Linking Ontological Resources Using Aggregatable Substance Identifiers to Organize Extracted Relations

B. Marshall, H. Su, D. McDonald, and H. Chen

Pacific Symposium on Biocomputing 10:162-173(2005)

LINKING ONTOLOGICAL RESOURCES USING AGGREGATABLE SUBSTANCE IDENTIFIERS TO ORGANIZE EXTRACTED RELATIONS

BYRON MARSHALL, HUA SU, DANIEL MCDONALD, HSINCHUN CHEN

Email: {byronm,hsu,dmm,hchen}@eller.arizona.edu
MIS Department, University of Arizona, McClelland Hall 430 1130 East Helen Street
Tucson, Arizona 85721, USA

Systems that extract biological regulatory pathway relations from free-text sources are intended to help researchers leverage vast and growing collections of research literature. Several systems to extract such relations have been developed but little work has focused on how those relations can be usefully organized (aggregated) to support visualization systems or analysis algorithms. Ontological resources that enumerate name strings for different types of biomedical objects should play a key role in the organization process. In this paper we delineate five potentially useful levels of relational granularity and propose the use of aggregatable substance identifiers to help reduce lexical ambiguity. An aggregatable substance identifier applies to a gene and its products. We merged 4 extensive lexicons and compared the extracted strings to the text of five million MEDLINE abstracts. We report on the ambiguity within and between name strings and common English words. Our results show an 89% reduction in ambiguity for the extracted human substance name strings when using an aggregatable substance approach.

1 Introduction

In the past few years several systems that extract biological relations from biomedical texts have been created. These systems are intended to help researchers leverage vast and growing collections of research literature. Extraction results are usually evaluated for accuracy but this is only one of several important considerations. Consumers of the extracted information (both humans and analysis algorithms) benefit when information is meaningfully organized. That is, (a) multiple references to the same biological substance or process are indexed, (b) substance references are marked so that they can be correctly associated with existing databases and resources, and (c) the context and granularity of the information is specified. These issues have not been ignored by other bioinformatics researchers. They are reflected in the development of semantic classes and ontologies [1], the use of existing lexicons for entity identification [2], the inclusion of context-related information in semantic frames [3], and the representation of extracted information in both binary and nested relations [4]. Still, techniques for aggregating and associating extracted information deserve additional investigation.

Creating a system to process biomedical texts and produce a useful network of relational information is a multi-faceted task [5]. As part of the GeneScene project (http://ai.bpa.arizona.edu) researchers have developed text mining tools to

automatically extract regulatory pathway relations from MEDLINE abstracts [6, 7]. The extracted information is used to create network visualizations and to support various data mining efforts. The work reported in this paper is part of an ongoing effort to improve the usefulness of extracted relations but the resulting aggregation methodologies are potentially interesting for the output of other information extraction systems as well. In particular, we evaluate an extensive aggregatable substance lexicon. Considerable attention is paid to lexical ambiguity which would hinder the matching of relation extraction system output to specific elements in existing lists of biological substances.

2 Relation Output Formats

The design of a relation aggregation system depends on the format of extracted relations. While relation extraction systems vary significantly in their output formats, labeled relational triples are frequently produced by extraction systems and are useful in analysis algorithms [8]. Output formats include predicate relations like those captured by GENIES and EDGAR [4, 9], inhibition relation pairs [10], labeled relational triples with negation extracted by the GeneScene Parser and the Arizona Relation Parser [6, 7], and binary object pairs with categorized relations in [11]. Binary biomedical pathway relations are frequently displayed as conceptual graphs and can be used with many analysis algorithms. More complex nested predicate relations can be stored as a knowledge base and queried but may be more difficult to visualize unless they are first reduced to a binary format. The balance of Section 2 describes a few of these systems.

Initial implementations of the GENIES system extracted simple binary relations such as *X activates Y*. Later the system was expanded to handle more complex relations with the output expressed as nested predicates. This format is precise and can express much of the relational information contained in a text. Relations extracted by GENIES are integrated into the larger GeneWays system after they are "unwound" into simple binary statements (e.g., "Interlukin-2 binds Interlukin-2 receptor" [4]).

The GeneScene parser [6], the Arizona Relation Parser (ARP) [7], and a system developed by Palakal et al. [11] extract relational triples with labeled links connecting labeled entities. The GeneScene parser focuses on closed class words to identify important relations while ARP uses a hybrid syntax and semantic parser. Both extract negation indicators. In the ARP results, the link labels consist primarily of verbs or verb phrases and a negation indicator. Entities are phrases (generally noun phrases) extracted from the text. Table 1 shows 4 examples of relations extracted by ARP from MEDLINE abstracts. ARP output does directly support relation nesting. However, the second entity in relation (4), *E1A-induced apoptosis,* includes a substance (*E1-A*), a function (*apoptosis*), and an associator (*-induced*). ARP frequently extracts these complex entities, capturing important

information beyond mere substance identification. Multiple relations can also be extracted from a single sentence as shown in the two relations extracted for the example sentence (4). This behavior represents a kind of relation nesting**.**

Table 1. Arizona Relation Parser Output

Original Sentence	Resulting Relation			
	Entity 1	Negation	Connector	Entity 2
(1) wild-type p53 tumor suppressor protein, which induces […] apoptosis…	wild-type p53 tumor suppressor protein	False	induces	Apoptosis
(2) Wt p53 also induced significant apoptosis	Wt p53	False	also induced	significant apoptosis
(3) oncogene mutant p53 suppresses apoptosis	oncogene mutant p53	False	suppresses	apoptosis
(4) mutant p53 blocked E1A-induced apoptosis	mutant p53	False	blocked	E1A-induced apoptosis
	E1-A	False	Induced	apoptosis
(5) mutant p53 […] does not induce […] apoptosis	mutant p53	True	does not induce	apoptosis

The system described in [11] adds tags to text marking the boundaries of identified biological objects. Relations are specified by a subject, an object, and a relational classification. Classifications include directional relationships such as "binds" and "inhibits" as well as a hierarchical relation "same" to designate an "is a" relationship. Relationships are expressed between simple objects such as *KiAA1009 protein* but can also involve more complex objects such as *nuclear mitotic apparatus protein* or *cellular transfer RNA for tryptophan*. In addition to reporting on their object and relation extraction, [11] proposes an algorithm for grouping object synonyms. The grouping process does limited co-reference resolution and expands contractions to identify multiple references to the same object within a document. This process reduced the number of unique objects by 24.7% with 92% recall and 82% grouping specificity. That is, 18% of the groupings made were incorrect. The paper included a number of suggestions for improving the grouping process.

3 Five Levels of Relational Granularity

Based on feedback from researchers using GeneScene's visualization tools and experience employing extracted information for data mining applications we have developed a list of five potentially useful levels of relational granularity. These levels define how extracted relations can appropriately be combined in

consolidating a network of extracted relations. They are intended to correspond to levels of aggregation that can be applied to extracted relations with reasonable accuracy. Thus, our selections were influenced by the needs of various users, the nature of available ontological resources, and the characteristics of the extracted relations. Once relations have been processed for aggregation it should be possible to provide a flexible query interface over various levels of granularity, extract simple pathway relation results for use in data mining and other analysis systems, and provide higher levels of analysis (e.g., identifying conflicting relations).

Our approach targets relations that describe how two biological objects interact. Each relation consists of two labeled entities (biological substances or functions) and a labeled connector (verbs). Entities may be assigned multiple features such as species, mutation, cellular component, and substance type. The intuition behind this approach is highlighted in a biologist's observation that we should combine references to wt MDM2, wild-type MDM2, and non-mutant MDM2. Each reference can be correctly understood as a non-mutated form of MDM2. Although some extraction systems may extract this kind of feature information as a relation, for our purposes we would want to assign the feature non-mutated to the entity and recognize the substance as a gene or protein related to the gene MDM2. Connectors are combined by grouping verbs. Appropriate methodologies need to be studied but we plan to begin our analysis using two fairly simple approaches: different morphological forms of a verb are combined, as in *induces* vs. *induce*, and a higher level of semantics can be applied to group verbs such as *inhibits* and *abrogates* into a common connector *inhibit*. Thus our aggregation approach allows for aggregation of complete relations as well as individual entities. We plan to support five levels of granularity in our aggregation system.

1) **Baseline** – the full text of the entity or connector labels must match
2) **Feature Match**
   a) all identified entity features must match
   b) morphological forms of the same connector verb are combined
3) **Typed Substance**
   a) entities with different identifiable substance types are not matched
   b) morphological forms of the same connector verb are combined
4) **Aggregatable Substance**
   a) references to a gene and its gene products are matched
   b) morphological forms of the same connector verb are combined
5) **Simple Pathway**
   a) references to a gene and its gene products are matched
   b) connector verbs are classified into one of 4 categories

Baseline aggregation is used to provide the maximum amount of differentiation in a visualization. We expect that it is most useful as a baseline for

comparison of aggregation system results. Baseline aggregation makes no attempt to combine equivalent objects unless they are labeled with exactly the same words. Thus neither the relation nor any of the elements of *Mdm2 – inhibits – apoptosis* would be matched with *Mdm2 genes – are involved in – regulation of apoptosis*. Baseline aggregation minimizes information loss but accomplishes very little consolidation.

Feature Match aggregation increases network consolidation by comparing feature values assigned to an entity. If a substance has been identified as mutated or recognized as present in a particular tissue type or cellular domain, it is matched only with similarly identified items. For instance, mdm2 *antisense oligodeoxynucleotide – induces – Apoptosis* and *anti-sense MDM2 — induces – apoptosis* would be aggregated because the connectors and second entity (*induces – apoptosis*) match and both MDM2 entities can be identified as mutated (*antisensed*) forms of the substance MDM2. This level of aggregation may be useful for detailed pathway analysis.

Typed Substance aggregation is somewhat comparable to the granularity found in some manually created databases. In a network aggregated at this level entities with different substance types are not matched thus relations involving protein P53 would not be combined with references to the TP53 gene. In this case, *tumour suppressor gene p53 – induces – apoptosis* would be aggregated with *p53 tumour suppressor gene – is known to induce – apoptosis* but not with *p53 protein – induces – apoptosis* because the gene would not be considered equivalent to the protein.

Aggregatable Substance aggregation assigns equivalence to references to a gene and its related gene products. At this level of aggregation no attempt is made to distinguish between interactions related to a particular gene and interactions for the protein that gene encodes. This is partly a practical matter. Across a set of abstracts the exact same phrase is frequently used to refer to a gene and to the related protein making it difficult to distinguish these references. Analysis of nearby words and other cues in the document can help address but not eliminate this ambiguity. Also, as a matter of application, a researcher studying effects of the gene TP53 might well be interested in references to the protein p53 because the presence of the protein is related to expression of the gene. Entities here may also be biological functions and connectors at this level and morphological forms of a verb are matched (e.g., *MDM2 – inhibits – apoptosis* and *MDM2 oncoprotein – has been shown to inhibit – apoptosis* are considered to be equivalent relations).

Simple Pathway aggregation creates an overview of the information extracted from a text. This kind of relation can be used for example, as input into a data mining algorithm. Connecting verbs would be classified as belonging to one of 4 categories: induce, inhibit, directional association, non-directional association. Relations at this level might be comparable to relations extracted by parsers that identify only single semantic types of relations as in the work reported in [8] where only inhibition relations are extracted. Relations could be filtered so that only items

with two recognizable substances and particular types of interaction are included. For a researcher, this representation of a pathway relation can be viewed as an outline or backbone of the regulatory network. For example, each of these relations,

MDM2 inhibits apoptosis

MDM2 oncoprotein abrogates apoptosis

Human MDM2 interferes with p53-mediated cell death

can be aggregated into a simple relation *MDM2 inhibits apoptosis*. *Gene MDM2* and its product *MDM2 oncoprotein* are matched to each other as the aggregatable substance *MDM2*, the verbs *inhibits, abrogates* and *interferes with* would all belong to the category inhibit, and *apoptosis* and *cell-death* would be identified as equivalent functions.

To implement our five level aggregation strategy, we need to identify multiple references to the same aggregatable substances and establish a method of linking those aggregatable substance references to other biomedical resources. Fortunately, several existing resources list and cross-reference genes, proteins, and other gene products. One primary task in building our aggregation system will be to construct an aggregatable substance lexicon from existing resources.

4 Merging Ontological Resources to Support Aggregation

Ontologies, lexicons, and controlled vocabularies are crucial to the aggregation process. Ontologies such as GO list concepts and concept classes and sometimes enumerate class instances (as GO does for biological functions). Other resources such as RefSeq, HUGO, LocusLink, and SGD also contribute to an overall ontology for the domain. For example, the concept class "Gene" is enumerated in the many human genes listed in LocusLink. Gene products such as proteins and mRNA are associated with Genes in the RefSeq repository. We will refer to these resources together as "ontological resources". Many different resources containing lists of biological object name strings have been created to support various tasks. Combining several lists will improve the coverage of an extraction system, but creating a lexicon merging process is a non-trivial task. One key merging problem is ambiguity. Ambiguity occurs (1) within a resource when a single name string is associated with more than one biological object, (2) between resources when two different resources associate the same name string with two or more different biological objects, and (3) when listed name strings are also commonly used English words such as *an*, *by*, *killer*, or *for*.

Two recent efforts to combine lists from multiple sources are documented in [12] and [2]. In [12] entries from selected mouse, fly, worm, and yeast lexicons were combined. Three kinds of ambiguity were measured: multiple name strings for the same gene, multiple genes identified by a single name string, and overlap with common English words. Ambiguity within each database was minimal, between 0% and 2.5%, except for the fly dataset with ambiguity found for 10% of the

references. Across datasets there was significantly more ambiguity, ranging between 4% and 20%. [2] combined name strings from four existing resources (HUGO, SWISSPROT, OMIM, and TREMBL) into an unambiguous list of name references. They describe their generally automatic but manually adjusted process for lexicon curation. To test their lexicon, they selected a set of documents from which relations had been manually extracted for the TRANSPATH database. They report that they matched 94% of the substance names and that dictionary curation improved precision from 78% to 90%.

To support our five level aggregation approach, we created an aggregatable substance lexicon consisting of substance name strings from several ontological resources including RefSeq, LocusLink, HUGO, and SGD. RefSeq is a comprehensive repository of curated reference sequences for transcripts, proteins and genomic regions [13]. LocusLink provides an interface to curated sequences and descriptive information about genes with links to gene-related resources [13]. Entrez Gene was recently deployed as a replacement for LocusLink; in future work we will adjust to the new input data format. HUGO [14] and SGD [15] are standard databases for human and yeast genes respectively. We used the following information: from LocusLink official gene names, official symbols and aliases, and products and aliases; from RefSeq gene/protein names and synonyms; from Hugo previous symbols, aliases, and previous gene names; and from the SGD yeast gene collection symbols, names, and synonyms. We used LocusLink IDs for Human and RefSeq accession numbers for other species as the primary identifier of an aggregatable substance. A few erroneous entries (noise) inevitable in large data sources were removed from the lexicon.

Many name strings occur in several of the ontological resources. In some cases this reflects ambiguity, in other cases simple redundancy. Table 2 documents the degree of name string overlap in the resources we used. In some cases overlap was very high. For example, 68.5% of the HUGO name strings were also found in the RefSeq database. Still, each database seems to have added some new name strings to the list. Because human and yeast abstracts are our primary targets at this time, we report on overlap for those species independently.

For our purposes, a name string is only useful if it actually appears in the analyzed text. We began by measuring how frequently the name strings in our resources occur in MEDLINE abstracts. To implement this overlap test, we preprocessed the texts and the name strings to normalize word boundaries and used left-to-right, longest-first phrase matching. The name string list (NSL) included 214,862 unique, unfiltered words or phrases extracted from RefSeq, LocusLink, HUGO, and SGD. We evaluated the NSL list as a collection of lists, one for each resource and as a combined list with all duplicates removed. These name strings were compared to more than five million MEDLINE abstracts. This set was prepared by excluding non-English MEDLINE records and records with no abstract. As shown in Table 3, we found 35,289 of the unique (combined) NSL items in the abstracts. Only a small portion (35,289 / 214,862 or 16.4%) of the

available name strings is directly useful in recognizing substance references in our target texts.

Table 2. Unique name string overlap between resources

	RefSeq	LocusLink	HUGO	SGD	Other Databases
RefSeq	169,312 a	24.2%	8.0%	1.1%	24.6%
LocusLink	58.8%	69,550 a	26.6%	2.3%	66.4%
HUGO	68.5%	94.0%	19,961 a	3.7%	94.8%
SGD	14.1%	11.1%	5.1%	14,089 a	15.6%

a - Diagonal entries list the number of unique name strings found e.g., 169,312 for RefSeq

The non-diagonal cells represent that portion of the entries from the row's lexicon also occurring the column's lexicon

The last column (Other Databases) shows the percentage of the entries from the row's lexicon that are found in any of the other resources

Table 3. Unique name strings found in 5 million MEDLINE abstracts by source and species

	Human	Yeast	Other	All Species
RefSeq	20,036	631	16,296	20,634
LocusLink	23,755			23,755
HUGO	9,070			9,070
SGD		6,192		6,192
Combined	25,570	6,588	16,296	35,289

While most of this overlap represents legitimate substance references, some of the matches are incorrect in that they result from overlap between name strings and common English words. We compared the extracted name strings to the MOBY list of common English words. Previous results show an overlap between 0 and 2.4% for substance names and common English words using the MOBY list and selected fly, worm, yeast, and mouse substance names [12]. Our results are comparable as shown in Table 4. We found overlap ranging from 0.16% to 3.19% between various resources and the MOBY list. Table 5 reports on only those NSL entries that actually occurred in MEDLINE abstracts and were common English words. Overlap between NSL and MOBY for human substance name strings has risen to 4.8% from the 1.29% reported in Table 4 for this important subset. Several systems have been created to identify biomedical substance references in free text (e.g. [16],[17]). These systems can use contextual cues to assign entity name boundaries; this might help them avoid identifying common English words as substances. One such system is the PROPER system [16]. This kind of processing may filter out some incorrect references (such as occurrences of common English words). Because PROPER is readily available on the internet, we also compared the NSL to a set of entities extracted by the PROPER system. Because some additional abstract preparation was required to run PROPER, we used a smaller set of 87,903 abstracts in this comparison. These abstracts were selected by searching MEDLINE for documents related to the p53 tumor suppressor gene. The PROPER system extracted 419,302 unique name strings from those abstracts. Of those, only 2.5%

(10,580 / 419,302) were found in the NSL. Table 6 breaks down the overlap by resource and species.

Table 4. Percentage of name strings that are common English words

	Human	Yeast	Other	All Species
RefSeq	1.59%	0%	0.77%	0.83%
LocusLink	1.68%			1.68%
HUGO	3.19%			3.19%
SGD		0.16%		0.16%
Combined	1.29%	0.16%	0.77%	0.74%

Table 5. Percentage of unique name strings from MEDLINE that are also common English words

	Human	Yeast	Other	All Species
RefSeq	5.3%	0%	5.3%	5.2%
LocusLink	4.7%			4.7%
HUGO	6.8%			6.8%
SGD		0.4%		0.4%
Combined	4.8%	0.4%	5.3%	4.2%

Table 6. Number of unique PROPER entity names found in the NSL

	Human	Yeast	Other	All Species
RefSeq	6,933	62	5,532	8,543
LocusLink	8,073			8,073
HUGO	3,730			3,730
SGD		1,661		1,661
Combined	8,492	1,688	5,532	10,580

5 Aggregatable Substances and Lexical Ambiguity

In addition to common English word ambiguity, we tabulated situations where a single name string is used to reference multiple substances in a single species either in the same resource or in different resources, and where a single name string refers to substances in different species. Ambiguity values of 2-20% have been reported in previous research [2, 12]. Our results are shown in Table 7. The 11.7% reported for human name strings in RefSeq represent cases where a name string is associated with more than one substance. An example of this kind of ambiguity involves *RAB38* which is associated with RefSeq id *NP_071732* (a protein) and *NM_022337* (an mRNA). The "combined" row of Table 7 counts cases where a name string is associated with different substances across multiple resources. For example, the 39.4% combined human ambiguity includes the name string *p53* which is associated with an mRNA in RefSeq and with a gene in LocusLink. The "All Species" column includes examples such as *ACP1* which represents a

particular human gene in LocusLink and a completely different yeast gene in SGD. The name strings that actually occur in MEDLINE abstracts are much more ambiguous as compared to the list of all available name strings. This is likely due to the common practice of referring to a gene and the protein it encodes with the same one word name string. These entries make up a substantial percentage of the items actually observed in MEDLINE abstracts.

Table 7. Percentage of unique name strings associated with multiple substances

	All NSL Entries				Only entries found in MEDLINE			
	Human	Yeast	Other	All Species	Human	Yeast	Other	All Species
RefSeq	11.7%	10.3%	5.0%	15.6%	21.7%	14.4%	11.4%	51.1%
LocusLink	3.0%			3.0%	6.8%			6.8%
HUGO	3.1%			3.1%	5.6%			5.6%
SGD		2.5%		2.5%		3.8%		3.8%
Combined	39.4%	3.8%	5.0%	23.1%	73.2%	4.7%	11.4%	62.0%

Our system implements a single aggregatable substance ID for a gene and its gene products. This was accomplished primarily through the cross-reference information in RefSeq. RefSeq lists gene products and frequently includes the LocusLink identifier of the related gene. When this occurs we record both identifiers but use the LocusLink ID as the aggregatable substance ID. For example RefSeq provides the name string PIRB as a synonym for 3 mRNA transcription variants associated with LocusLink ID 29990. Thus, our list includes both PILRB (the official symbol) and PIRB (the synonym) as name strings associated with LocusLink ID 29990. To see how much this consolidation reduces name string ambiguity, we recalculated the percentage of ambiguous name strings where a single name string refers to a substance outside of its aggregatable substance. Table 8 shows that ambiguity is substantially reduced using the aggregatable substance approach. In Table 7 we showed that 39.4% of the human name strings found in the combined NSL list were associated with more than one substance. When considering only those items found in MEDLINE, ambiguity is even higher at 73.2%. Using our aggregatable substance approach we found only 6.1% of the name strings to be ambiguous, an 85% improvement. An even higher (89%) improvement was recorded for items that occur in MEDLINE with only 7.8% of the name strings ambiguous as to aggregatable substance. Similar improvement (84%) was found for NSL items that were also extracted by PROPER from our p53–related abstract collection. An aggregatable substance lexicon can be used to combine entities at the aggregatable substance and simple pathway levels of a five level framework of relational granularities. Our results show that considering existing lexicons from this perspective substantially reduces lexical ambiguity.

Table 8. Reduction in ambiguity using the aggregatable substance approach.

	Name Strings Associated with Multiple Substances	Name Strings Associated with Multiple Aggregatable Substances	Improvement
NSL Entries	39.4%	6.1%	85%
NSL Entries found in MEDLINE abstracts	73.2%	7.8%	89%
NSL Entries Found by PROPER	78.6%	12.4%	84%

* This table reflects human substance name strings only

6 Discussion and Future Directions

Our merged NSL associates name strings with substances that would appear at different levels of a biomedical ontology. Some resources list genes, some proteins, and others mRNA. The close relationship that exists between certain substances is reflected in the words authors use in referring to those substances and in the name strings associated with those substances in biomedical lexicons. To address this resulting ambiguity we have proposed the use of a single aggregatable substance ID for a gene and its products. Our results show that this approach substantially reduces lexical ambiguity. Unfortunately, using this approach, a substance name matcher would be unable to differentiate between genes and proteins. This problem can be addressed a number of other ways. In many cases, nearby words are likely to provide useful clues that can be used to differentiate references so a substance reference could be matched to an aggregatable substance list and the substance type could then be clarified using other techniques. Furthermore, there are likely to be many tasks for which this differentiation is not essential. For example, a researcher studying the relationship between p53 and apoptosis might well be interested in literature connecting apoptosis to either the p53 protein or the TP53 gene. The aggregatable substance notion presented here will be an integral part of a larger system that organizes extracted relations to support human visualization and algorithmic analysis of information from biomedical texts. In future work we plan to consider extending the aggregatable substance notion to account for homologous genes and we have already begun development of an aggregation system to support our five level aggregation approach.

7 Acknowledgments

This project was supported by the following grant: NIH/NLM, 1 R33 LM07299-01, 2002-2005, ‘‘Genescene: a Toolkit for Gene Pathway Analysis’’

References

1. A. Rzhetsky et al., "A knowledge model for analysis and simulation of regulatory networks", *Bioinformatics* **16**, 1120-1128 (2000)
2. D. Hanisch et al., "Playing biology's name game: identifying protein names in scientific text", PSB 2003
3. R. Gaizauskas et al., "Protein structures and information extraction from biological texts: the PASTA system" , *Bioinformatics* **19**, 135-143 (2003)
4. C. Friedman et al., "GENIES:a natural-language processing system for the extraction of molecular pathways from journal articles", *Bioinformatics* **17**, 74-82 (2001)
5. A. Rzhetsky et al., "GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data", *J. of Biomedical Informatics* **37**, 43-53 (2004)
6. G. Leroy, H. Chen, and J. D. Martinez, "A shallow parser based on closed-class words to capture relations in biomedical text", *J. of Biomedical Informatics* **36**, 145-158 (2003)
7. D. M. McDonald et al., "Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser", *Bioinformatics* Forthcoming (2004)
8. G. R. G. Lanckriet et al., "Kernel-based data fusion and its application to protein function prediction in yeast", PSB 2004
9. T. C. Rindflesch et al., "EDGAR: extraction of drugs, genes and relations from the biomedical literature", PSB 2000.
10. J. Pustejovsky et al., "Robust relational parsing over biomedical literature: extracting inhibit relations", PSB 2002
11. M. Palakal et al., "Identification of biological relationships from text documents using efficient computational methods", *J. of Bioinformatics and Comp. Biology* **1**, 307-342 (2003)
12. O. Tuason et al., "Biological nomenclatures: a source of lexical knowledge and ambiguity", PSB 2004.
13. D. L. Wheeler et al., "Database resources of the National Center for Biotechnology Information: update", *Nucleic Acids Res.* **32**, D35-40 (2004)
14. H. M. Wain, "Genew: the Human Gene Nomenclature Database, 2004 updates", *Nucleic Acids Res.* **32**, D255-7 (2004)
15. K. R. Christie et al., "Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms", *Nucleic Acids Res.* **32**, D311-4 (2004)
16. K. Fukuda et al., "Toward information extraction: identifying protein names from biological papers" , PSB 1998
17. L. Tanabe and W. J. Wilbur, "Tagging Gene and Protein Names in Full Text Articles, Natural Language Processing in the Biomedical Domain", July 2002, Assoc. for Comp. Ling.