Confirming protein-protein interactions by text mining

D. Otasek[1], K. Brown[2], I. Jurisica[1,2,3]*

[1]Division of Signaling Biology, Ontario Cancer Institute, Toronto, Ontario, Canada
[2]University of Toronto, Departments of Medical Biophysics, and [3] Computer Science

ABSTRACT

Motivation: Although manual curation of protein-protein interactions from literature resulted in several large databases, many interactions are still available only in manuscripts. Though PubMed does include a search engine, protein-protein interactions remain difficult to find in an automated manner.

Results: OPHID Text Miner (OTM) is an information extraction system dedicated to finding specific protein-protein interactions in PubMed abstracts. Originally designed to validate predicted interactions, it can be used to provide additional support for researchers. Using several layers of pattern matching, OTM can extract proof for interactions between two proteins with 47% recall and 93% precision.

Availability: OTM's results have been integrated into OPHID (Online Predicted Human Interaction Database; http://ophid.utoronto.ca). Additional information regarding interaction terms and synonym databases are available upon request. **Contact**: juris@ai.utoronto.ca

Keywords: Protein-protein interactions; information extraction.

1 INTRODUCTION

Knowledge about protein-protein interactions (PPIs) is rapidly growing, with results from experiments available in diverse databases such as BIND, DIP, GRID, HPRD, and MINT. However, large number of PPIs is still available only in text format in PubMed, (http://www.ncbi.nlm.nih.gov/entrez), which is a centralized database that contains links to over 16 million papers in several languages. Attempts to manually curate such data resulted in BIND, DIP, HPRD, and MINT databases. Although accurate, manual curation does not provide optimum coverage, and is resource intensive. BIND for example had many curators devoted to the task of extracting information from PubMed articles. They estimated that approximately 2,000 interactions were reported gper month in the text of papers released on PubMed. For a researcher interested in specific interactions, finding information is compounded by the limitations of querying PubMed. For example, a search for tgf-beta produces over 29,000 matches. Attempting to reduce this to a more manageable number, a researcher could create a query for 'tgf-beta interactions', which still returns in excess of 2,000 articles, many of which will be irrelevant, but importantly, it may not cover all known interactions. Even for a single protein, finding PPIs in PubMed is difficult, further compounding the problem of finding supporting evidence for high-throughput experiments.

OTM (OPHID Text Mining) was designed to find supporting evidence for interaction prediction algorithms, such as those used in the Online Predicted Human Interaction Database (OPHID; http://ophid.utoronto.ca) [3]. OPHID contains human PPIs predicted by several methods from model organism PPI databases [1, 8, 9, 10, 14, 17, 26, 18], combined with human PPIs from high-throughput experiments [24, 21, 2, 13, 15] and from curated databases [28, 27, 20, 1]. Combined, it comprises 47,656 interactions among 10,652 proteins (February, 2006).

Several methods have been developed to make manual curation of PPI data from PubMed more manageable. PreBIND speeds up manual expert reviewing by providing confidence data for the existence of interactions in retrieved abstracts using a Support Vector Machine (SVM) algorithm [6]. SVM is a statistical approach, which appears to perform well in recognition and classification of phrases, without focusing on actual meaning. PreBIND is able to identify phrases containing interactions with 92% recall and 90% accuracy, although the details of the interactions are then extracted by human analysis [6]. By streamlining the tedious task of reviewing articles rather than fully-automating information extraction, PreBIND greatly improves the volume of information that can be retrieved in a given period of time. A step further is the iHOP system (Information Hyperliked Over Proteins), which implements a semantic network for PubMed, by linking genes and proteins to sentences and abstracts [11].

* To whom correspondence should be addressed.

To reduce the challenge of manual curation, one can use a combination of methods to directly extract PPI data from abstracts or full text. Some recent notable information extraction methods include GENIES [7], MedScan [19] and BioRAT [5], all of which implement a semantic approach. Generally, these are rule-based systems, relying on human-generated patterns to recognize phrases. This results in high precision rates, often in excess of 90%; however, this is sometimes at the expense of lower recall, which may range from 20% (BioRAT, MedScan) to 53% (GENIES), depending on the relationship to be extracted. Importantly, these systems can provide more in-depth information regarding the interactions they have found, such as the specific proteins involved, or the interaction type. It is clear that better tagging improves precision and recall (e.g., GENIES system that uses human-generated patterns achieves high both recall and precision). However, scalability of automated information extraction systems requires that all steps in the process are automated.

2 SYSTEM AND METHODS

2.1 System Overview

OPHID Text Miner (OTM) was developed to support validation of predicted interactions in the OPHID database, by automatically extracting human protein interactions from PubMed for protein pairs. OTM is implemented in Java. A local version of PubMed is stored in an IBM DB2 database, which also provides local versions of the SwissProt and Entrez Gene databases for the identification of proteins. OTM's general architecture has five basic modules (see also Figure 1):

1. **Synonym Generation.** We retrieve all synonyms for all the proteins listed in SwissProt, and also add gene names from the available data in Entrez Gene.
2. **Abstract retrieval.** Using OPHID's list of synonyms, OTM queries PubMed's abstract database to search for protein pair co-occurrences. In the case of OPHID, abstracts retrieved are limited to articles containing the MeSH term 'Humans' and synonyms related to Swiss-Prot proteins listed as human, OTM can also limit it's search by other organisms in this manner.
3. **Tagger.** Words in each abstract are classified using a series of dictionaries.
4. **Parser.** Individual phrases of interest are identified by passing the tagged abstracts through several layers of pattern recognition.
5. **Human Review (optional).** An expert may analyze tagged phrases, and update both the dictionaries and the patterns involved in tagging and parsing.

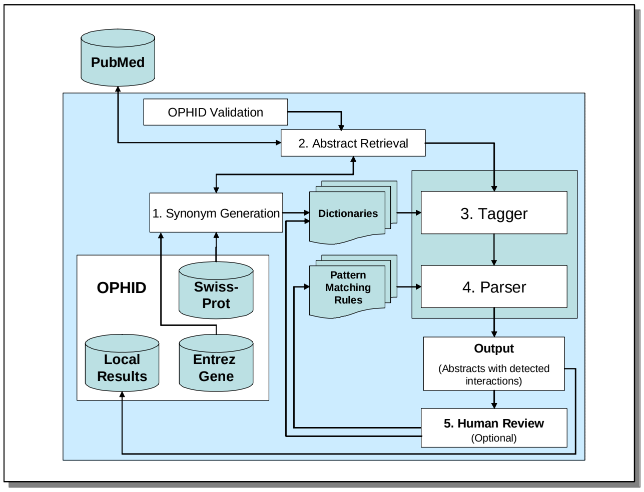

Figure 1. System architecture. OTM workflow starts with generating synonyms for a given interaction pair, retrieving abstracts from PubMed, tagging and parsing the abstracts.

Information extraction in OTM comprises two parts: a tagger and a parser. Both components use a vocabulary, originally developed using ideas from Temkin and Gilder [25]. In order to improve both precision and recall, we have modified the vocabularies and the grammar, as explained later in the manuscript.

2.2 Synonym Generation

The use of protein name synonyms represents a major challenge in the tagging of proteins. The lack of standards in research writing allows referencing a single protein in multiple ways, producing many-to-many relationships. For example: ppif, cyclophilin 3, cyclophilin III and cyp3 are all valid references to the same protein, immediately understandable to the researcher, but requiring extra effort to catalogue for information extraction. Making protein name recognition robust requires collecting all available synonyms. We have created our synonym list by integrating synonyms from Swiss-Prot (http://us.expasy.org/sprot) and Entrez (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene). However, even at 5 synonyms per protein on average, it does not provide full coverage. OTM still encounters abstracts where a protein is referenced by a name that isn't in our database. In some cases this is merely a formatting issue, involving arbitrary placement of dashes or spaces (tgf-beta vs. tgfbeta) or capital/lowercase lettering (TGF-beta vs. tgf-beta). We diminish this simple problem in the tagging stage by using a basic approximate string-matching algorithm, which ignores dashes, case, commas and spaces. OTM also produces additional synonyms by expanding arabic into roman numerals. However, in cases where authors use a yet undocumented synonym, such matching is not possible (tgf-b vs. tgf-beta).

Further hampering the generation of useable synonyms is the use of English language words in the creation of protein names. This is pronounced in numerous cases, such as 'in', 'and', 'pre', etc. These synonyms result in false positive matches and interfere in some cases with the tagging of other words necessary for parsing. Another unfortunate byproduct is the unnecessary examination of articles that likely have little to do with an interaction. There are several methods available to isolate and remove such overlaps with the English language, e.g., using a non-biological corpus.

Tagging a corpus that is unlikely to contain protein names, such as the Reuters or Brown corpus, and tracking incidences of retrieved matches provides useful indicators of false positives. The following examples were found with a fair amount of regularity (numbers correspond to word incidences in the Brown corpus: 'of' (5,210), 'in' (3,262), 'for' (1,881). For practical purposes, this is culled into a more accurate list by human review. The tagger then uses this list, and identifies false positives by capitalization and spacing features, allowing words more likely to be synonyms ('OF", 'FOR', etc.) to still be recognized.

In addition to tagging a corpus outside of the biomedical domain, the tagging of PubMed and counting word incidence can yield useable lists of impractical synonyms. In addition, the first incidence of a word within PubMed can be indicative of it's usefulness as a synonym. Words discovered earlier in PubMed can often be discarded, as they are fairly often generic English or biomedical terms. This approach again results in a human produced list of possible false positives, which is combined with the list produced from non biomedical corpus.

2.3 Abstract Retrieval

Although abstracts are less rich compared to a full article, they often contain the paper's most important interaction information [22]. OTM indexes PubMed abstracts using relevant MeSH (Medical Subject Headings) terms and synonyms found in each one. Abstracts with PPIs are printed out in either html or a tab-delimited text file, with their corresponding PubMed ID and the phrase that was detected.

2.4 Tagging

The parser relies on being limited to parsing single clauses, as the subject of a separate clause might be elusive. Thus, at the tagging stage, sentences and clauses are treated as separate entities and are parsed individually. Sentences are recognized as ending with a period followed by a space, and clauses are identified as being separated by a series of phrases called subordinating conjunctions, e.g., *after, although, because*, etc.

After separating a document into sentences, OTM uses several dictionaries to tag words and phrases for parsing with its hierarchical pattern recognition parser. They are divided into three primary categories: molecule names, interaction keywords, context indicators and category for miscellaneous nouns and verbs. As in Temkin and Gilder's context-free grammar (CFG) [25], unrecognized words are ignored.

To address issues of protein name formatting discussed in the synonym retrieval section, an approximate string matching algorithm was used. In the first stage of identifying words, OTM ignores commas, spaces, dashes, and uppercase/lowercase formatting. Beyond this, the algorithm is essentially an optimized naïve method that searches a tree of possible matches. While progressing character by character through the sentence string, if a match is found, the string must immediately terminate with a space or punctuation, or be considered a partial match and ignored. In addition, preference is given to longer matches.

To improve performance, tagging is performed hierarchically, as some protein names may overlap with other parts of the tagger's vocabulary:

1. Proteins – Synonyms for the proteins being searched by.
2. Keywords – Protein interaction keywords (see Table 1)
3. Miscellaneous nouns and verbs.
4. Context indicators.

For example: 'Protein Kinase A' could have the letter 'A' identified as the grammatical determinant 'a' as in '*Thyrotropin-dependent Complex Formation of Protein Kinase A Catalytic Subunit with IKB*'. Some protein names overlap with keywords (e.g., TAT-*binding* protein). Having all protein names tagged in advance eliminates this possibility, and also increases the need to ensure minimal synonym overlap with general English language words, particularly context indicators.

2.5 Parsing

We have initially implemented the parser using CFG from [25], which identified sentences containing any interaction, regardless of the proteins involved. Our intention was to alter the grammar to focus on specific interactions, and to achieve both a high precision and high recall. Initial modification of the grammar provided less than adequate results, largely due to ambiguities, and the difference of aims of each project.

Parsing of tagged abstracts is performed by layered pattern matching (Figure 2). Each successive layer serves to group information into more easily managed tags for processing by the next layer. For example, *'Molecule A AND Molecule X'* is a set of tokens that can be combined into a single tag, *'Group A'*, which represents groups of molecules containing *Molecule A.* This overcomes the difficulties encountered with using a CFG, in that rules are constrained to certain levels of parsing, removing the ambiguities that created problems with the original implementation. For example, the phrase 'KIAA0380 and LARG could bind plexin-B1' would be misinterpreted by a CFG, resulting in a parse that recognized the interaction as only 'LARG could bind plexin-B1' excluding KIAA0380. In addition, patterns were produced with automated tagging in mind, and in some cases account for errors in the tagging process.

Table 1. Common keywords. Words are also identified in other grammatical forms. For example, 'acetylate' is also recognized by 'acetylated', 'acetylates', 'acetylating' and 'acetylation'.

KEYWORDS	
acetylate, (co)activate	produce, modify
transactivate	impair, inactivate
associate, add	interact, react
bind, link	disassemble
catalyze, cleave	discharge, modulate
coimmunoprecipitate	substitute, dissociate
demethylate	ubiquitinate
dephosphorylate	heterodimerize
methylate	heterotrimerize
phosphorylate	immunoprecipitate

Molecule names: All molecules are identified via an approximate string matching algorithm.

Interaction Keywords: These words (verbs or nouns describing an interaction), are largely based on the list from [25]. Some keywords were removed from the original set as they were not suitable for our own definition of an interaction (e.g., 'cleaves', 'expressed', 'severed'). Additional tags (e.g., 'heterodimerize', 'coactivate') were added during the human review process upon observation of interaction phrases. Each tag maintains the grammatical tense/type of verb used. This expands on the original vocabulary [25], which treated each interaction keyword the same way, regardless of grammatical tense. Our approach makes it easier to distinguish between several types of phrases, which in turn allow more complex sentences to be correctly parsed. This is necessary for most of the rules involved in the earlier stages of parsing, and a parse of the testing set using only one tag for all verbs resulted in a significant increase in the rate of false positives with no increase in recall (56% accuracy, 47% recall).

Context Indicators: Words such as 'the', 'and', 'to' are commonly referred to as a closed class in parts of speech identification. While interaction keywords are used to give a phrase general meaning, closed class words specify contexts that make that meaning more specific.

The selection of phrases is performed by a hidden Markov model (HMM) trained only on phrases containing interactions. The emission states of the model are used to tag the beginnings and ends of phrases. The emission at each stage of the model is not weighted as in a traditional HMM; instead, all probable phrase endpoints are weighted by the probability of the region they envelop, based on a bi-gram Markov model trained with the same data:

$$(\sum^{i=0}_{i\leq n} P(a_i \rightarrow a_{i+1}))/\#S,$$

where $S=\{a_1, a_2, \ldots, a_n\}$ is the sequence of tokens/tags produced by a given sentence; $P(a_i \rightarrow a_{i+1})$ is the probability of token a_i being followed by token a_{i+1}.

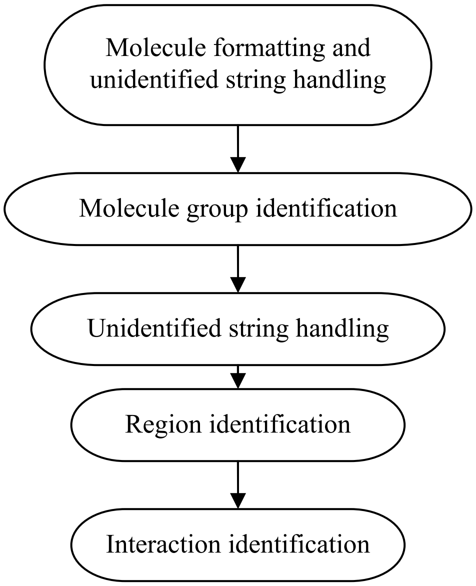
Molecule formatting and unidentified string handling
Molecule group identification
Unidentified string handling
Region identification
Interaction identification

Figure 2. Layered pattern matchers.

Molecule formatting and unidentified string handling – Organizes molecules into discreet units, and resolves some common issues regarding unidentified molecules.

Molecule group identification – Identifies and tags proteins that function as a single grammatical entity.

Unidentified string handling – Resolves common issues regarding unidentified molecules.

Region identification – Identifies references to domains and regions and binds them to their respective proteins.

Interaction identification – Identifies PPIs.

Out of possible combinations of endpoints in a sentence, the most probable arrangement is selected. This is particularly useful in sentences containing multiple interactions, where several parts of the sentence may be considered valid endpoints for interaction phrases, based on the tri-gram model. The value of each phrase is defined by an average as opposed to a true probability to account for differences in phrase length.

The use of a tri-gram model for initial phrase tagging and a bi-gram model for its evaluation proved optimal, since lower values for either resulted in excessive drops in precision, and higher values required far greater system resources with no substantial increase in either precision or recall.

Parsed results can be output at any stage into a human-readable plain-text file, with a format that is interchangeable with the files used to program the parser itself. If a file contains an interaction, the corresponding record in the local abstract database is updated, flagging it as containing an interaction, and recording the sentence found. The output to tagged plain-text was done to ensure versatility, and to provide as much opportunity for the iterative growth and debugging of the parse grammar as possible. The reviewer could at this stage simply copy a phrase from the output and insert it into the pattern matching configuration, or add a new keyword into the dictionaries. It also allows for the use of regular expression text processing to perform tasks on the parse grammar. An effort was made during development to isolate basic rules, most of them relating directly to English grammatical rules, and separate them from the actual interaction phrases.

The final results from OTM are records comprising the pairs of interacting synonyms combined with the PubMed ID's of the abstract in which they were found, a record of the interaction phrase, and a record of the parsing process that resulted in the interaction being identified. The synonyms can be linked to their corresponding SwissProt ID's, allowing more precise identification and comparison to PPI databases.

3 RESULTS AND ANALYSIS

3.1 Evaluation

OTM was tested on an expert reviewed corpus of 2,093 randomly selected abstracts from the superset of abstracts retrieved from PubMed for the purposes of validating OPHID predictions. It was observed that only 253 or 12% of the articles containing protein pairs actually contained interactions. Articles retrieved were found to contain synonyms referring to a pair of proteins predicted to interact by OPHID. To ensure good estimate of OTM's actual performance, the corpus was randomly split into testing and training sets (998 and 1,095 abstracts), with an approximately equal proportion of abstracts containing an interaction and abstract without an interaction per set.

To measure OTM's performance we used precision and recall, as defined bellow [23]:

$$\boldsymbol{Recall} = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$\boldsymbol{Precision} = \frac{TruePositive}{TruePositive + FalsePositive}.$$

True positive results were defined as abstracts that contained interaction phrases recognized by both OTM and the human reviewer. False positives comprised articles that did not contain a specified interaction but had phrases improperly identified as containing interactions. False negatives contained interactions that were not recognized by the OTM.

The current version of OTM achieves 93% precision and 47% recall on our testing set. Figure 3 suggests that increasing the size of the training set minutely affects precision, but has dramatic effects on recall.

To further improve recall, we have analyzed characteristics of false negative results. The largest group (65%) comprised phrases either not yet seen in the training set, or too complex to be handled. Improper tagging of molecules, where one of the two molecules searched by was not present resulted in 22% of false negatives. Finally, words not yet seen in the training set resulted in 7%, and interaction phrases spread among multiple sentences in 6% of false negatives.

False negative results are mostly due to sentence structure, and to a lesser extent improper tagging. Some interactions are split between two sentences, where in the one containing the interaction phrase, the protein in question is only implied as the subject. The more common case is simply the variety of ways an interaction can be stated using the English language. Commonly, this occurs because of compound or complex sentences, which contain multiple clauses and pronouns. For example "Molecule A interacts with Molecule C but under other conditions it interacts with Molecule B" or "Molecule A interacts with Molecule C and was not bound to phosphorylated Molecule B" would require far deeper understanding of the general English language. Since the OTM is limited in its language domain to protein interactions, certain constructs in the English language will elude it. Clearly, further training examples will help to reduce false negatives.

False positive results occur for similar reasons; however, the ambiguity of the English language often plays a much larger role. For example, the grammatical function of the word 'associated' is different in the cases: "BRCA1 and BRCA2 associated" and "BRCA1 and BRCA2 associated carcinomas".

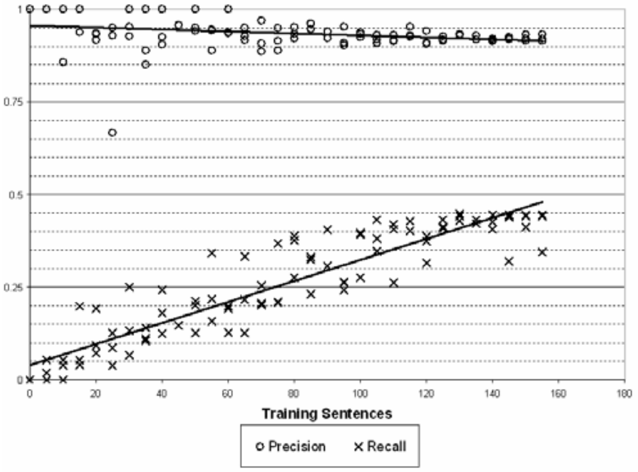

Figure 3. Precision and recall of OTM, as validated on an independent set of 998 abstracts. Three samples were randomly selected from the training set at each number of training sentences.

3.2 Results

OTM was used to extract human PPIs from PubMed (2005 baseline database), and resulted in 330,000 interaction phrases. 1,057 verified PPIs from OPHID. Of those, 50 were not previously documented, and are supported by 233 abstracts.

To further evaluate OTM's performance, we manually analyzed 80 of the novel interaction phrases returned from its overlap with OPHID, and found it retained an accuracy of 78%. In addition, we compared our results to several large manually curated resources (see Table 2). Manually curated PPI databases review full-text articles, which contain much more interaction information than the abstract itself, giving them a distinct advantage over OTM's abstract searches. To identify the overlap, the interacting pairs for each database were translated into SwissProt format to match the output of OTM. Totals represent PPI pairs that exist both in OTM's abstract database and the database in question.

Table 2. Comparing DIP, MINT, HPRD and OTM based on articles retrieved per interaction. Individual versions are as follows: MINT:2005-02-02, DIP:2004-02-11, HPRD:2004-09-30, BIND:2004-02-11.

Human Curated PPI Databases				
	HPRD	**BIND**	**MINT**	**DIP**
Unique PPIs	12,272	5,737	3,219	989
Evidence	12,398 Articles	7,531 Articles	3,354 Articles	1,102 Articles
Articles/PPI	1.0	1.3	1.0	1.1
OPHID				
Unique PPIs Overlap	875 (7%)	263 (5%)	164 (5%)	159 (16%)
Evidence	2,919 Articles	1,228 Articles	749 Articles	994 Articles
Articles/PPI	3.8	5.0	5.0	6.4

The relatively small overlap seems indicative of two features: OTM's recall, and the distribution of interaction phrases between abstracts and full text papers. Corney *et al.* similarly discovered that their recall was cut by half when their searches were restricted to abstracts as opposed to full-text papers [5]. The total numbers of articles found by OTM indicate that many of the human curated databases are not comprehensive. The discovery of multiple references to a single interaction may seem unnecessary at first, but a review of some manually extracted interactions suggests that this approach can advantageously be used to increase confidence, and thus precision. There is also evidence that regardless of precision achieved by the automated text mining system, the overall "quality" of discovered interactions cannot be guaranteed, and human review is beneficial filtering step.

The automated discovery of protein-protein interactions in literature, especially with high precision, is beneficial, directing researchers to articles referencing interactions they are researching or that they have predicted. High precision and full automation also make it possible to verify larger scale experiments and predictions with a greater degree of confidence in situations where the quantity of data makes it impractical to review manually (for example, OPHID with almost 50,000 human PPIs). Increasing the recall would make the body of data retrieved by text-mining more comprehensive, and in turn more useful to researchers.

3.3 Performance

OTM was run as a multi-threaded application on an IBM server under Linux, running 4 3.06 GHz Intel® Xeon™ CPU's with 1 Gb of RAM. OTM processes PubMed in individual files containing 30,000 abstracts each, with each file taking just over 30 minutes to tag and parse (in the local PubMed distribution, files representing articles entered earlier in PubMed's history may not have abstracts, resulting in faster processing times), and roughly a week to process all of PubMed. This highlights one of the major advantages of our method over human curation. As previously discussed, manual analysis of PubMed records for protein-protein interactions is not practical, while automated methods such as our own can cover the entirety of the literature in a relatively short time. Text mining still cannot replace a team of expert curators, but fine-tuning the information extraction system to either providing high precision or high recall enables useful applications. The first scenario is best suited for validation of predicted PPIs, while the second scenario is useful for integrated data mining and interpretation.

4 FUTURE DIRECTIONS

Our focus is on high precision retrieval. To improve OTM's recall without sacrificing precision, we are developing new approaches, mainly by expanding the diversity of sentence structures recognizable by the parser through additional training. However, continuous expansion of PPI phrases from the training set is a time-consuming process, and limited in its application. Some interesting headway has been made into automating it through the use of local alignment algorithms [12], but given the almost limitless variety of the English language, simple pattern matching may still fall short.

Continued development will focus on including a general English parser that can be adapted to finding interaction phrases. Many of the rules involved in the earlier stages of parsing need only to be mapped onto English language terms. Groups of molecules and phrases describing molecules are merely noun phrases. Producing parse rules on the same principles, and then applying domain knowledge could expand the usefulness and versatility of such a system. WordNet provides a vast amount of information [4], which will be used to expand our current dictionaries to better encompass the rest of the English language.

Improper tagging is another major problem, and can be handled in a number of ways. More complete synonym lists are only a partial solution, as long as the researchers use inconsistent naming conventions. Without standardization in the research world, this will continue to be a challenge. Tools such as BLAST, an alignment application devised for comparing strings of DNA, have proven successful as a means of identifying unrecognizable molecule names [16]. Customizing the algorithm for the task of protein identification could be taken further by assigning domain specific replacement rules, such as: b → beta, 2 → II.

Also of interest is the use of abstracts versus full text articles. Systems such as GENIES were tested on full articles, which contained multiple instances of interaction phrases, providing more possibilities for interaction extraction. PubMed now has the option of automated download of full articles, creating a richer data set, and great promise for improving recall.

ACKNOWLEDGEMENTS

The authors would like to thank Kathryn Harris and Dan Strumpf for their assistance in reviewing PubMed abstracts, which have made larger scale validation of OTM possible. We thank Graeme Hirst for a non-biological corpus. This research was supported in part by DOD grant #OC040038 and NSERC grant #203833. We acknowledge the hardware and software support from IBM Life Sciences through a Shared University Research Grant. Access to PubMed database was gratefully received from National Library of Medicine.

REFERENCES

[1] G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson and C. W. Hogue, *BIND--The Biomolecular Interaction Network Database*, Nucleic Acids Res, 29 (2001), pp. 242-5.

[2] M. Barrios-Rodiles, K. R. Brown, B. Ozdamar, R. Bose, Z. Liu, R. S. Donovan, F. Shinjo, Y. Liu, J. Dembowy, I. W. Taylor, V. Luga, N. Przulj, M. Robinson, H. Suzuki, Y. Hayashizaki, I. Jurisica and J. L. Wrana, *High-throughput mapping of a dynamic signaling network in mammalian cells*, Science, 307 (2005), pp. 1621-5.

[3] K. R. Brown and I. Jurisica, *Online predicted human interaction database*, Bioinformatics, 21 (2005), pp. 2076-82.

[4] A. Budanitsky and G. Hirst, *Evaluating WordNet-based Measures of Lexical Semantic Relatedness*, Computational Linguistics, 32 (2006).

[5] D. P. Corney, B. F. Buxton, W. B. Langdon and D. T. Jones, *BioRAT: extracting biological information from full-length papers*, Bioinformatics, 20 (2004), pp. 3206-13.

[6] I. Donaldson, J. Martin, B. de Bruijn, C. Wolting, V. Lay, B. Tuekam, S. Zhang, B. Baskin, G. D. Bader, K. Michalickova, T. Pawson and C. W. Hogue, *PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine*, BMC Bioinformatics, 4 (2003), pp. 11.

[7] C. Friedman, P. Kra, H. Yu, M. Krauthammer and A. Rzhetsky, *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles*, Bioinformatics, 17 Suppl 1 (2001), pp. S74-82.

[8] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer and G. Superti-Furga, *Functional organization of the*

yeast proteome by systematic analysis of protein complexes, Nature, 415 (2002), pp. 141-7.

[9] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, Jr., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant and J. M. Rothberg, *A protein interaction map of Drosophila melanogaster*, Science, 302 (2003), pp. 1727-36.

[10] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys and M. Tyers, *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry*, Nature, 415 (2002), pp. 180-3.

[11] R. Hoffmann and A. Valencia, *Implementing the iHOP concept for navigation of biomedical literature*, Bioinformatics, 21 Suppl 2 (2005), pp. ii252-ii258.

[12] M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu and M. Li, *Discovering patterns to extract protein-protein interactions from full texts*, Bioinformatics, 20 (2004), pp. 3604-12.

[13] R. J. Ingham, K. Colwill, C. Howard, S. Dettwiler, C. S. Lim, J. Yu, K. Hersi, J. Raaijmakers, G. Gish, G. Mbamalu, L. Taylor, B. Yeung, G. Vassilovski, M. Amin, F. Chen, L. Matskova, G. Winberg, I. Ernberg, R. Linding, P. O'Donnell, A. Starostine, W. Keller, P. Metalnikov, C. Stark and T. Pawson, *WW domains provide a platform for the assembly of multiprotein networks*, Mol Cell Biol, 25 (2005), pp. 7092-106.

[14] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara and Y. Sakaki, *Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins*, Proc Natl Acad Sci U S A, 97 (2000), pp. 1143-7.

[15] R. B. Jones, A. Gordus, J. A. Krall and G. MacBeath, *A quantitative protein interaction network for the ErbB receptors using protein microarrays*, Nature, 439 (2006), pp. 168-74.

[16] M. Krauthammer, A. Rzhetsky, P. Morozov and C. Friedman, *Using BLAST for indentifying gene and protein names in journal articles*, GENE (2000), pp. 245-252.

[17] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, J. D. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill and M. Vidal, *A map of the interactome network of the metazoan C. elegans*, Science, 303 (2004), pp. 540-3.

[18] H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd and B. Weil, *MIPS: a database for genomes and protein sequences*, Nucleic Acids Res, 30 (2002), pp. 31-4.

[19] S. Novichkova, S. Egorov and N. Daraselia, *MedScan, a natural language processing engine for MEDLINE abstracts*, Bioinformatics, 19 (2003), pp. 1699-1706.

[20] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Z. Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, M. Menezes, D. R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V. Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti and A. Pandey, *Development of human protein reference database as an initial platform for approaching systems biology in humans*, Genome Res, 13 (2003), pp. 2363-71.

[21] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth and M. Vidal, *Towards a proteome-scale map of the human protein-protein interaction network*, Nature, 437 (2005), pp. 1173-8.

[22] M. J. Schuemie, M. Weeber, B. J. A. Schijvenaars, E. M. v. Mulligen, C. C. v. d. Eijk, R. Jelier, B. Mons and J. A. Kors, *Distribution of information in biomedical abstracts and full-text publications*, Bioinformatics, 20 (2004), pp. 2597-2604.

[23] H. Shatkay and R. Feldman, *Mining the Biomedical Literature in the Genomic Era: An Overview*, Journal of Computational Biology, 10 (2003), pp. 821-856.

[24] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach and E. E. Wanker, *A human protein-protein interaction network: a resource for annotating the proteome*, Cell, 122 (2005), pp. 957-68.

[25] J. M. Temkin and M. R. Gilder, *Extraction of protein interaction information from unstructured text using a context-free grammar*, Bioinformatics, 19 (2003), pp. 2046-2053.

[26] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields and P. Bork, *Comparative assessment of large-scale data sets of protein-protein interactions*, Nature, 417 (2002), pp. 399-403.

[27] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. Eisenberg, *DIP: the database of interacting proteins*, Nucleic Acids Res, 28 (2000), pp. 289-91.

[28] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich and G. Cesareni, *MINT: a Molecular INTeraction database*, FEBS Lett, 513 (2002), pp. 135-40.