September 2006

An Automated Procedure to Identify Biomedical Articles that Contain Cancer-associated Gene Variants

Ryan McDonald
University of Pennsylvania

Raymond Scott Winters
The Children's Hospital of Philadelphia

Claire K. Ankuda
University of Vermont

Joan A. Murphy
University of Vermont

Amy E. Rogers
University of Vermont

See next page for additional authors

Postprint version. Published in *Human Mutation*, Volume 27, Issue 29, September 2006, pages 957-964.
Publisher URL: http://dx.doi.org/10.1002/humu.20363

An Automated Procedure to Identify Biomedical Articles that Contain Cancer-associated Gene Variants

Abstract

The proliferation of biomedical literature makes it increasingly difficult for researchers to find and manage relevant information. However, identifying research articles containing mutation data, a requisite first step in integrating large and complex mutation data sets, is currently tedious, time-consuming and imprecise. More effective mechanisms for identifying articles containing mutation information would be beneficial both for the curation of mutation databases and for individual researchers. We developed an automated method that uses information extraction, classifier, and relevance ranking techniques to determine the likelihood of MEDLINE abstracts containing information regarding genomic variation data suitable for inclusion in mutation databases. We targeted the CDKN2A (p16) gene and the procedure for document identification currently used by CDKN2A Database curators as a measure of feasibility. A set of abstracts was manually identified from a MEDLINE search as potentially containing specific CDKN2A mutation events. A subset of these abstracts was used as a training set for a maximum entropy classifier to identify text features distinguishing "relevant" from "not relevant" abstracts. Each document was represented as a set of indicative word, word pair, and entity tagger-derived genomic variation features. When applied to a test set of 200 candidate abstracts, the classifier predicted 88 articles as being relevant; of these, 29 of 32 manuscripts in which manual curation found CDKN2A sequence variants were positively predicted. Thus, the set of potentially useful articles that a manual curator would have to review was reduced by 56%, maintaining 91% recall (sensitivity) and more than doubling precision (positive predictive value). Subsequent expansion of the training set to 494 articles yielded similar precision and recall rates, and comparison of the original and expanded trials demonstrated that the average precision improved with the larger data set. Our results show that automated systems can effectively identify article subsets relevant to a given task and may prove to be powerful tools for the broader research community. This procedure can be readily adapted to any or all genes, organisms, or sets of documents.

Keywords

p16, database, bioinformatics, genomics, CDKN2A, text mining, conditional random fields, relevance ranking

Comments

Postprint version. Published in *Human Mutation*, Volume 27, Issue 29, September 2006, pages 957-964.
Publisher URL: http://dx.doi.org/10.1002/humu.20363

Author(s)

Ryan McDonald, Raymond Scott Winters, Claire K. Ankuda, Joan A. Murphy, Amy E. Rogers, Fernando C.N. Pereira, Mark S. Greenblatt, and Peter S. White

An automated procedure to identify biomedical articles that contain cancer-associated gene variants

Ryan McDonald[1], R. Scott Winters[2], Claire K. Ankuda[3], Joan A. Murphy[3], Amy E. Rogers[3], Fernando Pereira[1], Marc S. Greenblatt[3] and Peter S. White[2, 4]

[1]*Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut Street, Philadelphia PA 19104 USA*

[2]*Division of Oncology, The Children's Hospital of Philadelphia, 34th St and Civic Center Blvd, Philadelphia, PA 19104, USA*

[3]*Vermont Cancer Center, University of Vermont College of Medicine, Burlington VT 05401*

[4]*Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA*

Correspondence to: Peter S. White, Division of Oncology, Rm 1407 CHOP North, The Children's Hospital of Philadelphia, 34th St. and Civic Center Blvd, Philadelphia, PA 19104-4318, USA

Tel: 215-590-5241

Fax: 215-590-5245

Email: white@genome.chop.edu

Running Title: Classifying text for cancer mutations

Abstract

The proliferation of biomedical literature makes it increasingly difficult for researchers to find and manage relevant information. However, identifying research articles containing mutation data, a requisite first step in integrating large and complex mutation data sets, is currently tedious, time-consuming and imprecise. More effective mechanisms for identifying articles containing mutation information would be beneficial both for the curation of mutation databases and for individual researchers. We developed an automated method that uses information extraction, classifier, and relevance ranking techniques to determine the likelihood of MEDLINE abstracts containing mentions of genomic variation data suitable for inclusion in mutation databases. We targeted the CDKN2A (p16) gene and the procedure for document identification currently used by CDKN2A Database curators as a measure of feasibility. A set of abstracts was manually identified from a MEDLINE search as potentially containing specific CDKN2A mutation events. A subset of these abstracts was used as a training set for a maximum entropy classifier to identify text features distinguishing “relevant” from “not relevant” abstracts. Each document was represented as a set of indicative word, word pair, and entity tagger-derived genomic variation features. When applied to a test set of 200 candidate abstracts, the classifier predicted 88 articles as being relevant; of these, 29 of 32 manuscripts in which manual curation found CDKN2A sequence variants were positively predicted. Thus, the set of potentially useful articles that a manual curator would have to review was

reduced by 56%, maintaining 91% recall (sensitivity) and more than doubling precision (positive predictive value). Subsequent expansion of the training set to 494 articles yielded similar precision and recall rates, and comparison of the original and expanded trials demonstrated that the average precision improved with the larger data set. Our results show that automated systems can effectively identify article subsets relevant to a given task and may prove to be powerful tools for the broader research community. This procedure can be readily adapted to any or all genes, organisms, or set of documents.

Key Words

p16, database, bioinformatics, mutations, genomics, CDKN2A, text mining, conditional random fields, relevance ranking

Introduction

Recent acceleration in research activities have produced challenges for researchers to identify, synthesize, and utilize published information. The semi-structured nature of biomedical text is not readily amenable to systematic approaches for information retrieval and management. Public repositories of biomedical research articles such as MEDLINE (Bodenreider, 2004), and interfaces to query these document sets such as PubMed (McEntyre and Lipman, 2001) and OVID (http://www.ovid.com), play critical roles in allowing identification of relevant articles through user-directed queries. However, MEDLINE provides only shallow semantic and no syntactic annotation of its content, with the result that document retrieval and relevance ranking capabilities are limited. More sophisticated automated techniques to extract information from text hold great promise in assisting in the identification and management of this wealth of research information (Cohen and Hersh, 2005; Krallinger and Valencia, 2005).

The current limitations of biomedical text retrieval capabilities can be illustrated by mutation databases that collect global mutation events, such as OMIM, COSMIC, and the Human Gene Mutation Database (Forbes, et al., 2006; Stenson, et al., 2003; Wheeler, et al., 2006), as well as specialized locus-specific databases (LSDBs), which record disease-causing gene mutations and neutral

variants for single genes, malignancies, or disease types (Horaitis and Cotton, 2004). LSDBs in particular have become valuable resources in the study and clinical management of cancer and many other genetic diseases. Over 200 publicly available LSDBs have been created in recent years. Many LSDBs now integrate large and complex mutation data sets with clinical and biological features of gene function. For example, we have created and continue to curate a LSDB for the tumor suppressor gene CDKN2A (Murphy, et al., 2004). CDKN2A (OMIM:600160) encodes the cell cycle regulatory protein p16(Ink4A), which is frequently mutated in a variety of cancers (Kamb, et al., 1994; Sharpless, 2005). The CDK2NA Database is a compendium of germline and somatic CDKN2A sequence variants associated with cancer.

However, compiling and maintaining a mutation database is labor intensive. The first step in this process, the identification of research articles that contain mutation data from the vast biomedical literature, is especially tedious, time-consuming and imprecise. As part of our efforts to improve the CDKN2A Database curation process, we have recently explored automated methods for the efficient identification of appropriate research articles that contain mutation data. We sought to develop an automated information extraction technique that would predict manuscripts that contain variation data suitable for inclusion in the CDK2NA Database, but that would be readily adaptable to any document set potentially describing genomic variation information of particular interest. Here,

we describe a methodology for predicting and relevance ranking articles of interest. This process combines 1) a named entity recognition algorithm to identify mentions of genomic variation from free text, and 2) a text-feature classifier that performs similarity analysis of potentially interesting documents to predict likely relevance. This method was successfully employed to predict with high precision which articles were most likely to contain mentions of CDKN2A genomic variation events. The overall procedure is directly applicable to any task requiring the identification of articles describing genomic variations.

Materials and Methods

Literature Search

For Version 1.0 of the CDKN2A Database, PubMed queries were performed in August 2000, November 2002, and February 2003 to identify manuscripts of potential relevance published through December 2002. Search parameters were: p16, mutation, cancer, human. Together, the queries identified 419 manuscripts published between January 2000 and December 2002. This set was labeled as Dataset 1. An expert curator manually read abstracts looking for variants reported in human tumors or cell lines and/or mention of one of the common techniques used to detect mutations. The expert scanned articles sequentially, considering first the article title, then the abstract, and then the full text of the article only if the expert considered there to be a likelihood of relevant information after each successive determination. Variants were included only if genomic DNA or cDNA sequencing was performed. In each case the article was marked as "true" if it contained at least one CDKN2A variation instance; otherwise, it was marked as "false". A second data set (Dataset 2) comprising the full collection of Dataset 1 along with an additional 267 documents represented all identified articles from January 2000 through June 2004. These additional articles were identified (in August 2004 and January 2005) and marked for relevance with the identical query and evaluation procedures employed for Dataset 1. The use of a 2nd training set that entirely encompassed the first was employed to mimic how

the classifier would likely be applied, where a user would wish to maximize the machine-learning benefit by including all possible documents suitable for training.

Document Classifier

In the natural language processing and machine learning communities, there has been a flurry of research on the problem of document classification and ranking (Crammer and Singer, 2003; Joachims, 2002; Nigam, et al., 1999). Our model uses the maximum entropy classification principle (Nigam, et al., 1999); such models are equivalent to multi-nomial logistic regression (Berger, et al., 1996). A maximum entropy classifier defines the probability that a document, *x*, is classified by the label, *y*, as shown in Figure 1. As per this formula, the probability of a document being relevant is proportional to a weighted linear sum over a set of features, f_i. The denominator in this term is present merely to insure that the probability distribution is properly normalized.

The CDKN2A document classification task requires only binary classification. In other words, only one of two labels for each document is possible: either it is relevant (y=1) or it is not relevant (y=-1). Maximum entropy classification relies on the definition of a set of indicative features, f_i, to help guide classification. Our model uses two kinds of features.

1) Word features indicate the presence of a word or word pair in the document. For instance the feature "$f_i(x,y) = 1.0$ if document *x* contains the word *CDKN2A*" may be created. Conjunctions, such as, "*$f_i(x,y) = 1.0$* if document *x* contains the word-pair *point mutation*", may also be created. Frequency of mention, but not location within a document, was considered in the model. Word triplets were not considered due to the likelihood of feature over-fitting for the document set. Character-based features did not significantly increase performance of the model.

2) The second class of features, genomic variation features, indicate the presence of a specific component of a genomic variation. For instance, the feature "$f_i(x,y) = 1.0$ if document *x* contains the location *codon 12*" may be created. In order to determine the presence or absence of genomic variation components, a named entity tagger for identifying text mentions of genomic variation that was previously developed by our group was applied (McDonald, et al., 2004). Specifically, this tagger identifies and distinguishes between text mentions of genomic variation type (e.g., point mutation, deletion), location (e.g., base pair 25, exon 2), and nucleic acid and protein state (e.g., A to T, Ala→Val). All CDKN2A document abstracts under consideration were used as input for the genomic variation tagger. The tagger annotated each abstract for genomic variation mention

predictions, and these annotations were used as input for feature evaluation by the classifier.

After defining the set of relevant features for classification, the weight, w_i, for each feature is determined. If a set of training data is available, this can be done automatically by finding the weights that maximize the likelihood of the training data (Berger, et al., 1996). The Dataset 1 and 2 training sets consisted of 219 and 494 documents, respectively. All documents had been manually labeled as either relevant (contains CDKN2A mutation data) or not. Once the classifier was trained, it was then run on a set of evaluation documents comprising the remaining articles in the trial set (200 for Dataset I; 192 for Dataset 2). The MALLET implementation of maximum entropy was used to construct the system (http://mallet.cs.umass.edu/).

Since automatically trained classifiers cannot guarantee that all relevant documents are classified correctly, a useful method would return a ranking of documents with the more relevant documents nearer the top. Maximum entropy provides a natural mechanism for ranking the documents. In particular, maximum entropy defines a probability P(y=1 l x), which is the probability that the document, *x*, is relevant. Using this probability score, a ranking of the documents was determined in each trial.

Evaluation

To evaluate the metric, the ranking criterion of *average precision* was used. Average precision measures the average accuracy of the rank over each possible rank cut-off. For instance, in Figure 2, if the cut-off between “considered relevant” and “considered not relevant” was established as being before position 5, the result would yield 4 documents, three that are actually relevant and 1 that is not (as assessed by the expert evaluator). The accuracy at this cut-off is 75%. The average precision metric sums this calculation (true positives/all documents), performed for all cut-offs. Intuitively this metric represents the likelihood of seeing a relevant document in the ranking at an arbitrary cut-off. For each trial, the cutoff yielding the highest maximum average precision was used for evaluation of performance. For determination of classifier performance relative to manual curation, the standard text mining measures of precision and recall were used. Precision was calculated as the number of articles correctly classified as relevant divided by the number of articles classified as relevant. Recall was calculated as the number of articles classified as relevant divided by the number of articles determined as relevant by the expert evaluator.

Results

A set of 419 biomedical articles published between 1/2000 and 6/2002 were identified from MEDLINE using a query of several keywords associated with CDKN2A, malignancy, and genomic variation (see Methods). This set was named Dataset 1. These articles were then evaluated manually by a domain expert to determine whether they described CDKN2A mutation instances suitable for inclusion in Version 1.0 of the CDKN2A Database. Articles were manually scored as either containing or not containing CDKN2A mutation data. Seventy of the 419 manuscripts [16.7% precision (specificity)] were found by the expert to contain relevant variation data. This set was then randomly divided into a training set of 219 articles and an evaluation set of the remaining 200 articles.

The training data were used to estimate a maximum entropy classifier that distinguished relevant from not relevant abstracts. As described in the Methods, our classifier defines the probability that a document, *x*, is classified by the label, *y*, based on weighting of syntactic and semantically-derived word features. Each document was represented as a set of indicative word, word pair, and entity tagger-derived genomic variation features (McDonald, et al., 2004). The model established by the training set was then evaluated on the remaining 200 articles. Article titles and abstracts in the evaluation set were subjected to the classifier, and each document was accorded an overall probability score indicating the likelihood that the document contained CDKN2A mutation information. An

average precision metric was then calculated, which measures the average accuracy of the rank over each possible rank cut-off (Figure 2).

The domain expert manually determined that 32 of the 200 evaluation articles actually contained CDKN2A mutation information (precision of the PubMed search was 32/200=16.0%). The classifier determined that 88 of the 200 articles (44%) likely contained mutation information. Twenty-nine of the 32 articles considered positive by the domain expert were included in the 88 articles predicted by the classifier (precision of 29/88=33.0%; recall of 29/32=90.6%). Predictions for each article are shown in Supplemental Figure S1. Application of the classifier more than doubled precision (33% vs. 16%), which would reduce expert evaluation efforts by 56% (88 articles to consider versus 200).

To confirm these findings and to determine whether a larger training set would improve performance, a second evaluation (Dataset 2) was performed on a set of 686 CDKN2A documents identified in MEDLINE between 1/2000 and 6/2004 by using the same initial query strategy. For this evaluation, all 419 documents used in Dataset 1 and an additional 75 documents (total of 494 documents) were used as a training set for the classifier. A separate set of 192 new articles was used for evaluation. Within the evaluation set, 27 were considered as positive for CDKN2A mutation instance data by the domain expert (precision of 27/192=14.1%). The classifier determined that 69 of the 192 articles (35.9%)

likely contained mutation information. Twenty-three of the 27 articles considered positive by the domain expert were included in the 67 articles predicted by the classifier (precision of 23/69=33.3%; recall of 23/27=85.2%). In this trial, application of the classifier improved precision 2.4-fold (33.3% vs. 14.1%) over that obtained by expert evaluation, which would in turn reduce expert evaluation efforts by 64% (69 articles to consider rather than 192). An average precision plot of the results is shown in Figure 3. Comparison of the Dataset 1 and Dataset 2 results demonstrates an overall higher performance for the larger trial (Figure 4).

Finally, the eight mutation-containing articles that the classifier failed to identify were analyzed in greater detail to determine possible causes. Article PMID:11058911 (Moore, et al., 2000) describes in detail a specific germline mutation of CDKN2A, but while this information is apparent in the article's title, there is no abstract body. Article PMID:14507338 (Godfraind, et al., 2003) focuses upon chromosomal deletions. This abstract has only four non-standard references to mutation: "CDKN2A alterations" (one instance) and "(epi)genetic modifications" (three instances). Similarly, articles PMID: 12898359 (Ohtsubo, et al., 2003) and PMID: 12721243 (Schneider-Stock, et al., 2003) both mention "mutation(s)" and either "homozygous deletion" or "loss of heterozygosity" sporadically, but each usually instead refers to "abnormalities", and the focus of the articles are on methylation status and immunohistochemical analysis of tumors. PMID:11159196 (Schraml, et al., 2001), specifically mentions "mutation

analysis" and "24-bp deletion" as the only two direct instances of mutation mentions, while most of the abstract describes results of a chromosomal deletion analysis. Importantly, 9p allelic loss and LOH instances are not considered as entries for inclusion in the CDKN2A Database. Article PMID: 15128789 (Huang, et al., 2004) frequently discusses a "mutated" product rather than a mutation, and this word would likely be missed by the tagger (stemming is not currently employed as a feature set) and not considered as similar to standard mentions such as "mutation" or "mutations" by the similarity analysis. Similarly, article PMID:15173226 (Goldstein, et al., 2004) mentions "mutations" but provides no specificity as to mutation types or locations, or the state of the DNA or protein. Thus, the tagger did not identify any mentions of genomic variation in this abstract, as it is trained to identify instances rather than generalized terms. The final false negative article, PMID:10942797 (Tsuchiya, et al., 2000) has five standard mentions of specified mutation phrases identified by the tagger. This abstract is written in an unusual style with many gene abbreviations and frequencies, and it uses an unusual form of the p16 gene name (p16INK4). As the classifier measures text feature similarity of documents to positive articles, is likely that these unusual elements makes this abstract sufficiently dissimilar to the positive training instances to be unrecognized.

Discussion

Efforts by several groups to provide portals to genomic variation information, including Online Mammalian Inheritance in Man, the Human Genome Variation Database, and the Human Genome Variation Society, have assisted with consolidation and more effective retrieval of mutation instances for particular diseases (Fredman, et al., 2004; Hamosh, et al., 2005; Horaitis and Cotton, 2004). Similarly, ongoing genome-wide mutation screening and data curation projects are generating sizable numbers of mutation instances for particular malignancies (Bamford, et al., 2004; Gottlieb, et al., 2004; Murphy, et al., 2004; Van Dreden, et al., 1989). However, many mutation instances are reported in the scientific literature, and attributing functional significance of identified mutation events requires specialized curation. As a result, LSDBs such as the CDKN2A Database have proved to be important resources for cancer and other genetic disorders, as they commonly provide data critical for linking molecular causes of disease with biological and clinical outcome. However, the level of effort required to initiate and maintain LSDBs is high. Also, because LSDBs target relatively specialized audiences, support for these resources is often limited. Despite these obstacles, over 200 separate LSDBs have been established (Horaitis and Cotton, 2004), and this number is expected to increase as the human genome becomes more fully annotated in functional terms. Our classification method is readily adaptable to assist with literature curation for many of these databases, as well

as for more general applications to populate biomedical datasets with mutation information.

The results reported here suggest that use of a specialized document classifier can substantially assist with the time-consuming task of filtering relevant documents from a larger initial set. Collectively, our system was able to positively identify 51 of 59 articles (86.4% recall) mentioning CDKN2A mutation instances while reducing the number of articles under consideration from 419 to 157). This reduction of over 60% translates to a saving of many person-hours of effort in curation each year. Interestingly, this procedure used only article titles and abstract texts, indicating that in most cases the article summaries provide sufficient clues regarding the presence or absence of desired mutation instances in the full text. Analysis of the articles missed indicate that these abstracts often mentioned mutation events in unusual ways, such as using non-standard terms for describing the genomic variations. Our genomic variation tagger includes a specialist lexicon of commonly used synonyms for mutation and genomic alteration text mentions (McDonald, et al., 2004). Expansion of this list to include the mentions used in the missed articles, or inclusion of additional feature sets specific to these exceptional cases, would likely assist with identification of these articles. It would also be interesting to see if a similar approach using full-length articles as input would yield higher performance, or whether the documents would be more dissimilar due to a greater proportion of divergent and extraneous

text, differences in article formatting, and variation in writing style.

Comparison of the results of the original and expanded datasets showed modest improvement in precision and a marginal decline in recall, suggesting the possibility that larger training sets will positively influence performance. It is reasonable to expect that continued utilization of the classifier would provide more accurate results over time. However, determination of the significance and optimal size of the training set, as well as the iterative impact of the machine learning component, will require additional training data and analysis.

While term-based queries of MEDLINE are effective for many information retrieval tasks, use of this procedure for identifying specific text content that is often mentioned in various ways is inefficient, and to our knowledge, tools to assist with this process are not readily available to bioinformatics-limited groups at this time. For example, the MEDLINE web interface PubMed has a “Related articles” feature that pre-computes a word feature-based similarity for all MEDLINE documents, allowing a user to identify articles similar to a selected individual abstract (McEntyre and Lipman, 2001). However, this tool does not allow similarity to be performed within a selected set of documents. To determine how well the PubMed tool performs for our task, we determined the frequency with which a CDKN2A mutation-positive article in Dataset 1 was present in the “Related Articles” set for each CDKN2A-positive article in Dataset 2. The overall

precision (# of Dataset 1 positive articles identified/# of Dataset1-positive articles) and recall (# of Dataset 1 positive articles identified/# of articles in the “Related Articles set) for this feature were 11.4% and 13.1%, respectively. Because the “Related Articles” feature is calculated against all MEDLINE articles rather than a smaller set of likely candidates, a lower performance is expected. However, this result indicates that many CDKN2A-related articles are likely sufficiently dissimilar to require more domain-targeted approaches such as our method provides.

Machine learning-based document classification is a mechanism in wide use in other application domains, such as Internet searching and email spam detection (Robinson, 2004; Zhang, et al., 2004). However, for biomedical tasks, only a few groups have reported the use of classifiers to identify document subsets (Bartling, et al., 2003; Chapman, et al., 2005; Chapman, et al., 2003; Rubin, et al., 2005), and these systems do not utilize advanced natural language processing methods. Dobrokhotov and colleagues (Dobrokhotov, et al., 2005) successfully used a combination of lemmatization, morpho-syntactic pattern recognition, and either Support Vector Machine- or Probabilistic Latent-based classifiers to classify and relevance rank MEDLINE articles suitable for annotating protein sequences. In contrast, our approach combined a natural language processing technique that was trained specifically upon the domain of interest with a generalized classifier in order to improve performance. The high

recall from our method indicates that this approach is suitable as a convenient filtering step prior to manual assessment and retrieval of relevant CDKN2A mutation data. In addition, as our classifier provides a ranking function for each document, database curators can begin with the articles deemed most relevant and establish their own imposed cutoffs.

An advantage of our system over the Dobrokhotov approach is that tailoring the NLP-based retrieval component to a specialized domain of interest provides an opportunity for increased performance. However, specialization requires additional effort for each new domain encountered. Our genomic variation entity tagger is built upon a probabilistic model that can operate with high performance in the absence of domain-specific features, but which also requires specialized feature sets for optimal performance, as well as a moderate amount of hand-annotated text specific to the domain of interest. A more comprehensive tagging procedure which incorporates part-of-speech tagging and sentence-level syntactic parsing would likely improve the quality of the genomic variation features employed by the classifier. As mentioned above, additional lexicons and regular expressions specific to genomic variation would undoubtedly improve performance; analysis of false negatives from a larger set of documents could assist in identifying recurrent patterns to exploit. Alternatively, additional syntactic and semantic approaches could be applied to the text independently and their outputs incorporated as feature sets for the classifier. Moreover, pre-tagging the

entirety of MEDLINE with the genomic variation tagger to generate an exhaustive lexicon of genomic variation mentions would likely be a valuable classifier feature set. It would also be expected that training of a classifier such as the one described here on full-text articles would improve performance, especially as many variation events are described in detail only in manuscript tables.

While our classifier assists with document ranking, it does not assist with the identification of specific text sections relevant to curation and annotation tasks. A possible use of our classifier would be to utilize it in combination with a specialized biomedical literature indexing tool for extraction of sentences and phrases relevant to genomic variation. For example, Textpresso is a tool that provides advanced indexing capabilities that incorporate Gene Ontology terms, to allow a user to immediately identify sections of text matching pre-defined biological attributes (Muller, et al., 2004). Textpresso has been implemented in several model organism domains as an effective literature curation tool. Our classifier could be used to define and relevance rank the document set of interest, whereupon relevant contextual strings could be extracted or annotated using Textpresso or a similar tool. Furthermore, as our classifier utilizes a tagger that identifies short phrases describing genomic variation, a slight modification of the application would allow output to be marked up (e.g. by color-coded HTML tags) for phrases representing genomic variation.

Our classifier was designed specifically to be readily adaptable to a wide domain of knowledge. For the identification of articles potentially mentioning genomic variations or mutations of a specific gene, the system requires only 1) the classifier; 2) a set of training articles or abstracts that contain both positive and negative instances of the type of genomic mention of interest; and 3) our genomic variation tagger. Preliminary results have shown that performance is slightly but not substantially improved with the addition of the tagger. Furthermore, the classifier can be trained upon any set of documents in which a contextual distinction can be made, although the performance will likely vary depending upon how precisely the distinction between positive and negative instances can be defined.

In summary, specialized document classification is a powerful technique for assisting with the growing need for curation of biological and biomedical text. Automated systems can effectively identify article subsets relevant to a given task. Opportunities for specialized high-performance document classifiers exist for database population and curation, but also for data integration tasks such as the alignment of molecular and clinical objects with biomedical text records. The combination of a generalized classifier with a feature-based and domain-trained NLP engine provides a potential way to streamline curation and annotation tasks considerably.

Acknowledgments

This work was supported in part by NSF ITR grant 0205448, NIH/NCI grant CA 96536, the David Lawrence Altschuler Endowed Chair, and the Penn Genomics Institute. The authors gratefully acknowledge members of the Penn/CHOP BioIE team, especially M. Liberman, M. Mandel, Y. Jin, and K. Murphy for helpful discussions and technical assistance.

References

Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 91(2):355-8.

Bartling WC, Schleyer TK, Visweswaran S. 2003. Retrieval and classification of dental research articles. Adv Dent Res 17:115-20.

Berger A, Della Pietra S, Della Pietra V. 1996. A Maximum Entropy Approach to Natural Language Processing. Comput Linguist 22(1):39-71.

Bodenreider O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32(1):D267-70.

Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, Olszewski RT. 2005. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artif Intell Med 33(1):31-40.

Chapman WW, Cooper GF, Hanbury P, Chapman BE, Harrison LH, Wagner MM. 2003. Creating a text classifier to detect radiology reports describing mediastinal

findings associated with inhalational anthrax and other disorders. J Am Med Inform Assoc 10(5):494-503.

Cohen AM, Hersh WR. 2005. A survey of current work in biomedical text mining. Brief Bioinform 6(1):57-71.

Crammer K, Singer Y. 2003. A Family of Online Algorithms for Category Ranking. J Mach Learn Res 3:1025-1058.

Dobrokhotov PB, Goutte C, Veuthey AL, Gaussier E. 2005. Assisting medical annotation in Swiss-Prot using statistical classifiers. Int J Med Inform 74(2-4):317-24.

Forbes S, Clements J, Dawson E, Bamford S, Webb T, Dogan A, Flanagan A, Teague J, Wooster R, Futreal PA, Stratton MR. 2006. Cosmic 2005. Br J Cancer 94(2):318-22.

Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H, Brookes AJ. 2004. HGVbase: a curated resource describing human DNA variation and phenotype relationships. Nucleic Acids Res 32(Database issue):D516-9.

Godfraind C, Rousseau E, Ruchoux MM, Scaravilli F, Vikkula M. 2003. Tumour necrosis and microvascular proliferation are associated with 9p deletion and CDKN2A alterations in 1p/19q-deleted oligodendrogliomas. Neuropathol Appl Neurobiol 29(5):462-71.

Goldstein AM, Struewing JP, Fraser MC, Smith MW, Tucker MA. 2004. Prospective risk of cancer in CDKN2A germline mutation carriers. J Med Genet 41(6):421-4.

Gottlieb B, Beitel LK, Wu JH, Trifiro M. 2004. The androgen receptor gene mutations database (ARDB): 2004 update. Hum Mutat 23(6):527-33.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33(Database issue):D514-7.

Horaitis O, Cotton RG. 2004. The challenge of documenting mutation across the genome: the human genome variation society approach. Hum Mutat 23(5):447-52.

Huang J, El-Gamil M, Dudley ME, Li YF, Rosenberg SA, Robbins PF. 2004. T cells associated with tumor regression recognize frameshifted products of the

CDKN2A tumor suppressor gene locus and a mutated HLA class I gene product. J Immunol 172(10):6057-64.

Joachims T. 2002. Learning to Classify Text using Support Vector Machines [Ph.D. thesis]. Ithaca: Cornell University. 224 p.

Kamb A, Shattuck-Eidens D, Eeles R, Liu Q, Gruis NA, Ding W, Hussey C, Tran T, Miki Y, Weaver-Feldhaus J, et al. 1994. Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. Nat Genet 8(1):23-6.

Krallinger M, Valencia A. 2005. Text-mining and information-retrieval services for molecular biology. Genome Biol 6(7):224.

McDonald RT, Winters RS, Mandel M, Jin Y, White PS, Pereira F. 2004. An entity tagger for recognizing acquired genomic variations in cancer literature. Bioinformatics 20(17):3249-51.

McEntyre J, Lipman D. 2001. PubMed: bridging the information gap. CMAJ 164(9):1317-9.

Moore PS, Zamboni G, Falconi M, Bassi C, Scarpa A. 2000. A novel germline mutation, P48T, in the CDKN2A/p16 gene in a patient with pancreatic carcinoma. Hum Mutat 16(5):447-8.

Muller HM, Kenny EE, Sternberg PW. 2004. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2(11):e309.

Murphy JA, Barrantes-Reynolds R, Kocherlakota R, Bond JP, Greenblatt MS. 2004. The CDKN2A database: Integrating allelic variants with evolution, structure, function, and disease association. Hum Mutat 24(4):296-304.

Nigam K, Lafferty J, McCallum A. Using Maximum Entropy for Text Classification. International Joint Conference on Artificial Intelligence: Workshop on Information Filtering; 1999; Stockholm, Sweden. p 61-67.

Ohtsubo K, Watanabe H, Yamaguchi Y, Hu YX, Motoo Y, Okai T, Sawabu N. 2003. Abnormalities of tumor suppressor gene p16 in pancreatic carcinoma: immunohistochemical and genetic findings compared with clinicopathological parameters. J Gastroenterol 38(7):663-71.

Robinson S. 2004 The ongoing search for efficient web search algorithms. SIAM News. 37(9).

Rubin DL, Thorn CF, Klein TE, Altman RB. 2005. A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge. J Am Med Inform Assoc 12(2):121-9.

Schneider-Stock R, Boltze C, Lasota J, Miettinen M, Peters B, Pross M, Roessner A, Gunther T. 2003. High prognostic value of p16INK4 alterations in gastrointestinal stromal tumors. J Clin Oncol 21(9):1688-97.

Schraml P, Struckmann K, Bednar R, Fu W, Gasser T, Wilber K, Kononen J, Sauter G, Mihatsch MJ, Moch H. 2001. CDKNA2A mutation analysis, protein expression, and deletion mapping of chromosome 9p in conventional clear-cell renal carcinomas: evidence for a second tumor suppressor gene proximal to CDKN2A. Am J Pathol 158(2):593-601.

Sharpless NE. 2005. INK4a/ARF: a multifunctional tumor suppressor locus. Mutat Res 576(1-2):22-38.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21(6):577-81.

Tsuchiya T, Sekine K, Hinohara S, Namiki T, Nobori T, Kaneko Y. 2000. Analysis of the p16INK4, p14ARF, p15, TP53, and MDM2 genes and their prognostic implications in osteosarcoma and Ewing sarcoma. Cancer Genet Cytogenet 120(2):91-8.

Van Dreden P, Richard P, Gonzales J. 1989. Fructose and proteins in human semen. Andrologia 21(6):576-9.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, Kapustin Y, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. 2006. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 34(Database issue):D173-80.

Zhang L, Zhu J, Yao T. 2004. An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing 3(4):243-269.

Figure Legends

Figure 1. Equation used to define the probability that a document, *x*, is classified by the label, *y*. This equation states that the probability of a document being classified as “relevant” is proportional to a weighted linear sum over a set of features, *f*.

Figure 2. Average precision for Dataset 1. This Figure plots the average percentage of relevant documents returned as a function of the number of documents in total. Our system is compared to a baseline in which a relevance ranking of documents is randomly created.

Figure 3. Average precision for Dataset 2. This Figure plots the average percentage of relevant documents returned as a function of the number of documents in total. Our system is compared to a baseline in which a relevance ranking of documents is randomly created.

Figure 4. Comparison of the average precisions for Datasets 1 and 2. This Figure plots the average percentage of relevant documents returned as a function of the number of documents in total.