

Software

Open Access

OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression

Lawrence Hunter*¹, Zhiyong Lu², James Firby³, William A Baumgartner Jr¹, Helen L Johnson¹, Philip V Ogren⁴ and K Bretonnel Cohen¹

Address: ¹Center for Computational Pharmacology, University of Colorado School of Medicine, Aurora, CO 80045, USA, ²National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA, ³PowerSet, Inc., San Francisco, CA 94107, USA and ⁴Department of Computer Science, University of Colorado, Boulder, CO 80303, USA

Email: Lawrence Hunter* - Larry.Hunter@uchsc.edu; Zhiyong Lu - luzh@ncbi.nlm.nih.gov; James Firby - Jim.Firby@gmail.com; William A Baumgartner - William.Baumgartner@uchsc.edu; Helen L Johnson - Helen.Johnson@uchsc.edu; Philip V Ogren - Philip@Ogren.info; K Bretonnel Cohen - Kevin.Cohen@gmail.com

* Corresponding author

Published: 31 January 2008

Received: 29 August 2007

BMC Bioinformatics 2008, 9:78 doi:10.1186/1471-2105-9-78

Accepted: 31 January 2008

This article is available from: <http://www.biomedcentral.com/1471-2105/9/78>

© 2008 Hunter et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Information extraction (IE) efforts are widely acknowledged to be important in harnessing the rapid advance of biomedical knowledge, particularly in areas where important factual information is published in a diverse literature. Here we report on the design, implementation and several evaluations of OpenDMAP, an ontology-driven, integrated concept analysis system. It significantly advances the state of the art in information extraction by leveraging knowledge in ontological resources, integrating diverse text processing applications, and using an expanded pattern language that allows the mixing of syntactic and semantic elements and variable ordering.

Results: OpenDMAP information extraction systems were produced for extracting protein transport assertions (transport), protein-protein interaction assertions (interaction) and assertions that a gene is expressed in a cell type (expression). Evaluations were performed on each system, resulting in F-scores ranging from .26 – .72 (precision .39 – .85, recall .16 – .85). Additionally, each of these systems was run over all abstracts in MEDLINE, producing a total of 72,460 transport instances, 265,795 interaction instances and 176,153 expression instances.

Conclusion: OpenDMAP advances the performance standards for extracting protein-protein interaction predications from the full texts of biomedical research articles. Furthermore, this level of performance appears to generalize to other information extraction tasks, including extracting information about predicates of more than two arguments. The output of the information extraction system is always constructed from elements of an ontology, ensuring that the knowledge representation is grounded with respect to a carefully constructed model of reality. The results of these efforts can be used to increase the efficiency of manual curation efforts and to provide additional features in systems that integrate multiple sources for information extraction. The open source OpenDMAP code library is freely available at <http://bionlp.sourceforge.net/>

Background

Conceptual analysis is the process of mapping from natural language texts to a formal representation of the objects and predicates (together, the concepts) meant by the text. The history of attempts to build programs to do conceptual analysis dates back to at least 1967 [1]. Recent advances in the availability of high quality ontologies, in the ability to accurately recognize named entities in texts, and in language processing methods generally have made possible a significant advance in concept analysis, arguably the most difficult and general natural language processing task. Here we report on the design, implementation and several evaluations of OpenDMAP, an ontology-driven, integrated concept analysis system that significantly advances the state of the art. We also discuss its application to three important information extraction tasks in molecular biology.

Information extraction (IE) efforts are widely acknowledged to be important in harnessing the rapid advance of biomedical knowledge, particularly in areas where important factual information is published in a diverse literature. In a recent *PLoS Biology* essay Rebholz-Schuhmann [2] argued, "It is only a matter of time and effort before we are able to extract facts [from articles in the primary literature] automatically. The consequences are likely to be profound." Existing examples include extraction of information about gene-gene interactions [3], alternative splicing [4], functional analysis of mutations [5], phosphorylation sites [6], and regulatory sites [7]. The primary significance of OpenDMAP to these efforts is that it leverages the large-scale efforts being made in biomedical ontology development, such as the Open Biomedical Ontologies Foundry (OBO Foundry) [8].

Logical representations of reality, such as those built on the OBO Foundry, use a set of predicates that formally describe properties of, or relationships among, objects. Predicates are defined with a specific number and type of admissible arguments. For example, the predicate *expresses* might be specified to take two arguments, a gene and a cell type, meaning that the specified gene is expressed in all normal cells of the specified type. Such predicates can also be related to each other through abstraction ("is a") and packaging ("part of") hierarchies, as done in the OBO Foundry. The semantics defined by the predicates and hierarchies in such ontologies provide a powerful tool for natural language processing.

Independently constructed ontologies have played at best a modest role in prior natural language processing systems. Guarino [9] characterizes various uses of ontologies in information systems: only systems that use an ontology at run time (rather than during system construction) to explicitly represent the domain knowledge exploited by

the system qualified for what Guarino called an "ontology-driven information system proper." To our knowledge, OpenDMAP is the first system developed to exploit a community consensus ontology as the central organizing principle of an information extraction system; for example, none of the systems that participated in the 2004 TREC Genomics evaluation for recognizing instances of Gene Ontology terms in text [10] meet the Guarino definition. Other language processing systems have used either small, *ad hoc* conceptual representations developed specifically for the application, or structured linguistic resources, such as WordNet [11], which do not meet the logical requirements for an ontology. While the implementation reported below exploits only a small portion of the OBO Foundry, and the crucial Relationship Ontology component of the Foundry is still in an early stage of development, the organizing principles of OpenDMAP generalize straightforwardly.

The MetaMap system [12] identifies biomedical concepts from free-form textual inputs and maps them to entries in the Unified Medical Language System (UMLS) metathesaurus; SemRep [13] is a related system that maps to predications drawn from the UMLS semantic network, and SemGen [14,15] is another related system that is focused on mapping to UMLS terms relevant to the etiology of genetic disease. These systems and their extensions have been used to extract semantic relationships relevant to pharmacogenomics [16] and to compare alternative sources of information [17], among other applications. OpenDMAP is like MetaMap and its descendents in that it can only produce output drawn from a predefined semantic representation. The main difference is that MetaMap, SemRep and SemGen are structured as traditional NLP systems, with a lexicon that enumerates possible concepts that might be associated with a word or phrase. Multiple possible mappings are returned, with rankings. OpenDMAP provides an alternative method of organizing knowledge about language, so that each concept has associated with it a set of patterns that describe how that concept can be realized in language; there is no explicit lexicon.

To appreciate the differences between OpenDMAP and previous work in biomedical text mining, it is also useful to contrast its handling of syntactic structure and of semantic content with other systems. At one end of the spectrum are systems that employ essentially asyntactic representations. Early in the modern period of genomic natural language processing, some such systems were able to achieve significant (and in some cases ground-breaking) results using techniques based on text literals only. These include [18-20]. One line of subsequent work has attempted to increase the coverage of these early systems, which utilized manually-built patterns, by automatically acquiring considerably larger sets of patterns – see, for

example, Huang et al. 2004 [21]. Another line of subsequent work has focused on adding a modest, but still useful, level of linguistic abstraction by explicitly including either lexical categories (parts of speech), word stems, or both [22,23]. These systems were essentially agrammatical; in contrast, OpenDMAP utilizes a classic form of "semantic grammar," freely mixing text literals, semantically typed basal syntactic constituents, and semantically defined classes of entities.

Although OpenDMAP is capable of utilizing full syntactic parses, the patterns for the three separate tasks discussed in this paper utilize primarily shallow syntactic parses (the development phase of the transport project reports results using syntactic dependency information). It remains to be seen what depth of syntactic parsing is useful in biomedical text mining. Some early systems explored full parsing [24,25], but they were not generally fruitful, and typical systems have employed at most shallow parsing [26-28]; only recently has productive attention returned to syntactically ambitious approaches to biomedical text [29-31], much of it taking a dependency-based, rather than a constituent-based, approach.

All of the systems discussed thus far have in common the fact that they employ some notion of explicit patterns, be they agrammatical, syntactic, or semantic. In a separate line of work, patterns are entirely implicit – that is, they exist only to the extent that they are captured by orthogonal features. This work approaches relation extraction as a classification problem; a classic example is the work of Craven and Kumlein 1999 [32]. Bunescu et al. 2005 [33] presents a detailed analysis of a number of classification-based approaches; the state of the art is characterized by the participants in the recent BioCreative protein-protein interaction shared task [34].

OpenDMAP has been applied in three domains: protein transport, protein-protein interaction and the expression of a gene in a particular cell type. The three application domains are independently significant. Protein transport, the directed movement of proteins from one cellular compartment to another, is a broadly important biological phenomenon. Although protein subcellular localization information is centralized (e.g. through ontological annotations at NCBI and in various model organism databases), information about transport is not. Protein transport information is published throughout the scientific literature, but no previous method was able to capture it systematically. Protein-protein interaction extraction has been the subject of dozens of systems (see, e.g. a review in [35]). Widely used web resources such as IHOP [3] and Chilibot [36] are based entirely on automated extraction of protein-protein interactions from text. This task was used in the BioCreative community evalua-

tion, described below. The third application area, extraction of assertions that a particular gene is expressed in a particular cell type, is of significance since it appears to be the predicate found most frequently in the biomedical literature; a form of the verb "express," usually its nominalization "expression," appears in nearly 20% of NCBI's GeneRIFs [37].

The protein transport task is illustrative of another distinguishing aspect of the OpenDMAP approach: it provides mechanisms for handling relationships involving more than two entities. Note that the protein transport predicate has at least three arguments: what protein is transported, from where, and to where (our model also includes a fourth argument: the transporting protein). Although some linguistic expressions of the concept may elide an argument, the predicate itself inherently describes a greater than binary relationship. Wattarujeekrit et al. [38] and Cohen and Hunter [39] present evidence that many important predicates in biomedicine require more than two arguments. However, most previous efforts at extracting relationships from biomedical text have addressed exclusively binary relationships. Geneways [40] and RLMPSP [41] are the only other biomedical IE systems of which we are aware that extracted greater than binary relationships, and neither is ontology-driven.

Assessing the accuracy of an information extraction system is a very labor-intensive activity. In order to identify information that could have been extracted, but was not (a "false negative"), a person must go through a large volume of text to determine all of the relevant assertions. To estimate the reliability of these manually derived assertions, at least two people must complete that task to assess inter-rater reliability. Once such data is used for one evaluation and system developers have seen it, further use of the data will generate upwardly biased accuracy estimates as system developers fit their systems to it. For these reasons, large-scale community evaluations of information extraction systems are particularly important. The second Critical Assessment of Information Extraction in Biology, (BioCreative) [34,42], community evaluation included a test of systems designed to extract human protein-protein interaction information from the full texts of hundreds of journal articles, called the IPS task. Human curators from the IntAct database [43] manually extracted interaction assertions from these articles using the same curatorial standards as for the database. The results produced by human experts were compared to the results submitted from 45 systems developed by laboratories around the world, providing the best current assessment of the accuracy of protein interaction information extraction systems. The performance of OpenDMAP on the protein interaction task was evaluated as part of this shared task.

More limited evaluations of the accuracy in the other applications are also reported in the results section.

The accuracy of an information extraction system depends on the genre of texts on which it operates [44]. This report demonstrates the application of OpenDMAP to full texts of scientific journal articles, to Medline abstracts, and to GeneRIFs (single sentences or sentence fragments that are selected by human curators for relevance to the function of a particular gene product). GeneRIFs are particularly attractive targets for information extraction, due to their roughly sentential length (identified by [44] as the optimum), breadth of coverage, manual preselection for relevance, and association with at least one normalized gene reference. Despite these attractive features, this is the first report of an information extraction system targeting them.

Results

OpenDMAP information extraction systems were produced for extracting protein transport assertions (trans-

port), protein-protein interaction assertions (interaction) and assertions that a gene is expressed in a cell type (expression). Each of these systems was run over all abstracts in Medline as of June 18, 2007, producing a total of 72,460 transport instances, 265,795 interaction instances and 176,153 expression instances. These results are provided in RDF format in the Additional Files 1, 2, 3, 4.

One particularly striking result is the diversity of journals from which these assertions were mined. The transport relationships were extracted from 2,340 different journals; the interaction relationships from 4,103 different journals; and the expression relationships from 2,984 different journals. A total of 4,434 unique journals contributed to these results, nearly 40% of the journals indexed in Medline each year (see Figure 1).

For the BioCreative evaluation, the interaction system was run on the full texts of all of the 359 articles in the test set, producing 385 interaction assertions. Performance was

% Journals providing Transport, Expression, and/or PPI relations in their abstracts, by year

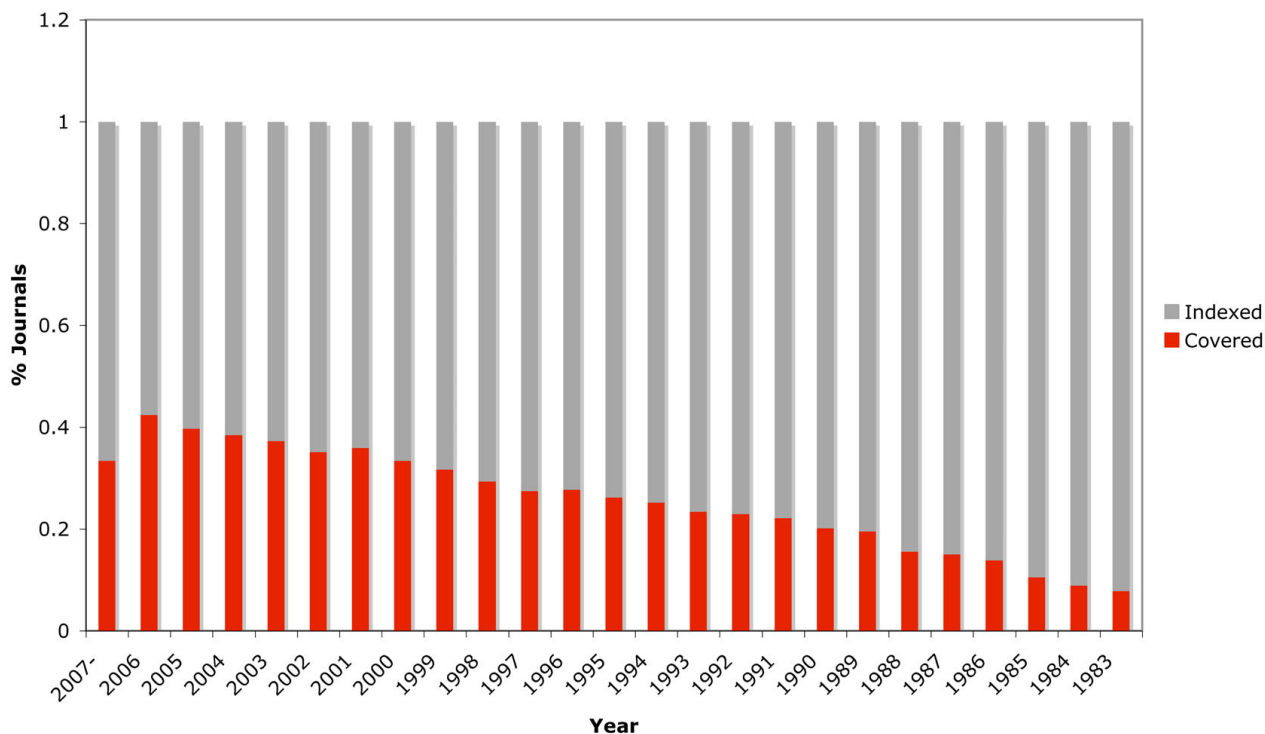


Figure 1
OpenDMAP coverage of MEDLINE. The gray bars indicate the number of journals indexed by MEDLINE each year. The red bars indicate the number of journal abstracts from which OpenDMAP extracted at least one assertion regarding transport, interaction or expression. In recent years, more than 40% of biomedical journals contain such information. 2007 is partial data (through July 1).

averaged per article, since a few articles had a very large number of interactions and would have dominated a per assertion calculation. OpenDMAP's average F-measure of 0.29 was 10% higher than the next best scoring system, and more than three standard deviations above the mean performance. OpenDMAP's recall was similar to the other high scoring systems; its advantage arose from being substantially more precise (fewer false positives), achieving an average precision of 0.39, more than 20% better than the next best system. Due to IntAct's curation criteria, which require clear experimental evidence for an interaction in the text, these results are quite conservative. Many "false positives" were in fact assertions of interactions, but fell short of the evidential requirements for IntAct curation.

A manual evaluation of the performance of the protein transport recognition system was based on all 570 GeneRIFs containing a form of the word "translocate" (382 of which were about protein transport, and 188 were about the transport of something else). Since transport is a greater than binary relationship, the extraction was only counted as correct if all of the components extracted matched the human annotation. For that strict criterion, OpenDMAP achieved precision of 0.75 and a recall of 0.49 (F-score of 0.59). If incomplete extractions are counted as correct, precision is unchanged at 0.75 and recall rises to 0.67 (F-score of 0.71). A substantial proportion of the errors were due to imperfect recognition of proteins; if OpenDMAP is given correct protein identifications as inputs, precision is 0.77, strict recall is 0.67 (F-score of 0.72) and incomplete recall is 0.85 (F-score of 0.81).

A manual evaluation of the performance of the expression recognition system was based on 324 GeneRIFs containing a form of the word "express," (these sentences contained 469 assertions about expression, 205 of which were about gene expression in 178 different cell types). OpenDMAP had a precision of 0.64, but missed many statements that annotators identified as expression assertions, achieving a recall of only 0.16 (F-score of 0.26). A substantial portion of these errors were due to imperfect recognition of gene names; if OpenDMAP is given correct gene identifications as input, precision is 0.85 and recall is 0.36 (F-score of 0.51). Many other failures to identify expression assertions were related to coordination; the test set had an average of more than two expression assertions per sentence, but the IE system extracted only about 1.3 assertions per sentence.

Discussion

As demonstrated by its performance in the community evaluation, OpenDMAP advances the state of the art for extracting protein-protein interaction predications from

the full texts of biomedical research articles. Furthermore, this level of performance appears to generalize to other information extraction tasks, including extracting information about predicates of more than two arguments.

There are several reasons why OpenDMAP exhibits better performance than any other biomedical information extraction system to date. OpenDMAP is an extension of the Direct Memory Access Parsing (DMAP) paradigm described in [45] and [46]. Three innovations distinguish the present work from those prior efforts. First, the ontology component of OpenDMAP is independent of the rest of the system. The knowledge representation component is the well-established, open source Protégé ontology development system [47,48], and OpenDMAP concept analyzers can be associated with any ontology compatible with Protégé, for example, the OBO Foundry. Second, OpenDMAP is fully integrated with the open source Unstructured Information Management Architecture, (UIMA) [49-51], which allows the results of any text processing application interfaced to UIMA to be exploited by the OpenDMAP system. As demonstrated below, this mechanism facilitates the use of many external language processing systems, including tokenizers, sentence boundary detectors, entity recognition systems, and syntactic parsers. Since the inputs and outputs of each system are mapped by UIMA to a common annotation structure accessed by OpenDMAP, the use, comparison and combination of various approaches to language processing can all be fully integrated into OpenDMAP patterns. The third innovation in the OpenDMAP system is an expanded pattern language for specifying how concepts can be expressed in text. The pattern language not only allows specifications of mixtures of any concepts available from either the ontology (e.g. a protein transport process) or the results of UIMA text processing (e.g. the head of a noun phrase), but it also has new features that allow more flexible concept ordering than previous DMAP analyzers (see the description of the pattern language in the Methods section for details).

The intimate connection between the ontology and the natural language processing system provides two significant advantages over prior information extraction systems generally. First, the output of the information extraction system is always constructed from elements of the ontology, ensuring that the knowledge representation is grounded with respect to a carefully constructed model of reality. In contrast, the outputs of most natural language processing systems are grounded only in substrings of text, not normalized to any model at all. Progress in normalizing biological entities recognized in text to specific database identifiers [52-54] has made the output of text processing systems much more valuable. Mapping the properties and relationships extracted to a community

ontology similarly provides a significant increment in the value of the output from text processing systems.

The second advantage of the OpenDMAP approach is that all of the knowledge used by the system to recognize concepts is structured by the ontology. In contrast, the nearly universal alternative approach is to embody knowledge of language into a lexicon, which associates individual lexical items with their possible semantic interpretations. In the OpenDMAP approach, information about which concepts are potentially relevant to the analysis of a particular text passage straightforwardly places limits on the linguistic knowledge relevant to analyzing that passage. This approach finesses many difficult ambiguity resolution problems faced by lexicon-driven systems, since these limits on the knowledge applied to conceptual analysis prevent many multiple interpretation problems from arising at all. For example, the string "hunk" refers to a cell type (human natural killer cells), a gene (hormonally upregulated Neu-associated kinase), and the general English word meaning a large piece of something without definite shape. A traditional, lexicon-driven system would have an explicit method for assigning the correct word sense to any occurrence of the string "hunk." However, OpenDMAP patterns specify expectations of semantic classes (e.g. in the transport application described below, the transported entity must be a protein or a molecular complex); if it is possible to construe a string as an instance of an expected class, the pattern matches. The fact that there might be possible alternative interpretations of the matching string has no consequence, and no explicit ambiguity resolution step is necessary. Ambiguity is a leading cause of errors in text processing systems, and this approach is one of the contributing factors to OpenDMAP's superior performance. Our top-down approach to restricting possible interpretations does not address all problems due to ambiguity in language; for example, errors in preprocessing systems (e.g. syntactic parsing, see below) are not effected.

The use of UIMA greatly facilitates the incorporation of various applications as input to OpenDMAP. The outputs of NLP tools integrated into the system are described by the extensible UIMA type system. In the case that a new type of information is produced by a preprocessor, OpenDMAP patterns would have to be modified to take advantage of the new type of information available. For example, the first time an external cell type tagging system is added, the UIMA type of the result of that processor must be linked to a cell-type concept in an OpenDMAP ontology in order for it to be used in patterns. However, if a new NLP tool produces a UIMA output type that has been used by OpenDMAP previously, then no changes in the ontology or patterns are needed.

We believe that the outputs of information extraction systems are not likely to be useful until the F-score (or at least the precision) is greater than about 0.85 [34], so the various sources of error in these systems must be addressed. A significant cause of errors in the OpenDMAP system as evaluated is incorrect identification of gene and protein names. The UIMA architecture makes it trivial to adopt and exploit better gene/protein recognition systems as they are developed. The best gene name identification and normalization systems from the BioCreative assessment achieved F-scores greater than 0.8, significantly above the ~0.7 F-score of the ABNER system [55] used by OpenDMAP to achieve the reported performance. Use of such a system should improve the performance of OpenDMAP.

Error analysis of the false positives in the transport data set indicates that more than 80% are due to errors in the syntactic analysis. For example, in the sentence "Rho protein regulates the tyrosine phosphorylation of FAK through translocation from the nucleus to the membrane," the subject of the translocation was incorrectly identified as FAK (rather than Rho) by the Stanford parser. That parser was developed for general English rather than biomedical text, so using specialized syntactic analysis systems may improve the precision of OpenDMAP. Remaining problems in false positives are due to problematic tokenization, failures to properly resolve anaphoric reference, and, rarely, negation. False negatives are due to gaps in concept recognition patterns, more than half of which arise from a failure to properly handle coordinated clauses and conjunctions. Addressing these issues remains an open area of research.

Another issue was that the Stanford parser was too slow to use in the application of the transport system to all of Medline, so it wasn't run. OpenDMAP ignores aspects of patterns that require inputs that aren't present, so the patterns that contained syntactic dependencies did not have to be altered. These syntactic constraints are important for accuracy, however. Tested on the gold standard set for the system without the parser precision drops to 0.62, while strict recall remains largely unchanged, rising to 0.51.

Conclusion

Despite OpenDMAP elevating the state of the art for biomedical information extraction significantly beyond previous levels, error rates remain high. In the most challenging BioCreative task, finding curatable assertions in full text documents, only about 29% of the relevant assertions were found, and only about 39% of the extracted assertions were completely correct. Such error rates mean that automatically generated databases cannot replace manual curation efforts. However, the evidence is quite clear that manual curation cannot keep up with the rate of data generation [56]. The surprisingly large

number of journals that contained information relevant to these three IE tasks suggests that the temporal approach taken in [56] may actually underestimate the severity of the problem.

Although the outputs produced by large-scale IE systems are not yet suitable for producing factual databases for direct use by biomedical researchers, the current level of performance provides two important facilities to the research community. First, the results of these efforts can be used to significantly increase the efficiency of manual curation efforts. Each extracted assertion is tied to a specific text, which can be used to direct the attention of manual curators both to relevant documents and to specific relevant passages within a document. Effective integration of IE results into curatorial workflows will require the development of new tools. OpenDMAP developers are working with curators at IntAct to address these issues. The open source availability of OpenDMAP will facilitate the work of others addressing this issue as well. The second important use of the sorts of results that IE systems

are currently able to generate is in statistical integration with multiple sources of noisy data, such as those described in [57] and [58]. As demonstrated in the latter, the proper addition of even noisy data from the literature substantially improves the quality and coverage of protein-protein interaction networks for several species.

Methods

OpenDMAP uses Protégé [47] to provide an object model for the possible concepts (predicates and objects) that might be found in a text. Protégé models concepts (including actions) as classes that participate in abstraction and packaging hierarchies, and relationships as class-specific slots. For example, protein transport is modeled as a class (called PROTEIN-TRANSPORT) and the relationship between a transport event and the protein transported in that event is represented as a slot in that class (called [TRANSPORTED-ENTITY]). Slots can take on values, which can be constrained to be instances of other classes. For example, the [TRANSPORTED-ENTITY] slot of the PROTEIN-TRANSPORT class is constrained to be an

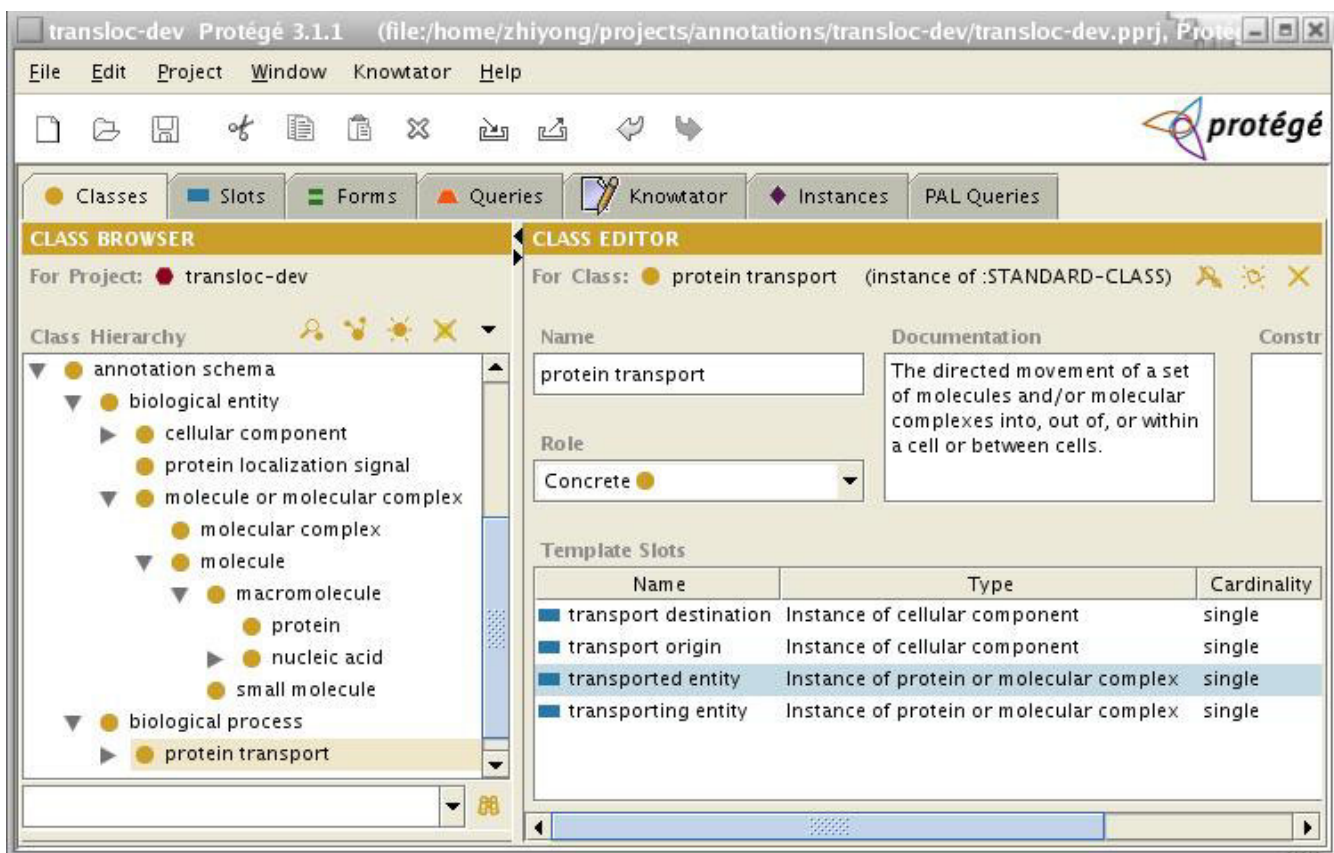


Figure 2
Screenshot of the Protégé ontology for the protein transport task. The slots of the protein transport class are shown in the lower right panel of this screen shot. Note that the subclasses of Cellular Component and Protein Transport are not shown.

instance of either of the classes PROTEIN or MOLECULAR-COMPLEX. Figure 2 shows a portion of the model used for the transport, which includes biological entities, such as molecular complexes and cellular components, and biological processes, particularly protein transport. This model is drawn almost entirely from the Gene Ontology (GO) [53] although the relationships that define the four slots shown in Figure 2 are from a provisional submission to the OBO Foundry Relationship Ontology and are not official. Preprocessing tools (ABNER [55] and LingPipe [59]) were applied to tag instances of proteins, genes, and cell types.

For the transport task, patterns were produced for 30 ontology concepts; eight directly related to transport and 22 others for cellular components that are the sources and destinations of transport. A large number of other concepts (e.g. genes, proteins and cell types) do not have explicit patterns associated with them, but are instead tagged as such by UIMA tools during preprocessing. The protein-protein interaction task involved producing patterns for nine concepts, and the cell expression task required patterns for six additional concepts.

The UIMA architecture [49] manages the processing of document sets. The collection of document processing tools interfaced through UIMA includes the LingPipe tools for sentence boundary detection and tokenization [59], both the LingPipe and ABNER [55] tools for recognizing protein mentions, ABNER for recognizing mentions of cell types, the Stanford Parser [60] to provide syntactic trees, and a locally produced implementation of the Modified Hobbs algorithm [61] for anaphora resolution. Various combinations of these tools were used in the different applications. For example, the GeneRIFs used in the transport application did not require sentence segmentation, and the applications to all medline abstracts did not use syntactic elements in patterns because the Stanford Parser was too slow to run over all of Medline. The results of this preprocessing are stored in UIMA's common annotation structure.

In order to be able to recognize a concept in text, OpenDMAP associates one or more patterns with each concept. A pattern describes the words, phrases, parts of speech, syntactic structures or concepts that should cause an instance of the associated concept to be recognized. A simple pattern, such as the one shown in equation 1, enumerates a disjunction of words that should trigger recognition of a concept.

$$\text{NUCLEUS} := \text{nucleus, nuclei, nuclear}; \quad (1)$$

The patterns for all of the CELLULAR-COMPONENT concepts were derived from the GO term names and syno-

nym, supplemented with derivational variants, such as the adjectival "nuclear" in equation 1. Twenty-two GO cellular component terms were used, along with 19 synonyms associated with the GO terms, and 78 additional derivational variants generated by inspection of the training corpus.

More complex patterns can include references to non-terminals, particularly other concepts. Equation 2 is one of the patterns for recognizing instances of the PROTEIN-TRANSPORT concept. This pattern specifies that a reference to PROTEIN-TRANSPORT can appear in text as a reference to a concept that can fill the [TRANSPORTED-ENTITY] slot, followed by the word "translocation" followed optionally by a phrase beginning with the word "from", possibly including a word with the part of speech determiner, and a concept that could fill the [TRANSPORT-ORIGIN] slot, also followed optionally by a similar phrase that regarding the [TRANSPORT-DESTINATION] slot.

$$\begin{aligned} \text{PROTEIN - TRANSPORT} := & [\text{TRANSPORTED - ENTITY}] \text{translocation} \\ & (\text{from}\{\text{det}\}?\{\text{TRANSPORT - ORIGIN}\})? \\ & (\text{to}\{\text{det}\}?\{\text{TRANSPORT - DESTINATION}\})?; \end{aligned} \quad (2)$$

When a pattern that includes a slot name is matched, the instance created has its slots filled with the concepts that matched the slot names in the pattern. For example, the above pattern matches the GeneRIF that contains "... Bax translocation to mitochondria..." (from Entrez GeneID 27113). Bax, which is recognized as a protein by ABNER, will cause an instance of the protein concept to be created; an instance of a protein matches one of the constraints on filler of the [TRANSPORTED-ENTITY] slot, which causes that slot to match to the Bax protein concept. The word "translocation" matches, and, while the optional "from" clause does not match, the "to" clause does match, since "mitochondria" matches one of the patterns for a subclass of CELLULAR-COMPONENT, the constraint on the filler of the [TRANSPORT-DESTINATION]. Since the entire pattern matches, an instance of PROTEIN-TRANSPORT is created, with the Bax protein concept in its [TRANSPORTED-ENTITY] slot and an instance of the mitochondria concept (from GO's cellular component hierarchy) in its [TRANSPORT-DESTINATION] slot.

OpenDMAP patterns can express variability in word and phrase order. Note, for example, that equation 2 would fail to match the phrase "Bax translocation to mitochondria from the cytosol." The special pattern marker @ is used to identify a set of subpatterns that are both optional and can occur before or after a required phrase; multiple @ marked phrases can occur in any order. For example,

equation 2 can be modified with this marker to recognize the above text:

```
PROTEIN-TRANSPORT := ([TRANSPORTED-ENTITY] translocation)
                    @ (from {det}? [TRANSPORT-ORIGIN])
                    @ (to {det}? [TRANSPORT-DESTINATION]);
(3)
```

Many sentences in the literature express multiple concepts, making extraction of even simple assertions problematic. Consider the following GeneRIF from GeneID:29560: "...HIF-1alpha which is present in glomus cells translocates to the nucleus..." The intervening phrase "which is present in glomus cells" prevents the pattern in equation 3 from matching that sentence. OpenDMAP does have a wildcard character (underscore) that could be added to the pattern in equation 3, between the [TRANSPORTED-ENTITY] concept and the word "translocation," allowing this sentence to be matched. However, using such a wild card would make any protein mentioned before the word "translocation" match the pattern, which is too promiscuous. To address this problem, OpenDMAP allows patterns to specify syntactic constraints on potential matches. For example, the [TRANSPORTED-ENTITY] slot can be constrained to have a syntactic dependency on the head of a phrase that contains the translocate action, thereby constraining it both semantically (it must be a protein or molecular complex) and syntactically (it must be the subject, object or modifier of the translocation). Furthermore, the reliance on the exact word "translocation" can be relaxed to be any reference to a transport action word, including both verbal and nominal forms of multiple terms (e.g., transported, translocation). The PROTEIN-TRANSPORT class is extended to have an [action] slot that specifies the type of transportation action, to keep track of the term that was used. Equation 4 demonstrates the pattern language for specifying syntactic constraints:

```
PROTEIN-TRANSPORT := ([TRANSPORTED-ENTITY dep:x] _
                      [action ACTION-TRANSPORT head:x])
                    @ (from {det}? [TRANSPORT-ORIGIN])
                    @ (to {det}? [TRANSPORT-DESTINATION]);
(4)
```

The use of the variable "x" in the specification identifies a specific syntactic unit, linking the dependency to the head of a phrase. Multiple variables can be used to specify constraints on different syntactic units within a sentence.

OpenDMAP patterns are very powerful. Only five such patterns, shown in equations 5–9 were required for the transport extraction system performance noted above. These patterns were devised manually, based on expert

knowledge of the domain and on a small training set of sample GeneRIFs.

```
PROTEIN-TRANSPORT := [TRANSPORT-DESTINATION] [action ACTION-TRANSPORT] _
                    (of {det}? [TRANSPORTED-ENTITY])?
                    (by {det}? [TRANSPORTING-ENTITY]);
(5)
```

```
PROTEIN-TRANSPORT := ([TRANSPORTED-ENTITY dep:x] _
                      [TRANSPORT-DESTINATION]
                      [action ACTION-TRANSPORT head:x]
                      (by {det}? [TRANSPORTING-ENTITY])?);
(6)
```

```
PROTEIN-TRANSPORT := [action ACTION-TRANSPORT]
                    @ (of {det}? [TRANSPORTED-ENTITY])
                    @ (by {det}? [TRANSPORTING-ENTITY])
                    @ (from {det}? [TRANSPORT-ORIGIN])
                    @ (to|toward|towards|into {det}?
                      [TRANSPORT-DESTINATION]);
(7)
```

```
PROTEIN-TRANSPORT := ([TRANSPORTED-ENTITY dep:x] _
                      [action ACTION-TRANSPORT head:x])
                    @ (by {det}? [TRANSPORTING-ENTITY])
                    @ (from {det}? [TRANSPORT-ORIGIN])
                    @ (to|toward|towards|into {det}?
                      [TRANSPORT-DESTINATION]);
(8)
```

```
PROTEIN-TRANSPORT := ([TRANSPORTED-ENTITY] (is|were|are|was)
                      [action ACTION-TRANSPORT-PASSIVE])
                    @ (by {det}? [TRANSPORTING-ENTITY])
                    @ (from {det}? [TRANSPORT-ORIGIN])
                    @ (to|toward|towards|into {det}?
                      [TRANSPORT-DESTINATION]);
(9)
```

These patterns were augmented with 119 cellular component patterns.

The test data used in the transport and expression evaluations were marked up by domain experts trained in conceptual annotation, using the Knowtator annotation tool [62].

Availability of data and software

The OpenDMAP platform-independent Java 1.5 source code, including UIMA wrappers for the tools used in this work and the patterns for the three tasks, is available from <http://bionlp.sourceforge.net/> under the Mozilla Public License v1.1 (OpenDMAP) and GPL v2.0 license (UIMA wrappers). The results of the information extraction effort are available as RDF format files in the Additional Files 1, 2, 3, 4.

Authors' contributions

LH conceived of the project, supervised the design and implementation of the system and wrote the manuscript. ZL was responsible for the transport project, including writing the patterns and analyzing the results; he also contributed suggestions for the interaction task. JRF implemented the OpenDMAP pattern recognition engine and its UIMA wrapper. WAB wrote all other infrastructure software, including the other UIMA wrappers, managed the data, applied OpenDMAP to all of MEDLINE, and designed and built other software. HLJ was responsible for the interaction and expression projects, including writing the patterns, analyzing the results, and doing the associated error analyses. PVO managed the creation of the gold standard data for transport and contributed to the design of the pattern language syntax. KBC managed the team, selected the preprocessing tools, coordinated and supervised the interaction and expression task efforts, and provided linguistic and software design contributions during all phases of the project. All authors have read and approved this manuscript.

Additional material

Additional file 1

Transport instances from MEDLINE. This file contains the RDF formatted instances of transport, mined from MEDLINE with OpenDMAP.
Click here for file
[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-78-S1.GZ>]

Additional file 2

Interaction instances from MEDLINE, part 1. The interaction data set is very large. This file contains the first half of RDF formatted instances of interaction, mined from MEDLINE with OpenDMAP.
Click here for file
[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-78-S2.GZ>]

Additional file 3

Interaction instances from MEDLINE, part 2. The interaction data set is very large. This file contains the second half of RDF formatted instances of interaction, mined from MEDLINE with OpenDMAP.
Click here for file
[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-78-S3.GZ>]

Additional file 4

Expression instances from MEDLINE. This file contains the RDF formatted instances of expression, mined from MEDLINE with OpenDMAP.
Click here for file
[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-78-S4.GZ>]

Acknowledgements

This work was funded by NIH grants R01NLM008111 and R01NLM009254 to LH. ZL was also supported in part by the Intramural Research Program of the NIH, NLM. We gratefully acknowledge Martha Palmer's invaluable advice regarding error analysis for the expression project; Martin Krallinger and Alfonso Valencia's coordination of the BioCreative 2 protein-protein interaction shared task; Michael Bada's management of the creation of the gold standard for the expression project; Jesse Paquette's development of the species identification code that we used in the interaction task; and Sue Brozowski and Fontein Davrazou's annotation of data for the transport and expression projects.

References

1. Sparck Jones K: **Natural language processing: A historical review.** *Current Issues in Computational Linguistics: in Honour of Don Walker* (Ed Zampolli, Calzolari and Palmer), Amsterdam: Kluwer 1994.
2. Rebholz-Schuhmann D, Kirsch H, Couto F: **Facts from text -- is text mining ready to deliver?** *PLoS Biol* 2005, **3(2)**:e65.
3. Hoffmann R, Valencia A: **A gene network for navigating the literature.** *Nat Genet* 2004/07/01 edition. 2004, **36(7)**:664.
4. Shah PK, Jensen LJ, Boué S, Bork P: **Extraction of transcript diversity from scientific literature.** *PLoS Comput Biol* 2005, **1(1)**:e10.
5. Horn F, Lau AL, Cohen FE: **Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors.** *Bioinformatics* 2004, **20(4)**:557-568.
6. Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, Wu CH: **Literature mining and database annotation of protein phosphorylation using a rule-based system.** *Bioinformatics* 2005/04/09 edition. 2005, **21(11)**:2759-2765.
7. Saric J, Jensen LJ, Ouzounova R, Rojas I, Bork P: **Extraction of regulatory gene/protein networks from Medline.** *Bioinformatics* 2005/07/28 edition. 2006, **22(6)**:645-650.
8. **Open Biomedical Ontologies Foundry** [<http://obofoundry.org>]
9. Guarino N: **Formal ontology in information systems.** Trento, Italy, IOS Press; 1998:3-15.
10. Hersh W, Bhupatiraju R, Ross L, Johnson P, Cohen A, Kraemer D: **TREC 2004 Genomics track overview.** National Institute of Standards and Technology; 2004.
11. Fellbaum C: **WordNet: An Electronic Lexical Database (Language, Speech, and Communication).** MIT Press; 1998.
12. Aronson A: **Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program.** *AMIA Annu Symp Proc* 2001:17-21.
13. Rindflesch TC, Fiszman M: **The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text.** *J Biomed Inform* 2003, **36(6)**:462-477.
14. Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H: **Semantic relations asserting the etiology of genetic diseases.** *AMIA Annu Symp Proc* 2004/01/20 edition. 2003:554-558.
15. Masseroli M, Kilicoglu H, Lang FM, Rindflesch TC: **Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease.** *BMC Bioinformatics* 2006/06/10 edition. 2006, **7**:291.
16. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC: **Extracting semantic predications from Medline citations for pharmacogenomics.** *Pac Symp Biocomput* 2007/11/10 edition. 2007:209-220.
17. Libbus B, Kilicoglu H, Rindflesch TC, Mork JG, Aronson AR, Hirschman L, Pustejovsky J: **Using Natural Language Processing, LocusLink and the Gene Ontology to Compare OMIM to MEDLINE.** *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases* 2004:69-76.
18. Blaschke C, Andrade MA, Ouzounis C, Valencia A: **Automated extraction of biological information from scientific text: protein-protein interactions.** *Proc Int Conf Intell Syst Mol Biol* 2000/04/29 edition. 1999:60-67.
19. Blaschke C, Oliveros JC, Valencia A: **Mining functional information associated with expression arrays.** *Funct Integr Genomics* 2002/01/17 edition. 2001, **1(4)**:256-268.
20. Blaschke C, Valencia A: **Can bibliographic pointers for known biological data be found automatically? Protein interactions**

- as a case study. *Comparative and Functional Genomics* 2001, **2(4)**:196-206.
21. Huang M, Zhu X, Hao Y, Payan DG, Qu K, Li M: **Discovering patterns to extract protein-protein interactions from biomedical full texts.** *Proc JNLPBA, COLING* 2004:22-28.
 22. Temkin JM, Gilder MR: **Extraction of protein interaction information from unstructured text using a context-free grammar.** In *Bioinformatics Volume 19*. Issue 16 Oxford Univ Press; 2003:2046-2053.
 23. Corney DP, Buxton BF, Langdon WB, Jones DT: **BioRAT: extracting biological information from full-length papers.** *Bioinformatics* 2004/07/03 edition. 2004, **20(17)**:3206-3213.
 24. Park JC, Kim HS, Kim JJ: **Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar.** *Pac Symp Biocomput* 2001:396-407.
 25. Yakushiji A, Tateisi Y, Miyao Y, Tsujii J: **Event extraction from biomedical papers using a full parser.** *Pac Symp Biocomput* 2001:408-419.
 26. Gaizauskas R, Demetriou G, Artymiuk PJ, Willett P: **Protein structures and information extraction from biological texts: the PASTA system.** *Bioinformatics* 2002/12/25 edition. 2003, **19(1)**:135-143.
 27. Leroy G, Chen H, Martinez JD: **A shallow parser based on closed-class words to capture relations in biomedical text.** *J Biomed Inform* 2003, **36(3)**:145-158.
 28. Koike A, Niwa Y, Takagi T: **Automatic extraction of gene/protein biological functions from biomedical text.** *Bioinformatics* 2004/10/29 edition. 2005, **21(7)**:1227-1236.
 29. Rinaldi F, Schneider G, Kaljurand K, Hess M, Romacker M: **An environment for relation mining over richly annotated corpora: the case of GENIA.** *BMC Bioinformatics* 2006, **7 Suppl 3**:S3.
 30. McInnes BT, Pedersen T, Pakhomov SV: **Determining the Syntactic Structure of Medical Terms in Clinical Notes.** *Proc Assoc Comp Ling* 2007.
 31. Pyysalo S, Ginter F, Haverinen K, Heimonen J, Salakoski T, Laippala V: **On the unification of syntactic annotations under the Stanford dependency scheme: A case study on BiolNfer and GENIA.** *Association for Computational Linguistics* 2007.
 32. Craven M, Kumlien J: **Constructing biological knowledge bases by extracting information from text sources.** *Proc Int Conf Intell Syst Mol Biol* 1999:77-86.
 33. Bunesco R, Ge R, Kate RJ, Marcotte EM, Mooney RJ, Ramani AK, Wong YW: **Comparative experiments on learning information extractors for proteins and their interactions.** *Artif Intell Med* 2005/04/07 edition. 2005, **33(2)**:139-155.
 34. Krallinger M, Leitner F, Valencia A: **Assessment of the second BioCreative PPI task: automatic extraction of protein-protein interactions.** 2007:41-55.
 35. Hunter L, Cohen KB: **Biomedical Language Processing: What's Beyond PubMed?** In *Molecular Cell Volume 21*. Cell; 2006:589-594.
 36. Chen H, Sharp BM: **Content-rich biological network constructed by mining PubMed abstracts.** *BMC Bioinformatics* 2004, **5**:147.
 37. **NCBI Entrez Gene GeneRIF Help** [<http://www.ncbi.nlm.nih.gov/projects/GeneRIF/>]
 38. Wattarujeekrit T, Shah PK, Collier N: **PASBio: predicate-argument structures for event extraction in molecular biology.** *BMC Bioinformatics* 2004, **5**:155-175.
 39. Cohen KB, Hunter L: **A critical review of PASBio's argument structures for biomedical verbs.** *BMC Bioinformatics* 2006, **7(Suppl. 3)**:S5.
 40. Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboue PA, Weng W, Wilbur WJ, Hatzivassiloglou V, Friedman C: **GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data.** *J Biomed Inform* 2004/03/16 edition. 2004, **37(1)**:43-53.
 41. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K: **Beyond the clause: extraction of phosphorylation information from medline abstracts.** *Bioinformatics* 2005, **21 Suppl 1**.
 42. **BioCreAtive Challenge Evaluation** [<http://biocreative.sourceforge.net/>]
 43. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R: **IntAct: an open source molecular interaction database.** *Nucleic Acids Res* 2003/12/19 edition. 2004, **32(Database issue)**:D452-5.
 44. Ding J, Berleant D, Nettleton D, Wurtele E: **Mining MEDLINE: Abstracts, Sentences, or Phrases?** *Pac Symp Biocomput* 2002, **7**:326-337.
 45. Martin C: **Direct Memory Access Parsing.** Yale University; 1992.
 46. Fitzgerald W: **Building Embedded Conceptual Parsers.** Northwestern University; 1994.
 47. Noy NF, Crubezy M, Ferguson RW, Knublauch H, Tu SW, Vendetti J, Musen MA: **Protege-2000: an open-source ontology-development and knowledge-acquisition environment.** *AMIA Annu Symp Proc* 2004/01/20 edition. 2003:953.
 48. **Protégé** [<http://protege.stanford.edu/>]
 49. Ferrucci D, Lally A: **Building an example application with the unstructured information management architecture.** In *IBM Systems Journal Volume 43*. Issue 3 IBM; 2004:455-475.
 50. **UIMA Java Framework** [<http://uima-framework.sourceforge.net/>]
 51. **Apache UIMA** [<http://incubator.apache.org/uima/>]
 52. Hirschman L, Colosimo M, Morgan A, Yeh A: **Overview of BioCreative task IB: normalized gene lists.** *BMC Bioinformatics* 2005/06/18 edition. 2005, **6 Suppl 1**:S11.
 53. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000/05/10 edition. 2000, **25(1)**:25-29.
 54. Morgan A, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenburg J, Sun C, Liu H, Torres R, Krauthammer M, Lau MW, Hsu CN, Schuemie M, Hirschman L: **Overview of BioCreative II Gene Normalization.** *Genome Biology* 2008.
 55. Settles B: **ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text.** *Bioinformatics* 2005, **21(14)**:3191-3192.
 56. Baumgartner WA Jr., Cohen KB, Fox LM, Acquaaah-Mensah G, Hunter L: **Manual curation is not sufficient for annotation of genomic databases.** *Bioinformatics* 2007, **23(14)**:e.
 57. Leach SM, Gabow AP, Hunter L, Goldberg D: **Assessing and combining reliability of protein interaction sources.** *Pac Symp Biocomp* 2007, **12**:433-444.
 58. Gabow AP, Leach SM, Baumgartner WA Jr., Hunter L, Goldberg D: **Improving Protein Function Prediction Methods with Integrated Literature Data.** 2003.
 59. Carpenter B: **Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval.** *13th Annual Text Retrieval Conference* 2004.
 60. Klein D, Manning CD: **Fast Exact Inference with a Factored Model for Natural Language Parsing.** In *Advances in Neural Information Processing Systems Volume 15*. MIT Press; 2003.
 61. Kehler A, Appelt D, Taylor L, Simma A: **The (non) utility of predicate-argument frequencies for pronoun interpretation.** *Proc of HLT-NAACL* 2004, **4**:289-296.
 62. Ogren P: **Knowtator: a Protege plugin for annotated copus construction.** *HLT-NAACL* 2006:273.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

