Automated extraction of precise protein expression patterns in lymphoma by text mining abstracts of immunohistochemical studies

Jia-Fu Chang¹, Mihail Popescu¹,², Gerald L. Arthur¹,³

¹MU Informatics Institute, ²Health Management and Informatics, ³Department of Pathology and Anatomical Sciences, University of Missouri, Columbia, USA

E-mail: *Gerald L. Arthur - arthurg@missouri.edu
*Corresponding author

Received: 01 March 13  
Accepted: 16 April 13  
Published: 31 July 13

Abstract

Background: In general, surgical pathology reviews report protein expression by tumors in a semi-quantitative manner, that is, –, –+/+, +/–, +. At the same time, the experimental pathology literature provides multiple examples of precise expression levels determined by immunohistochemical (IHC) tissue examination of populations of tumors. Natural language processing (NLP) techniques enable the automated extraction of such information through text mining. We propose establishing a database linking quantitative protein expression levels with specific tumor classifications through NLP.

Materials and Methods: Our method takes advantage of typical forms of representing experimental findings in terms of percentages of protein expression manifest by the tumor population under study. Characteristically, percentages are represented straightforwardly with the % symbol or as the number of positive findings of the total population. Such text is readily recognized using regular expressions and templates permitting extraction of sentences containing these forms for further analysis using grammatical structures and rule-based algorithms.

Results: Our pilot study is limited to the extraction of such information related to lymphomas. We achieved a satisfactory level of retrieval as reflected in scores of 69.91% precision and 57.25% recall with an F-score of 62.95%. In addition, we demonstrate the utility of a web-based curation tool for confirming and correcting our findings.

Conclusions: The experimental pathology literature represents a rich source of pathobiological information, which has been relatively underutilized. There has been a combinatorial explosion of knowledge within the pathology domain as represented by increasing numbers of immunophenotypes and disease subclassifications. NLP techniques support practical text mining techniques for extracting this knowledge and organizing it in forms appropriate for pathology decision support systems.

Key words: Natural language processing, protein expression, text mining

BACKGROUND

The interpretation of histologic sections of tissue biopsy material requires the evaluation of innumerable visual features in order to properly classify disease processes and, in turn, establish a prognosis and appropriate therapeutic plan. This process requires great domain knowledge and cognitive resources to accomplish. The sources of this knowledge include training visual interpretive skills through exposure to multiple examples of histological...
changes associated with specific pathologic processes, didactic instruction, and critically, the extraction of new information derived from experimental studies as reported in the literature. As a result of the explosion of knowledge reflected in the remarkable increase of research and review manuscripts and increasing numbers of professional journals, it has become evident that this high-dimensional flow of new information is beyond the ability of humans to adequately process and curate. As a consequence, text mining techniques have been developed to computationally extract and organize data in a form usable by biomedical professionals.

The techniques of natural language processing (NLP) for the identification and extraction of information have been widely applied in the biomedical domain and numerous articles have been published discussing the basic concepts and methodology.[1-4] In the pathology domain, text mining research has been primarily directed toward the use of pathology tissue reports as text corpora.[5-10]

We have explored the application of text mining to the extraction of information relating to protein expression by tumors as determined by immunohistochemical (IHC) staining. Our text mining task is unique not only in terms of application to the pathology domain, but also to our specific targeting task: A combination of expression relationship extraction and associated numerical statistics defining the expression level. To this end, we apply a collection of rules based on templates and keywords that is easy to comprehend and use and can be expanded and modified over time.

The construction of a knowledge base containing accurate rates of protein expression by highly specific lymphoma types will permit improved classification schemes resulting in more precise estimations of probability within differential diagnoses. Such information would provide a significant improvement over the frequent use of semi-quantitative estimations of protein expression in reviews of IHC studies. For example, a common format is the following scale: Negative, expressed by less than 5% of tumors examined; −/+ , protein expressed by 5-25% of tumors examined; +/− , expressed by 25-50% of tumors; and +, expressed by greater than 50% of tumors.[11,12]

It should be noted that precise expression data that is statistically actionable has been employed in some studies.[13] However, the retrieval and organization of this detailed information is extremely labor intensive and of necessity, limited to very specific diagnostic disease classes. Compiling comprehensive protein expression databases will require computer aided information extraction methods. The difficulty of extracting the correct relationship between tumors and proteins in terms of expression varies primarily with the author’s style of writing. Those with a more direct syntax utilizing frequent subject, predicate, and object constructions with sparse use of qualifying clauses present a more straightforward analytical process. By contrast, authors favoring more complex grammatical constructions present significant difficulties for automated knowledge extraction.

MATERIALS AND METHODS

Dataset

The dataset consists of abstracts obtained from PubMed covering “lymphoma” and is limited to English language articles. The training set is limited to abstracts containing the medical subject headings (MeSH) term “follicular lymphoma” as the major topic and up to year 2010. The test set uses keywords of “lymphoma” and “immunohistochemistry” as MeSH terms in abstracts up to the year 2011. The test set excludes any abstracts overlapping with the training set. The training set is designed to include every aspect of follicular lymphoma so we can build up rules and templates pertinent to immunohistochemistry. The training and test sets contain 1794 and 249 abstracts respectively. Validation of the results was performed by a pathologist using custom-designed web software which highlights the sentence of interest and the text mining results.

The aim of the study is to obtain protein expression data from the literature in the form of tuples which provide information about the disease and protein as well as expression level. A tuple is defined in the form of “disease, protein, expression level”. In other words, it structures the description of protein expression into a succinct form that can be stored and manipulated. For example, the sentence, “Eighty-three percent (83%) of anaplastic large cell lymphomas (ALCL) express ALK,” should be converted into the tuple “ALCL, ALK, 83%.” To extract correct expression relationships between disease, protein and expression level, we developed rules based on syntactic and proximity information. The workflow of the NLP procedures is depicted in Figure 1. The abstracts are fed into the preprocessing block which includes functions that perform sentence detection, regular expression identification of numbers and matching of the named entities and templates designed for the domain. After preprocessing, link grammar parser parses the abstracts sentence by sentence and we apply the rules aggregated from linkage, constituent tree and finding neighboring tokens rule sets. A final step of semantic filtering ensures the extracted tuple relates to protein expression. In the next sections, we will describe each processing block in more detail. Specifically, our methods take advantage of characteristic forms of numerical representation of the incidence of protein expression, which typically assume a limited variety of structures including, for example, “10 of 25 cases” or “40% of cases.” This permits the use
of regular expressions to identify sentences of interest and narrow the amount of text to be further interrogated using a variety of NLP techniques. A similar approach taking advantage of a limited dictionary of forms of expression of essential information about enzyme kinetics was employed by Heinen et al.\(^\text{[14]}\). Caporaso et al.\(^\text{[15]}\) relied upon regular expressions to identify point mutations in genes.

Potential Expression Level Extraction
Regular expressions offer a basic and highly flexible mechanism for the identification of specific forms of text and they form the cornerstone for identifying expression level in our text mining algorithms. We begin with recognizing numbers in the text by applying regular expression crafted for numerical number extraction. The design to capture number information is to match consecutive numerical digits (0, 1, 2,..., 8, 9) in text. In addition, there may be one decimal point among the numerical digits depending upon the precision of the expression. The written form of number is identified by a preset dictionary of English numbers. After numbers are identified, we attempt to recognize basic forms of representation of percentages using predefined templates. Templates are designed based on the observation of commonly encountered domain-typical word usage patterns found in the training set. At this stage, regular expressions and templates reflect linguistic structures based on pathology domain entities that demonstrate predefined word sequence orders. For example, the explicit form consisting of simple integers followed by the percentage sign and the implicit compound form of two integers stated in several forms such as 10 of 30 or 10/30. It should be noted that percentage-like forms of representation are encountered in multiple contexts other than those related to protein expression, for instance, survival statistics, sensitivity, specificity, etc. The exclusion of these false positive (FP) findings is achieved through the use of rules based on semantic filtering, which will be discussed later.

Templates and Domain Specific Entities
Other applications of templates permit the identification of particular common pathology phraseology. In particular, the complex disease entities are defined as subclasses of primary diseases by particular combinations of protein expression such as “CD5-CD10-follicular lymphoma”. As pathology knowledge becomes progressively refined and sub-classifications are based not only on protein expression but also by emerging features such as microarray determinations of RNA and miRNA expression, as well as the next generation sequencing features are recognized, it will be imperative to be able to encompass such complex disease entities. Templates also enable the recognition of common parallel grammatical constructions which, when identified, yield several protein-disease expression pairs from one sentence.

Dictionaries
Named entity recognition (NER) was focused on the identification of names of lymphomatous disease and
proteins. The disease name dictionary is extracted from the National Cancer Institute (NCI) thesaurus\cite{16} and protein names from the UniProt Knowledge Base\cite{17} limited to Homo sapiens and is also customized through domain expert input. In addition, many pathology immunohistochemistry articles use the terminology of clusters of differentiation (CD) which has been developed and widely implemented by the flow cytometry community, but are sometimes not included in the UniProt database. Therefore, a comprehensive name dictionary of CD antigens is also imported into the protein dictionary.\cite{18} All pertinent aliases of a given named entity entry, including synonyms, acronyms, etc., are included. The attributes by which diseases and proteins are linked require a separate dictionary encompassing such terms as “expresses”, “is positive for”, “is stained by”, and others. This is constructed as a customized keyword list largely reflecting domain knowledge.

Grammatical Parsing

**Link Grammar Linkage and Hierarchical Output Forms**

The link grammar parser (LGP),\cite{19} originally developed at Carnegie Mellon University, is an English language syntactic parser based on link grammar, an original theory of English syntax. LGP has been proven to work in various biomedical text mining tasks\cite{20-22} and has evolved into an open source project that is updated continuously. A recent update includes a better adaptation for biomedical text (BioLG).\cite{23} There are two types of parsing outputs from LGP: Linkage and hierarchical constituent output, both are rich in syntactic structure information. We store the two outputs for each sentence from the literature in the database and preserve their original structure forms in connected linkages and constituent tree format respectively. While most biomedical text mining methods adopt only the connected linkage or constituent tree formats of LGP output, our method implements customized sets of rules from both outputs allowing us to extract information from the most informative form.

**Rule Matching on Linkage**

Linkage rules have been adopted in biomedical text mining for relationship extraction of molecular interaction events.\cite{20} Active participants in the sentence can be found through certain combinations of links that connect word tokens. Application examples in the biomedical domain include finding participating proteins in biochemical interactions and protein-protein interactions.\cite{21} Using grammatical linkage rules permits the extraction of protein expression information with varying degrees of success depending upon the linkage. For instance, subject (S) and object (O) linkage connection via verb (V) between the appropriate named entities are very reliable, in our experience and that of others.\cite{22} Useful information can also be obtained based on the elucidation of noun phrases and numerous other grammatical linkages. We have constructed a number of heuristic rules to encapsulate the more reliable forms of syntactic structure as applied to varying styles of writing. To satisfy the linkage rules, the linkage must match at least part of the syntactic structure in the sentence. The word tokens visited by the linkage rule are marked and collected. These tokens are treated as words with potential relationships.

In Figure 2, linkage rule 1 deals with the active voice and can utilize the simple construction of subject, object, verb (S + O) in order to correctly extract the information that “lymphomas, express, immunoglobulin.” Rule 2 accommodates interpretation of the passive voice.

![Figure 2: Select linkage rules. Description: Rule 1 shows a straightforward connection between subject (lymphomas) and object (immunoglobulin). The verb “expressed” between S link and O link confirms the desired semantics of the sentence. Word tokens visited by the rule, “lymphomas”, “expressed” and “immunoglobulin”, are collected. Rule 2 also collects the verb “expressed” but demonstrates the passive voice. “Caspase-3”, “was”, “expressed”, “in” and “%” are collected by the rules. While the tokens relevant to the expression tuple consist only of the protein “Caspase-3,” the remaining tokens can act as links to the implementation of other rule sets in the rules aggregation stage.](image-url)
but utilizes a different grammatical linkage, $S + P + MV + J$, retrieving the bit of knowledge, “caspase-3” and “expressed.” To be more specific, the “S” link leads the subject (Caspase-3) to find the be verb (was).

The following “P” link connects be verb (was) to passive participle (expressed). The last two “MV” and “J” links in turn form the connection between the word “expressed” and the modifying phrase up to “%”. Note that both examples require the implementation of additional rules to identify the complete informational tuples which provides knowledge of the disease, protein expressed and the frequency of that expression.

**Rule Matching on Hierarchical Constituents**

The hierarchical phrase constituent is the other form of output generated by the link grammar parser. Here, word tokens are grouped into grammatical phrases, which are related to one another in the form of a tree structure. The word tokens within phrases at various levels of the tree have natural semantic relationships due to the inherent structure of grammar. By traversing through the constituent tree, both inter- and intra-phrase relationships between word tokens can be established. Rules related to the analysis of constituent trees are applied as SQL queries over trees stored in a database. The rules first identify a word token of interest through NER of proteins and diseases and then identify neighboring phrases through the use of the parent child relationships. Any word tokens within these phrases will be considered to be tokens likely to be participants in the relationships of interest. The basic constituent rules are as follows: For every token of interest (protein and disease names), find tokens within phrases that either contain the token itself or are the parent’s sibling, child of parent’s sibling, or grandparent’s sibling. For instance, in Figure 3, the noun phrase “detectable $P53$ expression” clearly establishes $P53$ as being expressed and traversing the constituent tree according to our rules leads to the percentage, 85.7%, which is a child of parent’s sibling node. The words found by constituent rules are collected for aggregation and further processing by other rule sets.

**Proximity Rules**

One approach to identifying semantic relationships is by co-occurrence, which suggests a relationship between terms when entities of interest coexist within a certain scope of the text. The scope may include co-occurrence within the same sentence, abstract or entire paper. Co-occurrence often provides the substrate for more sophisticated text mining methods. We build on co-occurrence with rule matching, which uses heuristics based on word order and syntactic structure. Proximity rules establish a window of predefined size centered on the word token of interest in order to analyze neighboring words. The size of the window varies with the length of the sentence and it is anticipated that words within the window are highly likely to convey additional information. Proximity rules compensate for link grammar’s limitations in analyzing complex sentences, particularly when multiple number representations are present. Tokens within the window are assigned distance measures based on a schema, which weights the intervening parts of speech. Stop words and punctuation marks are assigned different weights since stop words tend to connect neighboring words while punctuation marks indicate a higher degree of separation of meaning by forming sub-clauses. For example, as illustrated in Figure 4, “41%” is more likely to be associated with diffuse large B-cell lymphoma than with Burkitt lymphoma since the latter

---

**Figure 3: Hierarchical constituent rules.** Description: Here, the token of interest is the protein “$P53$” contained within a noun phrase which serves as the index node. The nodes to be further queried are indicated by the labels. In this manner, the potential expression level “85.7%” is found in the “child of parent’s sibling node”, another noun phrase. Abbreviations: PP (prepositional phrase), NP (noun phrase), VP (verb phrase), S (clause)
is separated by a comma and the former by a preposition. Furthermore, a shorter weighted path indicates a greater likelihood of finding a relationship between words. All word tokens inside a given window are evaluated before implementing the final rule aggregation.

**Rule Aggregation**

Applying each set of rules together will generate a collection of word tokens that share meaningful relationships and enable the construction of the final tuple, which combines the three facts of disease, protein, and percentage of expression. To this end, candidate expression tuples are chosen after evaluating the weighted edges between disease token, potential expression level token, and protein token. If more than one candidate tuple is found for a given expression level, we choose the tuple that has the shortest weighted path unless there is grouping involved. The process of applying each set of rules and aggregation is given in Table 1.

This aggregation process extends the scope of the individual rule set and facilitates the connection of tokens that are related remotely even if not identified by a single rule set. In Figure 5a and b, both linkage and constituent rules capture the percentage representation and the protein. However, it is the proximity rule in Figure 5c that finally connects these to the disease so that a complete tuple can be extracted. The tuple extraction is based on the premise of co-occurrence such that at least one potential expression level and protein are located in the sentence. There may be cases in which the disease name still needs to be identified even after the application of the aggregated rules in which case the following steps are applied: (1) Try to locate the disease name at the local sentence outside of the aggregated rules; (2) find the disease name within the previous two sentences, starting from the nearest one; (3) if there is no disease name found in steps 1 and 2, seek the disease name in the title.

**Grouping Correction Following the Rules Matching**

If the final tuple has any entities in an identified group, which can be recognized during the process of template matching, the tuple will be corrected. Entities in the group will be assigned to the corresponding expression level percentage according to the order of appearance. In the case of only one percentage with a group of entities, the percentage will be shared among the whole group.

**Semantics of Keywords to Identify Expression Relationship**

In the task of extracting expression relationships, knowledge of the IHC domain is fundamental in constructing the keyword dictionary. We have utilized this specialized knowledge to stratify terminology by relative degrees of certainty that the term applies to the IHC identification of protein expression in a disease state. This has been formalized by associating tuples with keywords with varying characteristics as follows:

**Table 1: Aggregation of the rules**

For each sentence \( \in \) the abstract
apply linkage rules
for each token \( \in \) the sentence
if (token \( \in \) {protein name or disease name})
apply constituent rules
if (token \( \in \) {candidate expression level number})
apply proximity rules
end for each
Rules aggregation:
\[
\text{arg min } (D(G,v,v')) \quad v \in V, v' \in V'
\]
end for each

\(D(G,v,v')\): Weighted distance between \( v \) and \( v' \), \( G \): Graph constructed by tokens satisfying the rules, \( V \): vertices, or tokens in the sentence representing potential expression level, \( V' \): Vertices, or tokens in the sentence representing protein or disease names. Collect the words from each rule set and select the final potential tuple

**Figure 4: Proximity rules. Description:** In this proximity window of size 5, “41%” can be connected to two disease names with differing weighted distances depending upon the intervening token types. However, diffuse large B-cell lymphoma is preferred over Burkitt lymphoma because the former has smaller weighted distance value.
1. Expression – Explicit
   The keywords in this category are expected to effectively predict description of protein expression events and include such terms as expression, over-expression, staining, immunohistochemistry, etc.

2. Expression – Implicit
   These keywords are usually associated with expression events, but not definitive. This includes terms such as antigen, protein, gene, etc.

3. Expression – Technique

---

Figure 5a: Rule aggregation example step 1 – applying linkage rules. Description: "48 out of 56" is matched by the templates as a potential expression level. Applying linkage S + P + MV + J (token visited): “Expression”, “was”, “detected”, “of”, “56” applying linkage M + J (token visited): “Expression”, “of”, “YB-1”

Figure 5b: Rule aggregation example step 2 – applying constituent rules. Description: Applying constituent rules from the index token, protein “YB-1” collects the tokens “Expression” (under grandparent’s sibling phrase), and “of” (under grandparent phrase) Constituent rules from disease “NHLs” (non Hodgkin’s lymphoma) do not collect any tokens as no token is found by the constituent rules

Figure 5c: Rule aggregation example step 3 – applying proximity rules. Description: The templates capture the potential expression level number description “48 out of 56”. Starting from the number within the proximity window, the nearest disease name according to the proximity rule is “NHLs.” Therefore, complete tuple from this sentence is “NHL, YB-1, 48/56”
These keywords may provide convincing evidence of expression based upon experimental techniques other than immunohistochemistry and includes terms such as polymerase chain reaction, and fluorescent in-situ hybridization.

4. Non-expression
Keywords that are associated with numerical representations typical of protein expression, but are actually not related to expression events. This situation is indicated by words such as mutation, deletion, translocation, and survival rate.

5. Non-specific evidence
Keywords reporting experimental settings, findings and analysis such as detect, findings, shows.

Semantics Filtering
After establishing these keyword categories, we can employ them as a filter to grade identified tuples by relative level of certainty. The dominant semantic level is determined by the keyword nearest to the expression level percentage representation from either “Expression” or “Non-expression” category. If keywords with differing semantic levels are equally distant based on proximity rules, additional keywords are examined within the scope of the same sentence with the final classification based on the semantics in majority.

Negation Detection: NegEx
To detect the presence of expression negation events, the NegEx algorithm is employed in every sentence where an expression tuple is located. Originally designed for predicting negated concepts in the discharge summary narratives, NegEx dynamically allocates its scope of negation in a short passage of text around the concept of interest.

RESULTS
Evaluation of the Accuracy of the Tuple Extraction
The test set contains 249 abstracts of which numerical representations including both digital and written forms are found in 230 all of which were identified by our algorithm. Among the abstracts containing numbers, the templates identified potential expression tuples in 112.

To evaluate the performance of expression tuple extraction, our results in the test set were compared to a gold standard annotation performed by a domain expert. This task was supported and facilitated by a custom web interface that presented the extracted expression tuples listed as well as the entities of the tuple highlighted in the context of the particular sentence of the abstract examined. The entire abstract is displayed as well providing contextual information. Furthermore, all numbers, numerical or written, in the entire abstract are also highlighted for the evaluation of false negatives (FNs). This curation tool has proven to be very user friendly and significantly extends the ability of a domain expert to confirm the accuracy of the findings prior to inclusion in a definitive knowledge base. For every extracted tuple, the domain expert classifies the tuple as “True” (correct expression level and disease with the correct gene/protein name), “False” (both incorrect disease and incorrect gene/protein or incorrect expression level), “Partially correct (missing protein),” “Partially correct (missing disease),” or “Not related to expression.” Any tuple that captures information about other techniques or statistics other than expression is classified as “Not related to expression.”

The protein expression tuple extraction methods that we present achieve a precision of 69.91% and recall of 57.25%. These results are based on tuples that are considered to be fully correct, that is, link the correct disease and protein names with the correct percentage of expression, a true positive (TP) result. FPs is defined as the finding of any extracted tuple where there is no proper match of any of these entities in the sentence of interest. Any valid expression tuple not captured by our algorithm is considered a FNs result.

Precision and recall are calculated according to the usual formulas:

\[ \text{Precision} = \frac{TP}{TP + FP} \]
\[ \text{Recall} = \frac{TP}{TP + FN} \]

The F-score, the harmonic mean, is derived using the following formula:

\[ F = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \]

With precision = 69.91% and recall = 57.25%, the F-score is 62.95%.

For the retrieved tuples, the distribution of sources of FPs is listed in Figure 6. Incorrectly identifying either the disease or gene/protein alone accounts for more than 50% of FPs. Incorrect identification of the disease occurs more frequently than the protein. This appears likely to result from the observation that whereas, proteins are more frequently than the protein. This appears likely to result from the observation that whereas, proteins are commonly located in the same sentence as the numerical data, this is not the case for disease names, which tend to be expressed less frequently. Therefore, there is typically a greater distance between the numerical representation and the disease name, even just in the title in some instances, resulting in a higher error rate for this component. In Table 2, we show examples of FPs. In the “missing protein” sentence, “87%” is in fact related to “classic nodular architecture” rather than a protein
Table 2: Examples of error sources for false positives

### Missing protein

Our results show that programmed death 1 positivity was found in the majority of nodular lymphocyte predominant Hodgkin lymphoma cases with a *classic nodular architecture* (87%) as compared with 50% for CD57 and was particularly helpful in identifying extranodal large atypical cells (PubMed ID: 20825974)

### Missing disease

The molecular subtype was correlated with the isotype, as 15/17 GCB cases expressed a secondary isotype (immunoglobulin (Ig) G or IgA), whereas 24/26 ABC cases expressed a primary isotype (IgM or IgD) (P<0.0001) (PubMed ID: 21233831)

### False

In PCNSL, expression of MDR1 Pgp and BCRP in tumor capillary endothelial cells was decreased in 63 and 93% of tumor areas examined, respectively, and these reduction levels differed significantly from those of gliomas (P<0.05) (PubMed ID: 21165575)

### Not related to expression

The serum anti-EBV antibody level was elevated (VCA-IgA titer≥1:40) in 6 of 12 available patients (PubMed ID: 21134987)

Missing protein: Nodular lymphocyte predominant Hodgkin lymphoma, CD57, 87%; Missing disease: Diffuse large B-cell lymphoma, (Ig) G or IgA, 15/17 and diffuse large B-cell lymphoma, (IgM or IgD) 24/26; False: PCNSL, MDR1 Pgp, 93%; Not related to expression: Non-Hodgkin lymphoma, VCA-IgA, 6/12. The keywords/phrases that cause the errors are highlighted. The falsely extracted tuples are shown in the format of disease, protein, expression level.

### Table 3: Tuple retrieval comparison

| Tuple extraction method                  | Precision (%) | Recall (%) | F-score (%) |
|------------------------------------------|---------------|------------|-------------|
| Co-occurrence                            | 24.43         | 46.38      | 32          |
| Rule sets without semantic filtering     | 52.17         | 60.87      | 56.19       |
| Rule sets with semantic filtering        | 69.91         | 57.25      | 62.95       |

Compare the co-occurrence (baseline) with the effect of applying rule sets with and without semantic filtering.

In order to evaluate the usefulness of our rule sets, we compare the final retrieved tuple results to the baseline results. The baseline is defined as the co-occurrence of disease names, protein names and valid expression level number representation within the same sentence. Therefore, the naive co-occurrence method may lose some entity names that do not locate in the current sentence. In Table 3, even before semantic filtering, the rule sets can raise both precision and recall by more than 25% and 14%, respectively. After applying semantic filtering to the tuples retrieved by the rule sets, the precision is significantly enhanced while recall experiences a minor drop. Precision apparently benefits from the semantic decision made by the nearest semantic keyword especially when some explicit keywords like “expression” are present. However, when sentences get more complicated and semantic keywords get denser within the same sentence, the semantic decision criteria could go wrong. For some tuples, semantic keywords may not even exist in the same sentence, but in the neighboring sentences, which makes the semantic decision relatively hard to make.

If we consider only the performance of valid expression level representation extraction without evaluating the associated gene/protein and disease information, the precision is 91.41% and recall is 81.25%. Here, we consider a TP to be defined as a correctly extracted instance of an expression level number representation that also is correctly identified as relating to protein expression. A FP is defined as an extracted instance of a number representation that does not indicate expression while a FN is a number representation related to an expression that was not retrieved. These results are encouraging in regard to human curation of protein expression in disease since it provides a reliable means of identifying sentences containing relevant information derived from IHC studies. The web-based annotation tool, which we have devised, would also be helpful in the review and validation of these data. The expression level alone...
can provide analytical data such as range, distribution and average that may generate useful visualization and comparison over different abstracts. In the future, we will utilize these data to create applications based on the database.

In reviewing the protein expression statistics, we noticed that there could be contradictions between correctly identified tuples. For example, the following two sentences are from different abstracts. While both talk about CD10 expression for angioimmunoblastic T-cell lymphoma (AITL), 75% and 47% are reported respectively.

“For AITL cases, the rate of CD10, BCL 6, PD-1, and CXCL 13 expression was 75.0% (36/48), 66.7% (32/48), 93.8% (45/48), and 97.9% (47/48), respectively.” (PubMed ID: 21350314).

“... diagnostic features of AITL. CD10 positivity (47%)... were common observations.” (PubMed ID: 21045384).

Expression level discrepancy is possibly the result of different experimental settings. Although it is out of the scope of this study, the background information or the contextual knowledge associated with protein expression is still important. Future work may be devoted to designing models of background knowledge to automatically choose the preferred expression statistics based on the user’s need.

**DISCUSSION AND CONCLUSION**

In this paper, we describe a pilot project for employing NLP techniques in the extraction of quantitative information relating the percentage of expression of specific proteins to specific tumors. The rule-based frame-work incorporates various templates to formulate the ultimate identification of disease/protein/expression tuples. While the levels of accuracy of our pilot project are not yet at optimal levels, we have demonstrated the feasibility of applying NLP to mining of the pathology literature to obtain specific numeric data. It appears evident that machine based information retrieval methods such as ours will be required to organize and present the enormous amount of new knowledge contained within the pathology literature. The text mining methods developed in this study are flexible and may be readily adapted to extraction of other forms of knowledge incorporating percentages, such as chromosomal translocations and oncogenic fusion genes/proteins. It is interesting to note that potential expression tuples are extracted in only 112 out of 230 abstracts that contain numbers. This discrepancy can be caused by the following factors:

1. The numbers may show values other than potential expression level, such as P value, sample size, year, demographic data, experimental values or other numerical data
2. The templates were constructed based on the training set so there may be new ways of describing potential expression level that were not included in the templates
3. The templates only apply to one sentence at a time. Potential expression level information scattered over different sentences are not captured by the templates.

Another aspect of our work is that it facilitates data retrieval in a much focused mode, which supports human curation in a more efficient manner. The tool for this provides the protein expression terms of interest highlighted in the context of the sentence in which it was identified. It is also linked to the entire abstract, which can easily be retrieved as well.

**ACKNOWLEDGMENT**

We would like to thank University of Missouri Informatics Institute for the support throughout the course of this study.

**REFERENCES**

1. Hunter L, Cohen KB. Biomedical language processing: What's beyond PubMed? Mol Cell 2006;21:589-94.
2. Krallinger M, Leitner F, Valencia A. Analysis of biological processes and diseases using text mining approaches. Methods Mol Biol 2010;593:341-82.
3. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: An introduction. J Am Med Inform Assoc 2011;18:544-51.
4. Yoo I, Alafairiet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, et al. Data mining in healthcare and biomedicine: A survey of the literature. J Med Syst 2012;36:2431-48.
5. Coden A, Savoia G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. J Biomed Inform 2009;42:937-49.
6. Currie AM, Fricke T, Gwane A, Johnston R, Liu J, Stein B. Automated extraction of free-text from pathology reports. AMIA Annu Symp Proc 2006:899.
7. Hanauer DA, Miela G, Chinnayyan AM, Chang AE, Blayney DW. The registry case finding engine: An automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. J Am Coll Surg 2010;205:690-7.
8. Liu K, Chapman W, Hwa R, Crowley RS. Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. J Am Med Inform Assoc 2007;14:641-50.
9. Yip V, Mete M, Topaloglu U, Kockara S. Concept discovery for pathology reports using an N-gram model. AMIA Summits Transl Sci Proc 2010;2010:43-7.
10. Buckley JM, Coopey SB, Shariq J, Polubriaginof F, Drohan B, Begg AK, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 2012:23:33.
11. Higgins RA, Blankenship JE, Kinney MC. Application of immunohistochemistry in the diagnosis of non-Hodgkin and Hodgkin lymphoma. Arch Pathol Lab Med 2008;132:441-61.
12. Patel DT, Rubin BP. Gastrointestinal stromal tumor: Advances in diagnosis and management. Arch Pathol Lab Med 2011;135:1298-310.
13. Rollins-Raval M, Chivukula M, Tseng GC, Jukic D, Dabbas DJ. An immunohistochemical panel to differentiate metastatic breast carcinoma to skin from primary sweat gland carcinomas with a review of the literature. Arch Pathol Lab Med 2011;135:975-83.
14. Heinen S, Thielen B, Schomburg D, KID – An algorithm for fast and efficient text mining used to automatically generate a database containing kinetic information of enzymes. BMC Bioinformatics 2010;11:375.
15. Caporaso JG, Baumgartner WA Jr, Randolph DA, Cohen KB, Hunter L. MutationFinder: A high-performance system for extracting point mutation mentions from text. Bioinformatics 2007;23:1862-5.
16. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40:30-43.

17. UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 2010;38:D142-8.

18. Human cell differentiation molecules, 2012. Available from: http://www.hcdm.org/MoleculeInformation/tabid/54/Default.aspx. [Last accessed on 2013 Feb 1].

19. Sleator CD, Temperley D. Parsing english with a link grammar. In: Third International Workshop on Parsing Technologies. 1993.

20. Ding J, Berleant D, Xu J, Fulmer AW. Extracting biochemical interactions from MEDLINE using a link grammar parser. In: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence. IEEE Computer Society; 2003.

21. Ahmed ST, Chidambaram D, Davulcu H, Baral C. IntEx: A syntactic role driven protein-protein interaction extractor for bio-medical text. In: Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. Detroit, Michigan: Association for Computational Linguistics; 2005.

22. Santos C, Eggle D, States DJ. Wnt pathway curation using automated natural language processing: Combining statistical methods with partial and full parse for knowledge extraction. Bioinformatics 2005;21:1653-8.

23. Pyysalo S, Salakoski T, Aubin S, Nazarenko A. Lexical adaptation of link grammar to the biomedical sublanguage: A comparative evaluation of three approaches. BMC Bioinformatics 2006;7 Suppl 3:S2.

24. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34:301-10.