BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 3 2011, pages 408–415
doi:10.1093/bioinformatics/btq667

Data and text mining

Advance Access publication December 7, 2010

Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature

Emily Doughty[1,†], Attila Kertesz-Farkas[1,2,†], Olivier Bodenreider[3], Gary Thompson[1], Asa Adadey[1], Thomas Peterson[1] and Maricel G. Kann[1,*]

[1]University of Maryland, Baltimore County, Baltimore, MD 21250, [2]Division of Imaging and Applied Mathematics, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD 20993 and [3]National Library of Medicine, Bethesda, MD 20894, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: A major goal of biomedical research in personalized medicine is to find relationships between mutations and their corresponding disease phenotypes. However, most of the disease-related mutational data are currently buried in the biomedical literature in textual form and lack the necessary structure to allow easy retrieval and visualization. We introduce a high-throughput computational method for the identification of relevant disease mutations in PubMed abstracts applied to prostate (PCa) and breast cancer (BCa) mutations.

Results: We developed the extractor of mutations (EMU) tool to identify mutations and their associated genes. We benchmarked EMU against MutationFinder—a tool to extract point mutations from text. Our results show that both methods achieve comparable performance on two manually curated datasets. We also benchmarked EMU's performance for extracting the complete mutational information and phenotype. Remarkably, we show that one of the steps in our approach, a filter based on sequence analysis, increases the precision for that task from 0.34 to 0.59 (PCa) and from 0.39 to 0.61 (BCa). We also show that this high-throughput approach can be extended to other diseases.

Discussion: Our method improves the current status of disease-mutation databases by significantly increasing the number of annotated mutations. We found 51 and 128 mutations manually verified to be related to PCa and Bca, respectively, that are not currently annotated for these cancer types in the OMIM or Swiss-Prot databases. EMU's retrieval performance represents a 2-fold improvement in the number of annotated mutations for PCa and BCa. We further show that our method can benefit from full-text analysis once there is an increase in Open Access availability of full-text articles.

Availability: Freely available at: http://bioinf.umbc.edu/EMU/ftp.

Contact: mkann@umbc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 29, 2010; revised on November 17, 2010; accepted on November 23, 2010

1 INTRODUCTION

Relating human genomic variation to disease risk is one of the major challenges of personalized medicine. During the past decade, hundreds of mutations in the human genome have been associated to disease phenotypes. When available, this information can be integrated into the analysis of genomic variation of individual genomes, which is crucial to developing personalized medicine approaches. Yet, mutational information is, for the most part, in textual form and not organized for easy identification and association to the corresponding disease phenotype. Moreover, the complexity of natural language processing and the continuous exponential growth of literature repositories (e.g. PubMed) make extracting this information challenging.

Publicly available databases like online Mendelian inheritance in man (OMIM) (Amberger *et al*., 2009) and Swiss-Prot (Boeckmann *et al*., 2003) contain categorized protein and DNA mutational information with explicit associations to cancer and other diseases. Other resources focus on specific diseases, such as cancer [e.g. Cancer Gene Index (caBIG, 2007)] or specific chromosomal locations (Claustres *et al*., 2002). These databases are currently constructed and curated manually, which is a slow process that limits the number of cancer mutations available to the biomedical community. Text-mining methods to find reported mutations in these databases have yielded up to 0.98 sensitivity (Caporaso *et al*., 2007) but, to date, no accurate automatic methods to find disease-related mutations are available.

Several algorithms have been developed to extract mutational data from the biomedical literature. All of these methods implement standard regular expressions to identify either the point mutations alone [e.g. MutationFinder (Caporaso *et al*., 2007)], or both the mutations and their associated gene and protein names [e.g. MEMA (Rebholz-Schuhmann *et al*., 2004); MuteXt (Horn *et al*., 2004)]. Most of the 'mutation + gene' recognition methods implement standard regular expressions and generate text collections of the point mutation information; some, however, provide algorithmic and interface alternatives. For instance, Mutation Grab (Lee *et al*., 2007) uses graph-based expressions to identify mutations, while MutationMiner (Baker and Witte, 2006) uses structural visualizations to display them. Overall, most methods have focused on the extraction of point mutations and their association with specific genes with reasonable accuracy. However, generating a resource of disease-related mutations is still a challenge.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

MuGeX (Erdogmus and Sezerman, 2007) extends the 'mutation + gene/protein' extraction capability to investigate the role of the given mutation in disease. MuGeX automatically finds mutation-gene pairs in MEDLINE abstracts for Alzheimer's disease and the authors claim that this functionality can be extended to any disease query. But, one of MuGeX's disadvantages is its inability to identify the exact correspondence between mutations and genes in abstracts featuring multiple mutations in multiple genes. For instance, if three mutations are present in three different genes, MuGeX reports the nine possible mutations, of which only three are correct. Most recently Yeniterzi and Sezerman (2009) developed EnzyMiner as an enzyme-specific mutation identification tool, which was applied to a set of disease-related abstracts to identify disease mutations. One of the disadvantages of EnzyMiner is that it only reports mutations related to enzymes. For that subset, however, it provides additional relevant information about the functional cause of the disorder. OSIRIS (Bonis *et al*., 2006), on the other hand,focuses on identifying abstracts linked to entries from the dbSNP (Sherry *et al*., 2001) database. While many of these entries have disease-associations, such associations are limited to the manually curated OMIM database. In a recent paper, Kuipers *et al*. introduced an automatic method to extract and validate mutations for Fabry disease (Kuipers *et al*., 2010).

In this article, we propose a novel, high-throughput approach for identifying point mutations and their relationships to disease phenotypes from the biomedical literature. Our approach combines text mining and sequence analysis to identify mutations and their associated genes and diseases. Our results show that when our approach is applied to the identification of mutations related to prostate cancer (PCa) and breast cancer (BCa), we obtain almost twice as many annotated mutations for these diseases compared to the current data in OMIM and Swiss-Prot. Application of our method to full-text articles is desirable, but currently limited by their availability in publicly available repositories. Furthermore, we show that this semi-automatic method can be applied to other diseases and discuss the remaining challenges for the complete automation of this methodology.

2 MATERIALS

We used the PubMed search engine to retrieve a set of abstracts that were potentially useful for identifying mutations from MEDLINE. Our PubMed query took advantage of the controlled vocabulary indexing in MEDLINE's Medical Subject Headings (MeSH). We submitted a simple query based on the MeSH descriptor 'mutation' (i.e. term = 'mutations[MeSH Terms]') and downloaded all PubMed citations for which an abstract was available. The citations were downloaded in XML format using the 'e-utilities' programming interface provided by NCBI.

The following explains the identification of disease terms using the MetaMap program. The identification of names of entities in text is generally referred to as entity recognition and often exploits terminologies and ontologies as a source of vocabulary (Krauthammer and Nenadic, 2004; Park and Jim, 2006). In the biomedical domain, the largest source of vocabulary is the Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004). The Metathesaurus integrates some 8-million terms from ∼150 source vocabularies. The MetaMap program (Aronson, 2001) is a biomedical entity-recognition software specifically designed to discover Metathesaurus concepts in text. We used MetaMap to identify disease names in the title and abstract of MEDLINE citations. Version 2008AB of the UMLS was used as the source of vocabulary. Specific options selected in MetaMap include the identification of the longest-spanning entities (e.g. identify 'androgen-independent prostate cancer' rather than 'androgen' and 'prostate cancer' in the phrase 'Molecular biology of androgen-independent prostate cancer'). Citations were processed on a small cluster of Solaris computers at the National Library of Medicine.

The Metathesaurus concepts identified by MetaMap in the titles and abstracts of MEDLINE citations were restricted to disease concepts. Each Metathesaurus concept is categorized using semantic types from the UMLS Semantic Network (McCray, 2003). For example, the concept 'adenocarcinoma of prostate' is categorized with the semantic type 'neoplastic process'. Groupings of semantic types, called semantic groups (Bodenreider and McCray, 2003), define coarser categories. We used the semantic group 'disorders' as a filter for selecting disease entities from the UMLS concepts identified by MetaMap in MEDLINE citations.

For specific diseases, we used hierarchical relations among UMLS concepts to select all specific kinds of cancers. Examples of kinds of prostate cancer include 'prostate cancer stage B', 'carcinoma in situ of prostate' and 'androgen-independent prostate cancer'. We identify as related to prostate cancer any citation in which any descendant, direct or not, of the Metathesaurus concept 'malignant neoplasm of prostate' (C0376358) is discovered by MetaMap. Analogously, we use 'malignant neoplasm of breast' (C0006142) for breast cancer. The resulting sets of citations compose the PCa_MetaMap and BCa_MetaMap datasets, respectively.

3 METHODS

Our approach for identifying disease mutations is depicted in Figure 1 and can be summarized as follows. First, the EMU mutation extraction tool is used to identify and retrieve mutations from the mutation-disease-related corpus. EMU also identifies gene names in the text. Second, a filter (SEQ_Filter) is applied to exclude all mutations for which the amino acids differ from those reported in the associated reference protein sequences. Finally, the pre- and post-SEQ_Filter results could be optionally curated manually to create a database of fully-annotated disease mutations. In the following sections, we provide a more detailed description of this multi-step approach.

3.1 EMU: a method for extracting mutations from the biomedical literature

3.1.1 Identifying mutations through text mining with EMU The EMU algorithm is a rule-based method that finds mutations in a given document using regular expression matching. The input is plain text and the output is a list of mutation terms. The algorithm is executed in two steps. First, EMU searches the input text for mutations using a set of regular expressions (called positive patterns) and stores the recognized terms in a list L. In the second step, EMU eliminates false positive terms from list L using another set of regular expressions (called fallible patterns). A set of 6541 cell line names was included in the fallible patterns. The regular expression schemes used in the positive and fallible patterns are available via our ftp site.

To develop the mutation patterns, we used the regular expressions in Pharmspresso (Garten and Altman, 2009) as a reference. We further refined Pharmspresso's patterns by manually extracting positive and fallible patterns from a set of 300 abstracts from PubMed. This 300-abstract set was chosen randomly from the 'MeSH = mutation' set we queried from PubMed. The 300-abstract set excluded the abstracts used in the 'PCa' and 'BCa' datasets or in any of the gold standards used to evaluate our methods in this article.

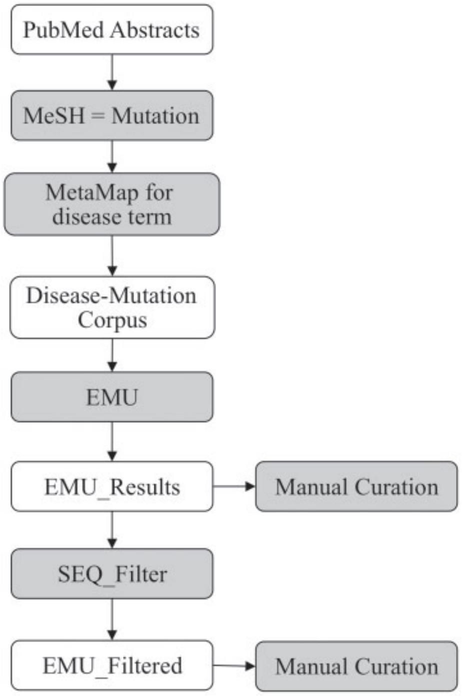

Fig. 1. Schematic of the overall methodology for disease-related extraction of mutational information (manual curation processes are optional).

In the literature, mutations are often recorded within word descriptions and not always in the (more easily parsed) alphanumerical mutation convention. For instance, an A21G mutation might be recorded as 'Ala was substituted by Gly in the VKORC1 gene at residue 21'. Thus, we created regular expressions to capture mutation annotations within word descriptions, as well as to identify cryptic residue positions (e.g. 'residue DD', 'codon DD', where DD stands for a residue position). When more than one residue position is found in the same sentence, EMU selects the position closest to the mutation pattern. In addition, EMU's algorithm also includes regular expressions that use the Human Genome Variation Society's (HGVS) nomenclature for residue substitution. EMU is implemented in Perl and it is freely available via our ftp site, along with the complete set of regular expressions used in this study (http://bioinf.umbc.edu/emu/ftp/).

3.1.2 Identifying gene information with EMU Historically, genes and proteins have not been recorded in the biomedical literature in a standardized manner. As a result, automating the extraction of gene and protein information from abstracts has remained a significant text-mining challenge. Our approach for extracting gene names is based on a string look-up on an in-house dictionary of human gene names from the Human Genome Organization (HUGO) and from the National Center for Biotechnology Information's (NCBI) gene database. All gene names that were identical to codon names were removed. The P53 gene name, which was absent in both gene dictionaries used, was added since it is a commonly used term.

It should be noted that EMU uses a dictionary of human genes; thus, EMU assumes that all reported genes are from human. Therefore, in our approach, during the first manual curation process, we removed all non-human-related abstracts.

3.2 Filtering mutations with inconsistencies in amino acid position with SEQ_Filter

For each gene name and mutation identified within an abstract we validated the corresponding protein information against NCBI's reference protein sequence set [RefSeq (Pruitt *et al.*, 2007)]. More specifically, given a gene, we identified all its associated proteins. For each protein, we verified that the wild-type amino acid recorded in the given mutation corresponded to the actual amino acid in the specified protein sequence position. In practice, we used NCBI's e-utilities to search for the gene in GenBank (Benson *et al.*, 2009) and to retrieve their linked proteins along with their sequences. The association between the gene and the mutation was deemed valid if there was a match for the wild-type (or mutated type) amino acid at the position of the mutation for at least one of the associated proteins.

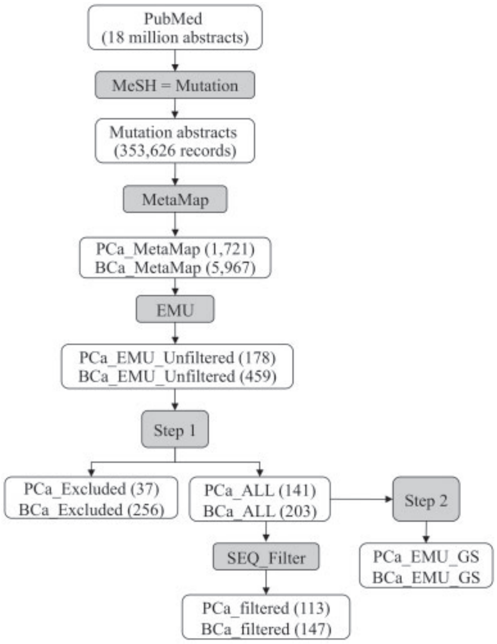

Fig. 2. Schematic of the process used to benchmark EMU. Steps 1 and 2 represent the manual curation processes to create the gold standards datasets for prostate cancer (PCa) and breast cancer (BCa), PCa_EMU_GS and BCa_EMU_GS, respectively.

3.3 Evaluation

3.3.1 Establishing gold standards All citations related to PCa or BCa identified by MetaMap, and in which EMU had identified a mutation, were manually curated by our team and further validated by the same team (seven curators including authors ED and GT). Curators identified mutation, gene name, and the disease-association for each citation. The curated data was then validated by two additional curators to ensure accuracy. If during the validation process, the validator could not reconcile the mutational information with that of the original curator, the entry was excluded from the final set. Abstracts lacking information about the relationship between the phenotype and mutation were also manually excluded from the benchmarking set. Likewise, abstracts with incomplete information about the mutation were excluded. All abstract exclusions are available in the PCa_Excluded and BCa_Excluded datasets (see diagrams in Fig. 2). The curators provided a numerical code for the categorization of the disease-mutation relationship. The manually curated datasets, PCa_EMU_GS and BCa_EMU_GS, could be used for training machine-learning methods for text mining other disease-mutation relationships, and are available upon request from the authors.

3.3.2 Evaluating mutation detection against MutationFinder In order to evaluate EMU's ability to retrieve non-synonymous single point mutation events, we compared its performance to that of MutationFinder

(Caporaso *et al*., 2007), which is a reference tool for identifying mutations. The performance evaluation was conducted on two manually-curated gold standard datasets: PCa (PCa_GS) and MutationFinder (MF_GS). Both gold standards were evaluated on both EMU and MutationFinder.

PCa gold standard (PCa_GS): We randomly selected 500 abstracts from the PCa_MetaMap mutation corpus subset. These citations were manually curated by our team, who annotated 95 mutations in 55 abstracts.

MutationFinder Gold Standard (MF_GS): This gold standard was downloaded from MutationFinder's supplementary data and it contains 476 unique mutations in 508 abstracts.

Metrics: To evaluate the mutation extraction algorithms we used the following standard information-retrieval metrics: recall, precision and F-measure. The recall R is defined as $R = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, the precision as $P = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and the F-measure as $F = 2\mathrm{PR}/(\mathrm{P}+\mathrm{R})$, where TP, FP and FN stand for the number of true positives, false positives and false negatives, respectively.

3.3.3 Benchmarking EMU's precision as identifying correct mutational information, genes, and disease association in PCa and BCa-related records To evaluate EMU's overall performance (i.e. not only the identification of the mutation, but also its relation to a gene and a disease), we used the sets PCa_EMU_GS and BCa_EMU_GS that were manually-curated records obtained according to Section 3.3.1 above. For each one of the data subsets (with and without SEQ_Filter, as shown in Fig. 2), the precision was calculated as $P = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$. In addition, the following three events were defined. A 'mutation' (mutation only) event involves the identification of the position in which the mutation took place and the identification of the wild-type and mutated amino acids. A 'complete mutation' (mutation + gene) event involves the identification of the mutation and a validation of the gene in which the mutation was reported to occur. In a 'disease mutation' event, the mutation, the gene and the disease to which the mutation is related are identified.

4 RESULTS

This section is organized by the different steps in our procedure to extract disease mutations: namely, creating the mutation-disease-related corpus, using EMU to extract mutations (which is benchmarked against MutationFinder), and estimating the overall performance (precision in terms of identifying correct mutational information, genes and disease association) of our approach.

4.1 Establishing a corpus of mutation-disease-related abstracts

A search of the 'mutation' MeSH descriptor in July 2008 retrieved 447 601 abstracts from which 353 626 were selected based on abstract availability (Fig. 2). Using a set of 218 UMLS concepts denoting forms of PCa as a filter for UMLS concepts identified by MetaMap in the title and abstract of MEDLINE citations, we selected 1721 citations annotated by MetaMap related to any of these PCa concepts (set PCa_MetaMap). The UMLS concepts most frequently identified are 'prostate carcinoma' (C0600139) and 'malignant neoplasm of prostate' (C0376358), and, more rarely, 'adenocarcinoma of prostate' (C0007112) and 'prostate cancer metastatic' (C0936223). Similarly, a set of 5967 citations was obtained with the filtering for breast cancer (set BCa_MetaMap).

4.2 Benchmarking EMU and MutationFinder

We compared EMU against MutationFinder using two manually curated gold standard datasets: PCa_GS and MF_GS, which were previously described in the Section 3. The results in Table 1 show that both methods have high precision at comparable levels ($P = 0.92$–0.99). EMU achieves a comparable F-measure in the MF_GS dataset (EMU: 0.89 versus MutationFinder: 0.88) and better F-measure for the PCa_GS dataset (EMU: 0.92 versus MutationFinder: 0.84).

Table 1. EMU against MutationFinder (MF) on two datasets

		TP	FP	FN	Precision	Recall	F-measure
PCa_GS	EMU	87 (53)	8 (6)	8 (3)	0.92	0.92	0.92
	MF	70 (45)	1 (1)	25 (14)	0.99	0.74	0.84
MF_GS	EMU	388 (164)	3 (3)	92 (37)	0.99	0.81	0.89
	MF	387 (162)	6 (4)	93 (36)	0.98	0.81	0.88

Results are based on the number of mutations extracted. The numbers in parentheses represent the number of abstracts in which the mutations were found.

Table 2. EMU's overall precision

Dataset	Mutation only: precision (TP, FP)	Complete mutation: precision (TP, FP)	Disease-mutation: precision (TP, FP)
PCa_ALL	0.97 (248, 8)	0.53 (207, 181)	0.39 (151, 237)
PCa_filtered	0.99 (195, 2)	0.80 (173, 43)	0.59 (127, 89)
BCa_ALL	0.94 (353, 23)	0.42 (300, 412)	0.34 (242, 470)
BCa_filtered	0.96 (249, 10)	0.74 (233, 81)	0.61 (193, 121)

4.3 Precision in retrieval of PCa and BCa-related records

As shown in Fig. 2, EMU's precision was evaluated in 141 and 203 abstracts obtained in the initial searches of PCa and BCa mutations, respectively (PCa_ALL and BCa_ALL). EMU's precision was also evaluated in 113 (PCa_filtered) and 147 (BCa_filtered) abstracts that passed the SEQ_Filter analysis.

Table 2 records EMU's performance at identifying the following events extracted from these datasets: mutation only, complete mutation (mutation + gene) and disease-mutation (mutation + gene + disease). In what follows, the numbers reported are numbers of mutations, not numbers of citations in which these mutations were identified.

4.3.1 Mutation only Our results show that EMU's precision for mutation-only events is high and ranging from 0.94 to 0.99. These results are consistent with our previous finding in the MF_GS and PCa_GS datasets.

4.3.2 Complete mutation The additional task of identifying the gene of the reported mutation results in a substantial loss of the method's precision: from 0.97 to 0.53 and from 0.94 to 0.42 for the PCa_ALL and BCa_ALL sets, respectively. However, applying the SEQ_Filter protocol significantly improves the correct retrieval of genes and increases precision (e.g. precision values of 0.80 and 0.74 were observed for the PCa_filtered and BCa_filtered datasets, respectively).

Table 3. EMU's overall precision using abstracts versus full text

Dataset	Mutation only: precision (TP, FP)	Complete mutation: precision (TP, FP)	Disease-association: precision (TP, FP)
BCa_Abs	0.95 (19, 1)	0.37 (17, 29)	0.37 (17, 29)
BCa_Abs_filtered	0.93 (14, 1)	0.55 (11, 9)	0.55 (11, 9)
BCa_Full	0.84 (47, 9)	0.55 (46, 37)	0.55 (46, 37)
BCa_Full_filtered	0.94 (17, 1)	0.77 (17, 5)	0.77 (17, 5)

4.3.3 Disease-mutation Lastly, to evaluate EMU's ability to retrieve disease-specific mutations, our curators manually annotated and validated all complete mutation-phenotype associations extracted by our method (i.e. EMU was combined with MetaMap to retrieve disease-related records). From this analysis, 151 out of the 207 correctly-identified mutations in the PCa-ALL set were found to be related to PCa. Similarly, 242/300 BCa_ALL, 127/173 PCa_filtered, and 193/233 BCa_filtered mutations were related to their corresponding diseases. Table 2 lists the overall precision of the automatic portion of the method in retrieving mutations that denote risk to the disease (No SEQ_Filter: 0.39 and 0.34, SEQ_Filter: 0.59 and 0.61, for the PCa and Bca, respectively).

4.4 Comparing EMU's overall precision with abstracts versus full-text

We performed a preliminary analysis on a subset of 10 full-text articles related to BCa (see additional details in Supplementary information). The performance of EMU and SEQ_Filter in the abstracts of the 10 articles (BCa_Abs and BCa_Abs_filtered sets) was compared against their performance in the body of the text (BCa_Full and BCa_Full_filtered). Table 3 depicts EMU's performance at identifying the following events extracted from these datasets: mutation only, complete mutation (mutation + gene) and disease-mutation (mutation + gene + disease).

4.4.1 Mutation only Our results show that EMU's precision for mutation-only events is comparable between abstracts and full text (0.93 versus 0.94 for the BCa_Abs_filtered and BCa_Full_filtered sets, respectively).

4.4.2 Complete mutation The additional task of identifying the gene of the reported mutation results in a substantial loss of the method's precision: from 0.93 to 0.37 and from 0.94 to 0.55 for the BCa_Abs and BCa_Full sets, respectively. However, applying the SEQ_Filter protocol significantly improves the correct retrieval of genes and increases precision from 0.37 to 0.55 and from 0.55 to 0.77 for BCa_Abs_filtered and BCa_Full_filtered sets, respectively. In comparing the precision between BCa_Abs_filtered and BCa_Full_filtered, the latter showed a 0.22 increase in precision.

4.4.3 Disease-mutation Our manual curation showed that all complete mutations in both BCa_Abs_filtered and BCa_Full were related to BCa. The BCa_Full_filtered contained 17 true positives BCa mutations while the BCa_Abs_filtered contain 11 BCa mutations. When comparing both sets, there were seven additional true positives in the BCa_Full_filtered set. The BCa_Abs_filtered set contained one true positive not present in the BCa_Full_filtered set.

5 DISCUSSION

In this section, we discuss the results for the comparison of: (i) two methods to extract mutations (EMU versus MutationFinder) and (ii) the advantages and disadvantages of using SEQ_Filter. We also discuss mapping the mutations found by EMU to disease-mutation reference databases. In addition, we discuss the feasibility of applying our approach to other diseases and to full-text articles, as well as the main challenges we foresee in the complete automation of the annotation of disease mutations extracted from PubMed abstracts.

5.1 Evaluating the ability of EMU and MutationFinder to identify mutations

We compared our method for identifying mutations (EMU) against a reference, high performance method [MutationFinder, (Caporaso *et al*., 2007)]. Our results show that EMU achieves a comparable *F*-measure and a better recall (0.92 versus 0.74) for the PCa_GS dataset. Analysis of the records in which MutationFinder missed the mutations and EMU did not suggests that MutationFinder missed those whose protein locations are cryptically specified (e.g. by the mutation's codon number). Moreover, several mutation patterns [e.g. '(d)A>A'] seem to be missing in the MutationFinder method. EMU failed to extract eight mutations from three abstracts that were also not found by MutationFinder. These mutations come from the following sentences: (i) 'by substituting alanine for six residues in the proposed IGF binding site, Ile(56)/Tyr(57)/Arg(75)/Leu(77)/Leu(80)/Leu(81)' (Hong *et al*., 2002), (ii) 'mutants with Leu47 replaced by serine' (Ray *et al*., 1988) and (iii) 'substitution of glutamic acid for the wild-type glutamine at position 798' (Evans *et al*., 1996). The first case contains six Ala to Ile mutations in several different positions, which EMU failed to identify. The last two cases contain expressions for mutation changes and locations that were not included in the current version of EMU.

Lastly, EMU retrieved eight false positives, two of which are shown below. The first phrase read: 'codon 119 (G→T), codon 432 (C→G), codon 449 (C→T)' (Tanaka *et al*., 2002). From this sentence, EMU extracted the order of the positions incorrectly, which resulted in two incorrectly-recorded mutations: codon 449 (C→G) and codon 432 (G→T). The second phrase involved a synonymous mutation—this type of mutation was excluded since it is not reported by MutationFinder—and it read: 'revealed six 201G>A (R201R) polymorphisms' (Koivisto *et al*., 2004). EMU reports a 'G>A' that results in a synonymous mutation, and while EMU did not report R201R, it still reported '201G>A'. The additional five false positives are either insertions or transitions that could be easily corrected in the next release by modifying the fallible patterns.

In summary, both MutationFinder and EMU show high precision in extracting mutations. These findings are comparable to the high mutation-pattern accuracy results reported for (Erdogmus and Sezerman, 2007) and suggest that these methods capture unique identifying features of mutations that separate them from other natural language samples.

5.2 Evaluating EMU's performance at identifying gene names

One of the most challenging tasks in natural language processing is that of retrieving gene and protein names from the literature. We documented two main issues responsible for the difficulty of this task: (i) identifying the correct gene name and (ii) identifying the correspondence between genes and mutations when multiple mutations are reported within an abstract.

First, we evaluated how well our dictionary-based method correctly identified the gene name. To do so, we manually identified all the records for which EMU recognized a mutation without its associated gene name. Our team of curators identified only 23 (out of 300) BCa-mutations and 12 (out of 205) PCa-mutations associated to genes that were not in the dictionary used. Examples of gene names that were not in our dictionary include the ER (official symbol: ESR) and Caveolin 1 (official symbol: CAV-1). In conclusion, our approach can successfully identify gene names when the standard gene name is used in the abstract. More sophisticated approaches, such as BANNER (Leaman and Gonzalez, 2008), GeNo (Wermter *et al*., 2009) and others (Neves *et al*., 2010; Tanabe and Wilbur, 2002) could be used to identify and normalize name entities when alternative names are used for the gene name.

Second, our results show that the application of the SEQ_Filter protocol greatly increases the correct identification of gene-mutation relationships, particularly when several genes and mutations are reported together. We found that, in most cases, when a gene name was incorrectly assigned to a mutation, the reported amino acids did not match the residues on the reference sequence. For these cases, inaccurate gene name assignments were ruled out when the SEQ_Filter was applied. We expect the SEQ_Filter to rule out almost all DNA mutations with only ∼10% error due to incorrect assignment of a DNA base mutation to a protein mutation. Our results in BCa show that this estimates are correct: only 30 out of the 314 BCa_filtered mutations are DNA mutations. We show that the application of the SEQ_Filter greatly improves the precision of the method. However, the SEQ_Filter limits the extraction to protein-only mutations. For an expanded search of mutations reported at the DNA level other techniques need to be implemented. A more sophisticated approach to match each gene/protein name inside text to its corresponding mutation is needed to make full use of this method.

5.3 Mapping mutations to the OMIM, dbSNP and Swiss-Prot reference databases

In addition to the evaluation of EMU, we compared EMU's results against two databases of disease-related protein mutations (Swiss-Prot and OMIM), to assess the impact of the combined approach on the current annotation of PCa and BCa mutations. When a mutation's gene, position, wild-type amino acids and mutated amino acids were identical to those in the reference database, the mutation was classified as a match to that database entry. A non-redundant count of mutation matches to OMIM and Swiss-Prot was reported in our results. Matches to dbSNP were used only to provide a SNP identifier (rs id) when it was available.

Table 4 lists the mutations found by EMU (in conjunction with MetaMap) that match entries in the reference databases. We show that our method finds a total of 87 PCa-mutations and 189 BCa-mutations in the set of abstracts studied in this work. Out of these mutations, 26 PCa (from both PCa_ALL and PCa_filtered) and 66 BCa mutations from BCa_ALL and 58 BCa mutations from BCa_filtered matched disease mutations already listed in the OMIM and Swiss-Prot databases. Furthermore, 12 of the 26 PCa-mutations and 49 of the 66 and 42 of the 58 BCa-mutations (from BCa_ALL and BCa_filtered, respectively) were not previously annotated specifically for PCa or Bca, respectively. In addition, we manually verified that 51 and 128 mutations extracted by EMU were previously unlisted PCa and BCa protein mutations (SEQ_Filter applied). Thus, using EMU combined with manual curation resulted in the annotation of 51 (12 + 39) PCa-mutations and 128 (42 + 86) BCa-mutations not previously annotated for these cancer types.

Table 4. Mapping EMU's disease mutations to those in OMIM and Swiss-Prot

	TP (unique mutations)	TP in disease databases	TP not in disease databases
BCa_ALL	189	66	123
BCa_filtered	144	58	86
PCa_ALL	87	26	61
PCa_filtered	65	26	39

Two mutations were considered identical if they occurred in the same gene, location, and involved the same amino acids.

In addition, we classified all disease mutations in the reference databases and found that there are 22 PCa mutations and 23 BCa mutations in OMIM (July 2009), and 75 PCa- and 123 BCa-mutations in Swiss-Prot (July 2009). The combined total of unique mutations in both databases comes to 83 PCa-mutations and 139 BCa-mutations. The new set of mutations correctly identified by EMU almost duplicates the existing size of the annotated datasets for both diseases. When the SEQ_Filter is not applied, both DNA and protein mutations are extracted by EMU and, thus, the total annotated mutations are 73 (12 + 61) PCa-mutations and 172 (49 + 123) BCa-mutations. All the mutations, i.e. those currently annotated in the reference databases (83 PCa and 139 BCa mutations) and those by EMU (51 PCa and 128 BCa-mutations), can be accessed on our website (http://bioinf.umbc.edu/EMU).

5.4 Limitations of the current approach and future work

As discussed before, correctly matching the gene-mutation relationship is a challenging task when text-mining mutations from abstracts. Another, perhaps more difficult, task is that of extracting the degree of association (or risk) between the mutation and the disease phenotype from the abstract text alone. For instance, our method classified a mutation as relevant to PCa even though the result was a negative outcome for the association (e.g. 'the existence of this mutation in PCa patients was not associated with any of the clinical or pathological characteristics of the disease'). Another error occurs when the mutation is assigned to an incorrect disease. This is a common mistake when processing abstracts reporting genes or mutations with multiple disease associations. This problem is particularly relevant in cancer, since cancer genes are known to participate in a myriad of other diseases or cancer types. We found 46 and 40 additional false positives when we moved from the evaluation of complete mutations to disease-mutations in the PCa_filtered

Table 5. Time estimation for manual curation of disease mutations for 10 diseases

Disease	Number of abstracts	Number of abstracts from EMU's output	Number of abstracts from EMU's output with filter
Prostate cancer	2081	251 (40)	158 (25)
Breast cancer	8097	804 (128)	445 (72)
Diabetes mellitus	4788	1513 (242)	505 (81)
Alzheimer's disease	2743	814 (130)	316 (51)
Hemochromatosis	1415	884 (141)	664 (106)
Lung cancer	4774	579 (93)	319 (51)
Acute myeloid leukemia	6987	304 (48)	148 (24)
Pancreatic cancer	1741	212 (34)	112 (18)
Colon cancer	4240	448 (72)	220 (35)
Cystic fibrosis	2341	587 (94)	497 (79)

The numbers in parentheses represent the number of hours estimated to be needed for manual curation and validation of abstracts.

and BCa_filtered datasets, respectively. From the additional false positives, we found 17 and 34, respectively, to be associated with another disease. This suggests that the precision values reported in Table 2 represent a lower bound of the method's performance. We expect that, for diseases involving genes that are more specific to the particular disease, the method's precision will improve (i.e. closer to the performance for extracting the complete mutation).

In conclusion, the automatic procedure drastically reduces the number of relevant records (from over 350 000 to ∼200 PCa and 500 BCa related abstracts), facilitating manual curation step. We manually curated all the PCa- and BCa-related records by annotating the disease-risk of the given mutations. We benchmarked the time needed to manually curate EMU's output for 50 abstracts. It took our curators an average of 3 h to annotate the mutation, gene name and disease risk information from 50 abstracts. Completing the curation process requires two additional validation steps, and each step takes less time than the initial curation. We have estimated that the total time needed to completely validate 50 abstracts is ∼8 h. Table 5 shows an estimation of the time it would take to curate and validate the mutation information from PubMed abstracts related to 10 arbitrarily chosen diseases. Abstracts for each disease were selected based on a search that combines MetaMap with a MeSH indexing for each disease.

Table 5 includes the number of abstracts obtained when EMU is applied to the mutation corpus and when the additional filter is incorporated. Overall, the filtering procedure reduces the number of abstracts to be processed in half. We estimate that it would take three curators roughly ∼3 months to process all the abstracts related to these 10 diseases. However, this time can be shortened to 1.5 months if only the filtered abstracts were analyzed. From our studies of PCa and BCa mutations, we estimate that 75% of the mutations will still be recovered after the filter is applied, which indicates that the filter provides a good compromise between time and sensitivity.

We used MetaMap in order to identify disease terms, but MeSH indexing could also have been used to select for citations for a specific disease or diseases. We compared both methods to identify disease-related citations. Although most citations were found by both methods, we suggest a combined approach to extend the analysis to other diseases and have maximal coverage.

Finally, our results are based on processing abstracts and limited by the number of mutations reported within them. We expect a significant number of mutations to be mentioned only in the body section of the article. Below we discuss two issues concerning the application of our method to the full text of manuscripts, namely the availability of full-text articles for text mining and the performance of the method in finding disease mutations.

Despite efforts from publishers and PubMed Central (PMC) to provide free, open access to full-text articles, there are still a large number of manuscripts without open access. From the original 1721 PCa_MetaMap and 5967 BCa_MetaMap, there are only 19 and 168 articles, respectively, with PMC open access. Furthermore, the PCa_ALL and BCa_ALL sets contain only 12 and 4 articles, respectively, with PMC open access. Using our approach (i.e. EMU and SEQ_filter), we were able to extract 14 unique BCa-mutations (10 of them were both in the abstract and body of the manuscript and one was mentioned only in the abstract) with a precision that increased from 0.55 (abstracts only) to 0.77 (body of manuscripts). In conclusion, there is a gain in coverage when using the body of the document (i.e. three new unique mutations) with a significant increase in precision. However, this approach is severely limited by the availability of full-text articles for text mining in open access repositories.

6 CONCLUSIONS

Identifying the correct gene-mutation pair and the mutation-phenotype relationship remains one of the main challenges in automating the extraction and classification of disease mutations. In this work, we introduce an automatic method for the extraction of mutations, which is coupled with manual validation of the gene and of the disease risk of the mutation. We show that this high-throughput approach for extracting disease mutations is scalable and holds great potential for contributing to a systematic repository of mutations with disease-phenotype associations. Our preliminary results also indicate that our text mining approach will greatly benefit from the increasing open access availability of full-text articles.

ACKNOWLEDGEMENTS

The authors are grateful to Aster Kenea, Brittany Beach, Hanna Fenta, Arianna Maharaj, Adenike Orogade, Veer Gariwala, Ajmer Randhawa and Olayinka Savage for helping E.D. and G.T. to manually curate hundreds of PubMed abstracts, to Yanan Sun, Trevor Suznick, Genaro Hernandez and Michael Povolotzky for their help with programing and to Mileidy Gonzalez for her invaluable editorial assistance. We wish to thank Carol Friedman for helpful discussions at the onset of this study, as well as Lan Aronson and Jim Mork at the National Library of Medicine for processing the citations with MetaMap.

Funding: This work was supported by the National Institutes of Health (NIH) [1K22CA143148, MGK (PI) and R01LM009722, MGK (collaborator)]. A.K.F. acknowledges funding from the Research Fellowship Program of the Centre for Devices and Radiological Health, administrated by the Oak Ridge Associated Universities through a contact with the US Food and Drug Administration. Additional funding (OB) was provided by the

Intramural Research Program of the NIH, National Library of Medicine.

Conflict of Interest: none declared.

REFERENCES

Amberger,J. *et al.* (2009) McKusick's online Mendelian inheritance in man (OMIM). *Nucleic Acids Res.,* **37**, D793–D796.

Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, 17–21.

Baker,C.J.O. and Witte,W. (2006) Mutation mining–a prospector's tale. *Information Systems Frontiers,* **8**, 47–57.

Benson,D.A. *et al.* (2009) GenBank. *Nucleic Acids Res.,* **37**, D26–D31.

Bodenreider,O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.,* **32**, D267–D270.

Bodenreider,O. and McCray,A.T. (2003) Exploring semantic groups through visual approaches. *J. Biomed. Inform.,* **36**, 414–432.

Boeckmann,B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.,* **31**, 365–370.

Bonis,J. *et al.* (2006) OSIRIS: a tool for retrieving literature about sequence variants. *Bioinformatics,* **22**, 2567–9.

caBIG (2007) The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. *Stud. Health Technol. Inform.,* **129**, 330–334.

Caporaso,J.G. *et al.* (2007) MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics,* **23**, 1862–1865.

Claustres,M. *et al.* (2002) Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res.,* **12**, 680–688.

Erdogmus,M. and Sezerman,O.U. (2007) Application of automatic mutation-gene pair extraction to diseases. *J. Bioinform. Comput. Biol.,* **5**, 1261–1275.

Evans,B.A. *et al.* (1996) Low incidence of androgen receptor gene mutations in human prostatic tumors using single strand conformation polymorphism analysis. *Prostate,* **28**, 162–171.

Garten,Y. and Altman,R.B. (2009) Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics,* **10** (Suppl 2), S6.

Hong,J. *et al.* (2002) Insulin-like growth factor (IGF)-binding protein-3 mutants that do not bind IGF-I or IGF-II stimulate apoptosis in human prostate cancer cells. *J. Biol. Chem.,* **277**, 10489–10497.

Horn,F. *et al.* (2004) Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics,* **20**, 557–568.

Koivisto,P.A. *et al.* (2004) Kruppel-like factor 6 germ-line mutations are infrequent in Finnish hereditary prostate cancer. *J. Urol.,* **172**, 506–507.

Krauthammer,M. and Nenadic,G. (2004) Term identification in the biomedical literature. *J. Biomed. Inform.,* **37**, 512–526.

Kuipers,R. *et al.* (2010) Novel tools for extraction and validation of disease-related mutations applied to fabry disease. *Hum. Mutat.,* **31**, 1026–1032.

Leaman,R. and Gonzalez,G. (2008) BANNER: an executable survey of advances in biomedical named entity recognition. *Pac. Symp. Biocomput.*, **13**, 652–663.

Lee,L.C. *et al.* (2007) Automatic extraction of protein point mutations using a graph bigram association. *PLoS Comput. Biol.,* **3**, e16.

McCray,A.T. (2003) An upper-level ontology for the biomedical domain. *Comp. Funct. Genomics,* **4**, 80–84.

Neves,M.L. *et al.* (2010) Moara: a Java library for extracting and normalizing gene and protein mentions. *BMC Bioinformatics,* **11**, 157.

Park,J.C. and Jim,J.-j. (2006) Named entity recognition. In: Ananiadou,S. and McNaught,J. (eds.) *Text Mining for Biology and Biomedicine*. Artech House, Boston, pp. 121–142.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.,* **35**, D61–D65.

Ray,P. *et al.* (1988) Structure-function studies of murine epidermal growth factor: expression and site-directed mutagenesis of epidermal growth factor gene. *Biochemistry,* **27**, 7289–7295.

Rebholz-Schuhmann,D. *et al.* (2004) Automatic extraction of mutations from Medline and cross-validation with OMIM. *Nucleic Acids Res.,* **32**, 135–42.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.,* **29**, 308–311.

Tanabe,L. and Wilbur,W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics,* **18**, 1124–1132.

Tanaka,Y. *et al.* (2002) Polymorphisms of the CYP1B1 gene have higher risk for prostate cancer. *Biochem. Biophys. Res. Commun.,* **296**, 820–826.

Wermter,J. *et al.* (2009) High-performance gene name normalization with GeNo. *Bioinformatics,* **25**, 815–821.

Yeniterzi,S. and Sezerman,U. (2009) EnzyMiner: automatic identification of protein level mutations and their impact on target enzymes from PubMed abstracts. *BMC Bioinformatics*, **10** (Suppl 8), S2.