14013).

Chapter 8

BioPAX – Data Exchange Ontology for Biological Pathway Databases

BioPAX Group
http://www.BioPAX.org

8.1 Introduction

With the completion of many large-scale genome-sequencing projects, efforts to annotate the gene products are generating increased interest in the biological community. These efforts have produced databases that capture functional information about gene products. Many of these databases can be categorized as "pathway" databases because the data they capture is evocative, either in whole or in part, of the biological notion of a pathway.

Traditionally, pathway databases have been populated through manual means. This typically involves one or more curators who review the scientific literature for experimental results. The curators then translate the conclusions drawn from those results and store the information in a database entry. This is a long and tedious process that accumulates pathway data at a slow rate.

Recently, however, a number of methods that expedite the process of pathway database population have been developed. For example, several tools now exist that use Natural Language Processing (NLP) to automatically generate pathway data from scientific publications. In addition, pathway data from high-throughput experimental methods (such as for detection of protein-protein interactions) are easy to translate into databases. There are also computational methods that generate pathway data by inferring new relationships from existing datasets.

These new methods have increased the size and number of existing databases. In fact, the rate of new pathway database creation appears to be increasing.

The value of these new databases to the research community would be much greater if the data were easily accessible in a standard format.

8.2 Objectives and Scope

The main objective of the BioPAX initiative is to develop a data exchange format for biological pathways that is flexible, extensible, optionally encapsulated and compatible with other standards and can be widely adopted in a timely manner.

Our definition of a biological pathway is "a network of biological relationships". This general definition encompasses metabolic pathways, signal transduction pathways, gene regulatory pathways, genetic interaction pathways, networks of word relationships found using text mining, *in silico* predicted pathways and pathways of cellular interactions. It does not cover pathways in biological systems that are higher order than the cell, such as physiological pathways. Thus the scope of BioPAX is defined to be all pathways relating to cellular and molecular biology. Initially, however, BioPAX will focus on metabolic, signal transduction and gene regulatory pathways, as most existing data fall into one of these three categories.

8.3 Use Cases

The primary use case of the BioPAX data exchange format will be to facilitate data sharing between pathway databases like aMAZE, BIND, DIP, EcoCyc, IntAct, and WIT. BioPAX could also facilitate the creation of a centralized public repository for pathway data (in fact, the desire for such a repository was one of the driving forces behind formation of the BioPAX effort).

Another intended use of the BioPAX format is to provide a standard format for software tools that must access pathway data. A few examples of such tools include pathway visualization programs, expression analysis tools, and pathway simulation tools.

It is important to note that although the systems biology mark-up languages, such as CellML (`http:www.cellml.org`) and SBML (`http://www.sbml.org`), already address the pathway simulation use case, they are not designed to handle the wider range of use cases intended for BioPAX. Still, BioPAX is being designed to be compatible with these and other important standards, such as PSI (`http://psidev.sourceforge.net`), in the areas where they overlap. This should decrease the burden on end-users of having yet another standard and allow for the possibility of a future common standard.

8.4 Ontology Development

BioPAX is being developed as an OWL ontology using the Generic Knowledge Base (GKB, `http://www.ai.sri.com/~gkb/`) Editor (developed by Peter Karp's group) and concurrently as an XML Schema using XMLSpy. We are trying to design both so that XSLT translators can flawlessly translate between them.

The rationale behind dual syntaxes for the exchange format stems from the fact that OWL is a more powerful representation language, but XML Schema is currently in much wider use. It is impossible to predict whether the advantages OWL provides will result in it becoming commonly used. Since the utility of a data exchange format depends on it being widely accepted, we felt it was necessary to create a version of BioPAX in both of these syntaxes.

The first drafts of both the OWL and XML Schema versions of BioPAX are planned for release in June 2003.

8.5 Summary of Progress to Date

The first milestone of the BioPAX group was the decision to use Chemical Markup Language (CML) for the representation of small molecules and a subsequent proof-of-concept that involved sending a set of small molecules from EcoCyc to a visualization program used by the Shah lab. Another important milestone was the completion of an informal description of the PAX framework. The framework (now called a PAX record) can represent both metabolic pathways and signal transduction pathways.

We have established contact with the SBML, CellML, and PSI groups and have agreed to work together to ensure compatibility. Other accomplishments include the evaluation of OWL as a syntax language and the creation of several subgroups. These subgroups address specific issues, such as the best way to represent molecular states.

8.6 History of the BioPAX Initiative

The BioPAX initiative grew out of open discussions about sharing pathway database information organized by Chris Sander at ISMB 2001 in Copenhagen and again at the BioPathways Consortium satellite meeting at ISMB 2002 in Edmonton, Canada.

The core work group was formed in October 2003 and is composed of a representative mix of end-users, database developers, and software developers, and includes several BioPathways Consortium (BPC) representatives. To ensure adequate representation of the biological pathways community, BioPAX core members actively reach out to community members for feedback. In addition, BioPAX forms subgroups that include additional community members. The core group holds biweekly conference calls and face- to-face meetings monthly to bi-monthly at rotating locations in the U.S. Minutes and presentations are posted on at `http://www.BioPAX.org`.

8.7 How to Contact BioPAX

Members of the community interested in BioPAX are encouraged to participate, promote and provide feedback to BioPAX. The BioPAX web site (`http://www.BioPAX.org`) contains documentation, activities, and mailing lists for feedback and discussion. Participation in the BioPAX core group currently requires an invitation as we try to keep the core group small and efficient, however please contact us for participation in subgroups. Subgroups are formed to extend the BioPAX specification and address specific development issues. We currently have *Small Molecule*, *Molecular State* and *Pathway Data Examples* subgroups.

Chapter 9

Structural Classification in the Gene Ontology

Cliff Joslyn[1], Susan Mniszewski[2], Andy Fulmer[3], and Gary Heaton[4]
[1] Knowledge Systems and Computational Biology Team;
Modeling, Algorithms, and Informatics Group (CCS-3);
Mail Stop B265, Los Alamos National Laboratory,
Los Alamos, NM 87545, USA,
joslyn@lanl.gov,
http://www.c3.lanl.gov/~joslyn,
[2] Quantum and Classical Computing; Modeling, Algorithms, and Informatics Group (CCS-3);
Mail Stop B265, Los Alamos National Laboratory,
Los Alamos, NM 87545, USA,
smm@lanl.gov,
(505) 667-0790
[3] Biotechnology R&D, Procter & Gamble Corporation,
fulmer.aw@pg.com
[4] Bioinformatics IT, Procter & Gamble Corporation,
heaton.gg@pg.com

While ontologies are becoming important in all the knowledge-based sciences, their penetration into computational biology is deeper than most others, and, moreover, becoming more necessary to the normal course of scientific development. Use of ontological structures such as the Gene Ontology (GO) (1; Gene Ontology) are increasingly a standard part of a typical biologist's work day.

We have been pursuing work in structural classification of the GO. That is, given a list of genes of interest, how are they organized with respect to the GO? Are they centralized, dispersed, grouped in one or more clusters? With respect to the biological functions which make up the GO, do the genes represent a collection of more general or more specific functions, a coherent collection of functions or distinct functions?

Existing approaches to these questions (3; 5) have relied on the statistics of how ontology nodes are generally populated, and/or use a distance based on the minimal path length between two nodes (4). Our approach (2) is based on the following principles:

- While we welcome the use of node statistics as supplementary information, it is also important to be able to provide answers when such information is lacking, that is, based only on the *structural* relations among the nodes in the ontology implicated by the genes of interest.
- It's also important to complement our intuitive approaches to methodological development with sound mathematical reasoning. Ontologies such as the GO share with object-oriented hierarchies a common mathematical structure called a **labeled poset**: a collection of partially ordered sets (posets, equivalent to Directed Acyclic Graphs (DAGs)), each one representing a different semantic category. In the case of the GO, and in many other ontological structures as well, there are two posets, composed of `is-a` and `has-part` links respectively. Compared to more familiar mathematical structures such as trees, lattices, networks, or Euclidean spaces, we have fewer good intuitions and techniques about posets. Therefore it's important to base methods in the mathematics of posets specifically. For example, minimal path length among any two nodes is a network-theoretical approach, and really makes no sense in a poset.

In this talk we will discuss our approach to structural classification in the GO based on **pseudo-distances in posets**. Our system, the Gene Ontology Clusterer (GOC), uses pseudo-distances between comparable nodes only, in conjunction with scoring algorithms, to rank-order the GO nodes with respect to the requested genes. By iterating this process, we support biologists in the process of knowledge discovery within the GO.

Aside from describing our system in general, our primary purpose in the talk will be to discuss the lessons we've thereby learned about working with the GO, in particular the following kinds of issues:

- Our experience in trying to identify categories in the GO based on lists of genes of interest has led us to appreciate two interrelated concepts: **coverage** is the idea that a given node should cover as many of the genes of interest as possible, while **specificity** is the idea that it should do so as precisely or specifically as possible. These ideas are conflicting: the top of the ontology always provides complete coverage but minimal specificity, while identifying any individual node containing a gene of interest provides complete specificity with minimal coverage. Identification of clusters is thus not unequivocal, but rather a user-dependent judgment about the tradeoff of specificity and coverage.
- The GO is widely and legitimately championed as being superior to other systems in that it is DAG-based, and not a tree. But the

consequences of this are not nearly as widely recognized. In particular, our intuitions about concepts of how "levels" work, and the relations among nodes at different levels, can be quite deceptive; and statistical approaches which are quite clean in trees can become non-additive with DAGs.

- Moreover, tree-based software dominates GO interfaces. Visualization of DAGs is especially important, for example to interpret the outputs of our GOC system.
- There is a need for a broader mathematical and statistical analytical base of understanding of the GO. The kinds of questions we have, and perhaps will address, or perhaps others have already addressed, include: the distributions of leaves and roots; the distribution of up-branching and down-branching; overall path length statistics; distribution of genes, both through the GO and with respect to multiple genes per node; and areas of tree-ish and lattice-like regions within the full GO.

References

[1] Ashburner, M., Ball, C., and Blake, J. e. (2000). Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25(1):25–29.

[Gene Ontology] Gene Ontology. `http://www.geneontology.org`.

[2] Joslyn, C., Mniszewski, S., Fulmer, X., and *et al* (2003). logical Spaces of Biological Function. In *Pacific Symposium on Biocomputing PSB 03*. `ftp://ftp.c3.lanl.gov/pub/users/joslyn/psb03f.pdf`.

[3] Lord, P., Stevens, R., Brass, A., and *et al* (2002). Semantic Similarity Measures Across the Gene Ontology: Relating Sequence to Annotation. Fifth Annual Bio-Ontologies Meeting, Satellite SIG of Proc. Intelligent Systems for Molecular Biology (ISMB 02).

[4] Rada, R., Mili, H., Bicknell, E., and *et al* (1989). Development and Application of a Metric on Semantic Nets. *IEEE Trans. on Systems, Man and Cybernetics*, 19(1):17–30.

[5] Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems in Ambiguity in Natural Language. *J. Artificial Intelligence Research*, 11:95–130.

Chapter 10

A Report on the Sequence Ontology

Suzanna E. Lewis, Karen Eilbeck, Michael Ashburner, Judith Blake, Michele Clamp, Richard Durbin, Lincoln Stein, Colin Wiel, Mark Yandell, and Christopher J. Mungall

The Sequence Ontology is a structured controlled vocabulary that provides a lexicon to describe details of biological sequences. It is provided as a common resource for the bioinformatics community. As multiple genome annotation groups converge on shared semantics to describe their primary annotations for sequence data, the bioinformatics community will benefit in many obvious ways:

- With a shared agreement on how annotations are described, the software that parses this data will become vastly simpler to write and maintain. Developers will no longer need to write special software versions for each source of annotations, *e.g.* for data collected from different DAS servers.
- With a shared agreement on how annotations are described, the annotations will be easier to query and the results that are returned will represent the same type of annotation regardless of the data source. When sequences within model organism databases are annotated with these terms, it will be possible to robustly query all these databases asking, for example, to see all gene sequences with edited transcripts, or trans-spliced, or which are bound by a particular protein.
- With a hierarchical set of terms describing sequence annotations, users can query at many levels of granularity using either broadly and narrowly defined terms to describe a search, for example, to retrieve annotations of sequence variations. The annotations will be recorded at the appropriate level, and the query can be constructed at a different level.

Within the Gene Ontology (GO) each term is an atomic (indivisible) unit, *i.e.* the curator is not able to dynamically construct their own phrase by combining existing words or qualifiers. As a consequence of this constraint, indivisible phrases have been introduced into GO as a temporary expedient to provide the expressiveness that is needed for curation. If we continue introducing complex compound terms the system will become progressively more redundant, less flexible and increasingly difficult to manage. This is a recognized concern that is being addressed for the future of the GO. The Sequence Ontology, offers the opportunity to implement and utilize a more flexible and dynamic methodology from the outset. Our first move is to implement attributes or *slots* for terms. "Slots" provide a formalism for dealing with the classification of finely granular terms by allowing the flexible creation of phrases. Another way to look at this is as a progression beyond a vocabulary of fixed phrases to a grammar for the creation of phrases that will offer the biological curators a *structured* means of composing phrases to use during annotation. The development strategies for the SO will reflect back to aid the evolution of the GO.

This talk will report on our current progress to this end and discuss the issues and choices that we have made. We will also review some of the semantic issues that have generated discussion within the consortium.

Chapter 11

Ontologies for the Physiome Project

Poul Nielsen, Matt Halstead, and Peter Hunter
Bioengineering Institute
University of Auckland
New Zealand
`http://www.bioeng.auckland.ac.nz/physiome/physiome.php`
`p.nielsen@auckland.ac.nz`

The Physiome Project is an attempt to build a comprehensive framework for computational modelling of human biochemistry, biophysics, and anatomy. The goal of this project, sponsored by the *International Union of Physiological Sciences* (IUPS) and the *IEEE Engineering in Medicine and Biology Society* (EMBS), is to use computational modelling to analyze integrative physiological function in terms of underlying biological structure and processes. Web-accessible databases of model-related data at the organ system, organ, tissue, and cellular levels have been established to support the project. These databases currently include quantitative descriptions of anatomy, mathematical characterisations of physiological processes, and associated bibliographic information.

CellML and CellML Metadata are XML-based languages used to describe the underlying mathematics and topology of a wide variety of biological models. CellML characterises the structure and underlying mathematics of the model, while CellML Metadata provides supporting information about the scope and context of the model. The structure of CellML is simple, yet powerful enough to provide a common basis for describing a wide variety of model types. It is currently being used to define electrophysiological, mechanical, signal transduction and metabolic pathway models in a publicly accessible database of over 140 published models.

CellML models are constructed as a network of interconnected components. Components, which may represent physical compartments, collections of related entities, or a convenient modeling abstraction, are the basic functional units of a CellML model. Each component may contain variables, the mathematics that describes the relationships between those variables, and metadata. For instance, a pathway model might be organized into components that represent the various species and the reactions they participate in. The components may be grouped logically or physically, allowing the encapsulation of model functionality and the specification of geometric relationships.

CellML models are able to import other models, enabling model hierarchies to be constructed. In this case the basic components themselves become models with a supermodel importing and connecting all the component-models together to form a more complex system. This mechanism encourages component re-use because models containing representations of biological entities that are common to other models may be incorporated into many different supermodels. This mechanism for reuse can thus be used to create libraries of components and units.

Knowledge implicitly associated with a model, however, is not normally included in the CellML representation. In order to address this problem facilities to include ontologies have been added to CellML. An ontology is, in essence, a controlled vocabulary of terms that are related to one another in class hierarchies and are associated by a set of rules. Ontologies are powerful because they give computer applications the ability to infer meaning about a particular set of data based on how the data set associates with the ontology. Ontologies extend the current capabilities of CellML by adding classing mechanisms to CellML components and variables.

We are exploring how CellML may benefit from the incorporation of ontologies by defining the base CellML language, the reaction subset of the CellML language, and a conceptual rendering of a reaction as ontologies with rules about how they interact. A biochemical reaction is broken down into its participants and the expressions that relate the participants. These three branches of the ontology are part of a wider effort to build a formal knowledge representation of the physiome, with all entrants into the ontology being peer reviewed. The ontologies are defined using systems based on the Open Knowledge Base Connectivity (OKBC) protocol and exported in a variety of standard formats, such as DAML+OIL and the W3C's Web Ontology Language, OWL (`http://www.w3c.org/2001/sw/WebOnt/`).

The benefits of using the CellML ontologies are numerous. The reaction ontologies serve as an interface between the scientist and the programmer by allowing the scientist to describe reaction pathways in a way that is biologically familiar and by breaking down the components of a reaction in a way that is conceptually significant and easy for the programmer to implement. For instance, the biologist can describe an enzyme-catalysed reaction with competitive inhibition using a pathway editor by creating an instance of the *competitive inhibition* class, a subclass of the *Michaelis-Menten* class. Because the *competitive inhibition* class is part of the reaction ontology, the editor knows that the reaction involves a substrate, enzyme, product, and inhibitor, and certain other parameters must be entered before the component is complete. What differentiates this methodology from other existing software is that the ontology is not application-specific. The same ontology may be shared and processed by many applications as long as the program can understand standard ontology representations. Furthermore, once an application is capable of processing ontologies, users may define and integrate their own ontologies for use by the program, or incorporate a number of existing ontologies.

The current ontologies created for use with CellML are both powerful and versatile. In the future further ontologies will be constructed to enable graphical information to be assigned to a component, provide better model validation techniques, and associate a model with other models or templates. For updates on how ontologies are being incorporated into CellML, see `http://www.cellml.org/`.

Chapter 12

Disease Ontology: Structuring Medical Billing Codes for Medical Record Mining and Disease Gene Association

Patricia Dyck and Rex Chisholm
Center for Genetic Medicine
Northwestern University

The Disease Ontology is an ontology defining subsumptive relationships in human disease. The goal behind the disease ontology is to create a comprehensive hierarchical vocabulary to represent all disease states. The terms in the ontology were originally based on and are mapped to ICD9 codes in order to facilitate medical record mining. In a manner similar to the Gene Ontology process of curation and public forum, the ontology will be continually extended and revised in order to broadly encompass disease phenotypes. Future plans include mapping these terms to other medical billing code systems. The structure of the ontology is strongly connected to anatomy and cell type, recapitulating gene expression. Disease terms within the ontology are linked by associations to gene products whose aberrant expression or allelic variation causes or contributes to disease. As these gene products additionally have Gene Ontology terms, one may conversely look at disease genes by molecular rather than anatomical groupings.

The ontology is being developed as part of the NUGene project at Northwestern University. Similar to BioBank (UK), The NUGene project collects DNA samples from participants, through an informed consent process, in conjunction with access to their medical history in the form of billing codes. The ontology was initially created in order to impart a logical hierarchical order to the billing terms and eradicate redundancies for data mining purposes. Gene products may have many disease associations such as in the case of a plieotropic Mendelian disorder or have contribution to a single multifactorial disease. These associations have evidence codes similar to those of the Gene Ontology.

This ontology provides many uses for the discovery of new biological knowledge. The subsumptive structure of this ontology favors data mining approaches over an automated reasoning approach because of the statistical basis of medical record mining. This basis was a key factor in designing the hierarchical structure and relationship weakness of the ontology. From the medical records, statistical associations between disparate disease occurrences may be mined. Overlap in biological processes or molecular function may also be mined or inferred from the GO associations of gene products known to cause particular diseases. This will lead to better identification of candidate disease genes and enable us to identify new candidate genes through co-operation in processes or similarity in function.

In it's current incarnation, the ontology holds approximately 6800 terms. In addition to the expansion by adding new terms and mapping to different billing codes, there will be continual restructuring of the ontology to reflect logic, eradicate redundancy and unnecessary groupings. The ontology will be available on Source Forge. Future plans include the release of the ontology, it's gene products associations, and their respective GO associations to the public through the AmiGO interface as a public tool for disease gene browsing.

Chapter 13

Integration of Ontology Data onto the Rat Genome Database (RGD) System

Norberto B. de la Cruz, Simon Twigger, Jedidiah Mathis, and Peter J. Tonellato
Bioinformatics Research Center
Medical College of Wisconsin
8701 Watertown Plank Rd.
Milwaukee
Wisconsin
53226
USA
www.rgd.mcw.edu noriede@mcw.edu

The Rat Genome Database (RGD) is an NIH-funded project whose stated mission is "to collect, consolidate and integrate data generated from ongoing rat genetics and genomic research efforts and make these data widely available to the scientific community". In a collaboration between the Bioinformatics Research Center (BRC) at the Medical College of Wisconsin, the Jackson Laboratory and the National Center for Biotechnology Information, RGD was created to meet these stated aims. The rat is uniquely suited to its role as a model of human disease and the primary focus of RGD is to aid researchers in their study of the rat and in applying their results to studies in a wider context. In support of this we have integrated a large amount of rat genetic and genomic resources in RGD and these are constantly being expanded through ongoing literature and bulk dataset curation. The current version of RGD, includes curated data on rat genes, sequences, quantitative trail loci (QTL), microsatellite markers and rat strains used in genetic and genomic research, plus a variety of tools for mapping comparative analysis.

To extend the usability of RGD, new approaches as to how the information in the database is regrouped, cataloged and presented have to be developed. One of the newest approaches to creating another angle toward knowledge representation is the development of ontologies that classify related concepts within hierarchies. We decided to use the ontologies as a means to present and view data in the RGD website. In light of its maturity and current use by RGD, the initial ontology to be implemented will be the Gene Ontology. Later on, given that work on phenotype and disease ontologies is under way at RGD and other institutions, we will look to incorporate these new vocabularies over the coming year as well.

Implementation of any feature onto a website requires a careful analysis of the audience of the site, along with other stakeholders, and available resources and site operations so as to provide the most effective use of the its features and tools. For this project we analyzed the need of the stakeholders from each of the operations needed to keep the site functional, current and useful. The site operations were identified as follows:

- **Data Presentation** - The Rat Genome Database's user interface is a website and thus ease of use is vital. It is important that users are able to view information readily without sacrificing scientific rigor in the site's organization.
- **Data Curation** - Keeping the data current is one of the important facets of RGD's operation. Proper and ready access to the ontology and annotated information by curators will help ensure that the data presented is as current and as accurate as possible.
- **Data Upload** - Operating a large scale database requires handling copious amounts of for upload and integration. Addition of ontologies into the database requires revision of the process to ensure data quality and to ensure referential integrity.
- **Enhancements** - Biology and bioinformatics are constantly evolving fields and keeping RGD current necessitates enhancement to the site to reflect such developments. For this effort alone, planning to incorporate upcoming ontologies requires us to implement schemas and scripts that allow us to readily upload and integrate them as they come.

The needs of the various stakeholders from each operation, if applicable were determined through interviews. The stakeholders that were identified were:

- **Site visitors** - The site us primarily aimed at the scientific community and the rat, genome and genetics researchers in particular.
- **Development Personnel** - People primarily responsible for upgrading the site and implementing the necessary enhancements.
- **Curation Staff** - Personnel responsible for gathering data for upload into the RGD site. User friendliness and scientific
- Other BRC teams - In addition to RGD, BRC engages in other activities related to processing, consolidating, and integrating genetic and

genomic data. Allowing those teams to readily access, download, and use RGD data is an important facet in RGD's operations.

- **Technical Personnel** - The incorporation of ontologies needs to be consistent with the current scheme for uploading and storing RGD data.
- **BRC Administration** - Website features, as envisioned by the BRC administration need to be considered for current and future implementations.

After categorizing the needs of the various stakeholders. We decided on a course as to how best to integrate ontologies into RGD. We needed to ensure that the implementation is user friendly as well as scientifically sound. Following these considerations, a use case diagram to lay out the function of the ontology information was created. The use case diagram envisions the ontology system serving as an engine by which RGD provides new angles for viewing information and extracting knowledge. The ontology system will provide functionalities to three different ways of accessing information in RGD. For the general search, ontologies can be used as controlled vocabularies from which keywords can be culled. Moreover, objects annotated to specific ontology terms and its descendants can be also be retrieved through this method. The second search method is through an ontology browser that allows users to navigate the vocabularies and obtain term details as well as lists of its annotated objects. In the third search method, ontology terms can be used to focus object searches to specific concepts. All these search strategies aim to provide links to information on the appropriate objects and their instances in the form of reports. Upon reaching the reports, links to information on related objects allow users to explore its biology and navigate further into the site. Future implementations will also integrate ontology functionalities into RGD tools like VCMap and the Genome Browser. Subsequently, current and future state diagrams of the RGD Website were drawn out to provide a concrete illustration for the use the ontology data.

The development of ontologies gives model organism databases a powerful tool for the classification, organization and presentation of knowledge. Following MGI's lead and adopting their technologies, RGD will implement a system that allows any ontology to be fielded. The strategy is to ensure that the schema, upload system, and dynamic web pages are abstract enough so that any ontology using the DAG structure can be accommodated and fielded with relative ease.