Journal of
Health & Medical Informatics

Schenck et al., J Health Med Informat 2012, S2
DOI: 10.4172/2157-7420.S2-002

Research Article Open Access

Extraction of Genetic Mutations Associated with Cancer from Public Literature

Martin Schenck[1], Oliver Politz[2] and Philip Groth[2]*

[1]Database Systems and Information Management, Technical University Berlin, 10587 Berlin, Germany
[2]Therapeutic Research Group Oncology, Bayer Healthcare Pharmaceuticals, 13353 Berlin, Germany

Abstract

Genomic mutations may result in severe diseases, e.g. cancer, a disease with a significant genetic component. The mutation state of cancer tissues is e.g. being determined experimentally in order to find the most likely response to a drug treatment. Results of such experiments are typically published in scientific literature.

We have developed a workflow of several text-mining algorithms, in order to harvest this wealth of information relevant to developing novel therapeutic approaches in cancer. Our workflow has successfully scanned over 150,000 abstracts related to cancer and genetic mutations. New information on mutated genes in cancer could be extracted with a precision and recall of 86.8% and 30.3%, respectively. By applying the workflow, novel associations of mutations in specific cancer tissues could be extracted for 264 genes.

Keywords: Text mining; Mutations; Cancer

Abbreviations: API: Application Programming Interface; COSMIC: Catalogue of Somatic Mutations in Cancer; HUGO: Human Genome Organization; MeSH: Medical Subject Headings; NER: Named Entity Recognition; NCBI: National Center for Biotechnology Information; NCIBI: National Centre for Integrative Biomedical Informatics; OMIM: Online Mendelian Inheritance in Man; SNP: Single Nucleotide Polymorphism; SAX: Simple API for XML; XML: Extensible Markup Language

Introduction

Genomic mutations may result in severe diseases, e.g. cancer, one of the most widespread diseases in which a significant genetic component has been widely recognized [1]. With such a property, the mutation state of cancer tissue can be used, e.g. to discriminate the most likely response to a drug treatment [2,3] (Figure 1, adapted from Sharma et al. [3]). Furthermore, relating human genomic variation to disease risk is one of the major challenges of personalized medicine [4]. Therefore, large-scale access to data on cancer tissues and types with their associated genomic mutations is required in order to develop novel treatments in areas with a strong medical need.

The vast biomedical literature repository of PubMed [5] is a resource granting such access. In November of 2011, 159,221 abstracts matching the query "(cancer OR carcinoma OR neoplasm) AND (mutation OR SNP OR polymorphism)" could be found. Thus, the information is available in large scale, but in an unstructured form. The largest structured publicly available resource is the Catalogue of Somatic Mutations in Cancer (COSMIC) [6]. It is mostly fed through the manual curation of selected articles, also available via PubMed and has experienced a vast growth within recent years. In November 2011, it contains over 67,000 unique mutations from almost 13,000 curated studies, covering mutation information for over 200 cell lines on known "hot spot" cancer driver genes (such as KRAS or PTEN), but also whole genome information from a limited number of studies. Thus, it has become the international de facto standard repository for mutation information.

As of now, only manually maintained databases exist, containing information on the concept "mutations of genes in cancer". It becomes evident that the data deluge from publications (almost 10,000 new

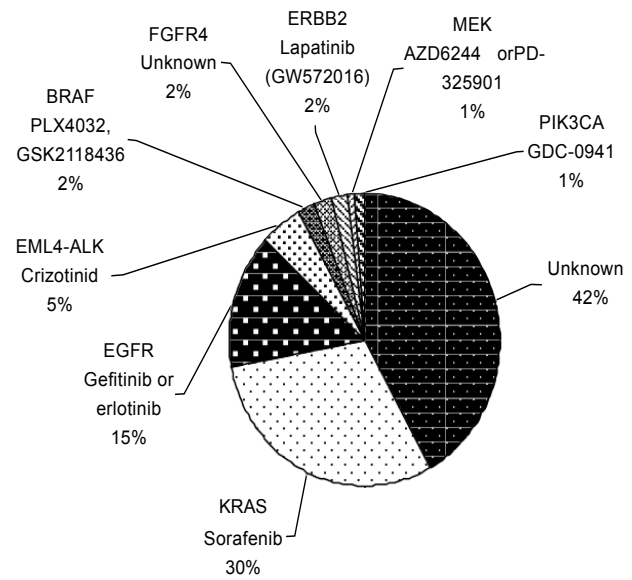

Figure 1: Figure was adapted from Sharma et al. [3]. Distribution of various reported activating oncogenic mutations in a survey of 139 non-small-cell lung cancer (NSCLC)-derived cell lines. Also shown for all the activating mutations (except KRAS) are inhibitors (supplemented by the author) that selectively target the activated oncoproteins, yielding the most effective growth inhibition and/or apoptosis of cancer cell lines. There are currently no inhibitors that target oncogenic KRAS [3], but it has been reported that Sorafenib is the best treatment option for those lung cancer patients with KRAS mutations [2].

***Corresponding author:** Philip Groth, Therapeutic Research Group Oncology, Bayer Healthcare Pharmaceuticals, 13353 Berlin, Germany, Tel: +49 30 468 196073; Fax: +49 30 468 18069; E-mail: philip.groth@bayer.com

Received November 23, 2011; **Accepted** December 25, 2011; **Published** January 03, 2012

Citation: Schenck M, Politz O, Groth P (2012) Extraction of Genetic Mutations Associated with Cancer from Public Literature. J Health Med Informat S2. doi:10.4172/2157-7420.S2-002

publications each year) cannot be fully compensated by manual efforts such as COSMIC (where the total number of curated papers of the past 5 years has just reached 12,000). Due to this discrepancy between the available free text and the already annotated publications, Human annotators are bound to lose overview over the vast number of publications containing the sought-after information. Thus, automated efforts are necessary to overcome this gap between available and (machine-) usable data. Hence, the creation of a method to automatically mine the available biomedical literature and populate a database with information on mutations, genes, cell lines and their relation to one another can bring a significant advantage to the field.

Several algorithms have already been developed to extract either gene or disease or mutation data from the biomedical literature. The aim of these lie in helping researchers to access information contained in scientific publications in a faster, easier and more complete way [7]. However, to further advance the task described above, there is a need to evaluate the most efficient of these algorithms and combine them into one workflow, not only considering extraction of each of these entities alone, but especially when they are mentioned in context to one another.

The goal of the present work is to incorporate different text mining methods towards an integrative workflow extracting and associating genes, mutations and diseases (or disease models, i.e. cancer cell lines) and to evaluate and benchmark their results. For evaluation, the programmatic results are compared to a manually annotated text corpus derived from COSMIC. Furthermore, results are improved by editing some of the available tools, optimizing precision and at the same time retaining a reasonable recall. The source code of the resulting workflow of tools and databases is available within the supplementary material of this publication and some examples of what has been achieved by applying this workflow is given within this study.

Materials and Methods

Overview

The result of the present work is a tool we call *gemuline*, usable for extracting genetic mutations in cancer models (i.e. cell lines) or cancer sub-types from text. Our workflow extracts either cancer model or cancer sub-type if mentioned in the text. It prefers the name of a cell line (i.e. cancer model) over the mention of a cancer sub-type since an identified mutation in a specific model is typically more useful for experimental settings.

For input, *gemuline* first needs to get a set of texts (i.e. scientific publications) to be analyzed. Then it extracts entities (i.e. mutations, genes and diseases) and their relation to one another. Finally, it writes the results to a database. Incorporated tagger and verification methods include amongst others GNAT [8], the NCIBI Name Tagger [9], Mutation Finder [10], Array Express [11], COSMIC [6] and Uniprot [12]. For evaluation, a manually annotated corpus was added to the database and compared to the workflow's results. Results were improved by adding more regular expressions to Mutation Finder. The

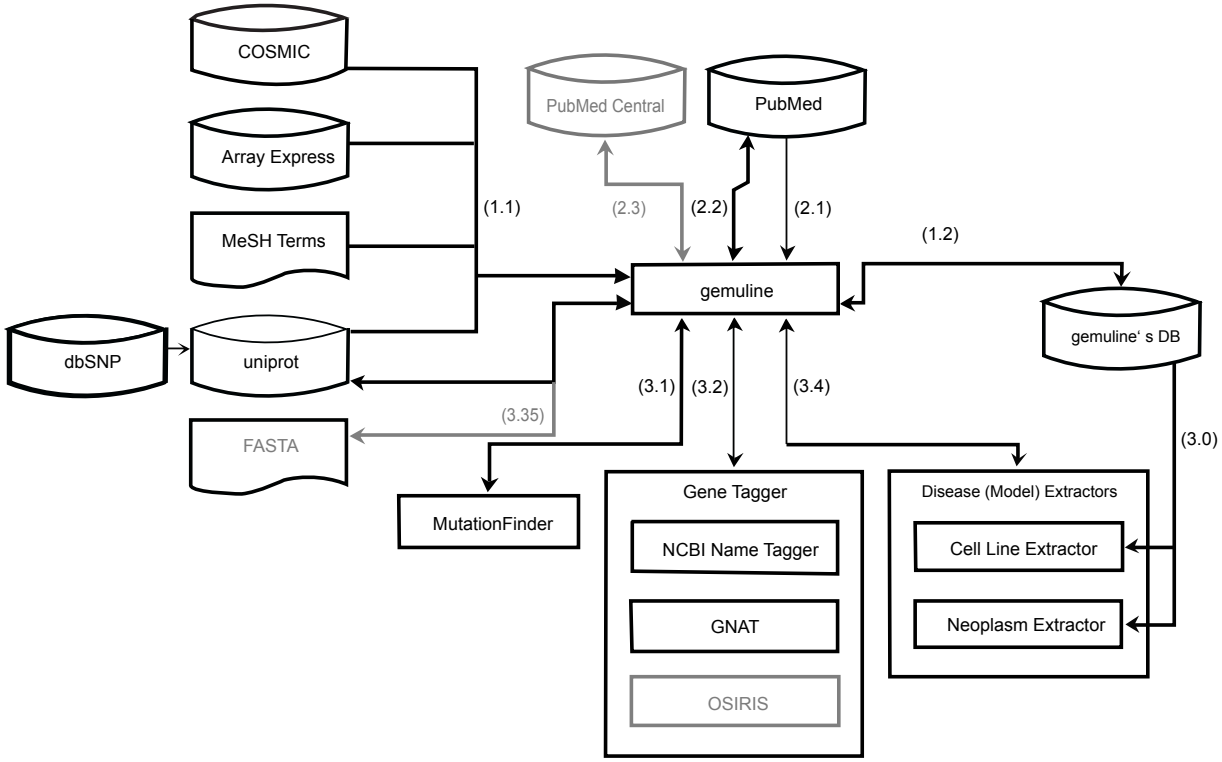

Figure 2: Workflow of all processes of *gemuline* in an overview. (1) Updating the data warehouse. Extracting information from data sources (1.1) and writing the processed, combined data into the data warehouse (1.2). (2) Text retrieval. Acquiring PubMed ids from an online query (2.1) and downloading the abstracts (2.2). Optionally, retrieving full text versions (2.3). (3) Information retrieval. First, loading the cell line and neoplasm extractor with aliases from the data warehouse (3.0). Running MutationFinder to get mutation candidates (3.1). Tagging genes with either the NCBI Name Tagger or GNAT (3.2). Originally it was planned to also integrate OSIRIS, but the tool is currently not publically available (see chapter 2.1.10). Matching mutations to genes with Uniprot (3.3). If (3.3) is not possible, gene sequences are checked for a possible mutation at the given position (3.35). Finally, extraction and matching to cell lines or MeSH terms (3.4) takes place.

best-performing tools identified in the course of this study are Mutation Finder and GNAT, thus they are used in the final implementation of the workflow. For disease extraction, a new extractor based on dictionaries was implemented by the authors as no working disease-term extractor exists at the time. Figure 2 depicts the overall workflow in detail.

This chapter discusses at first the main components of the implemented workflow: the database, the interface, text retrieval and information retrieval. Then, the implementation is described in more detail.

Database

A database is the foundation of the tool (Figure 3). It contains information on genes, diseases and gene expressions in cell lines. The tool extracts gene information from COSMIC and Uniprot. COSMIC provides most of the information, i.e. primary gene symbol, Entrez Gene ID [13] and synonyms for a gene. However, the Ensembl ID [14] is not listed in COSMIC. It is therefore fetched from Uniprot for every gene. COSMIC furthermore supplies a comprehensive set of cell line names. A disease term list supplementing the cell line names is extracted from Medical Subject Headings (MeSH) [15]. Since the external MeSH database cannot be queried, a file in Extensible Markup Language (XML) downloaded from the MeSH website was used to import these data. MeSH covers far more terms than needed for this study. Thus, the content of the XML was filtered to the terms under the "neoplasm" (i.e. cancer) branch. Gene expression values in cell lines can be found in the Array Express database. XML files were fetched from Array Express, one file per gene. In the XML files, each gene's expressions in various cell lines are stated. The tool extracts the expression values and stores them in the database as well. An expression is always linked to its gene and its cell line respectively. The tool can update the database at any time. However, database access cannot be guaranteed during the update. Updating drops all current entries for diseases, genes and expressions and fetches them from the underlying data sources anew.

Interface

gemuline includes a web interface available to access the database. Figure 4 shows a screenshot of the web interface with results for a query. The web interface was built using PHP and javascript (including jQuery and jQueryUI). It allows searching for certain mutations, genes, diseases or a combination of those. The results can be manually validated to be true or false. Results manually set to false will only be shown as results of a query if the user explicitly wants to see them. The web interface links to all known sources like PubMed, Entrez Gene or Array Express.

Text retrieval

The tool receives documents on the basis of PubMed IDs as input. These can either be chosen randomly from all available PubMed IDs listed in COSMIC or read from a file listing them as chosen. A file listing all relevant PubMed IDs can e.g. be obtained with any PubMed search from the website http://www.ncbi.nlm.nih.gov/pubmed.

gemuline uses the PubMed IDs to retrieve XML files using the provided eFetch utility [16]. Figure 5 shows an overview over text retrieval. The abstracts are being extracted from the XML files subsequently. The entire abstract of a publication is part of the XML. The user can choose to also fetch available full text articles from PubMed central [16]. The tool will automatically convert PubMed IDs to PubMed central IDs using an online interface made available through the NCBI and download the full text PDF files. Conversion from PDF to plain text then is the next step. For conversion, Apache PDFBox (http://pdfbox.apache.org) is integrated into the workflow.

Information retrieval

Running MutationFinder on the plain text abstracts yields a list of all possible point mutations including their position in the text. To find genes within the text, either GNAT or the NCIBI Name Tagger can be

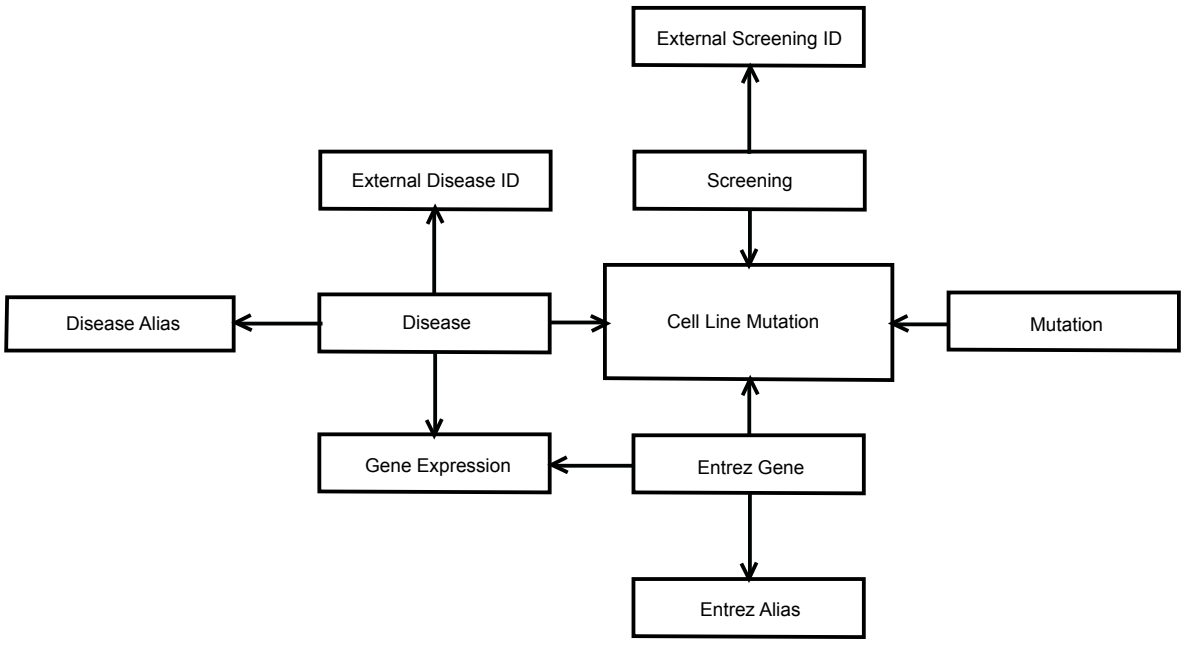

Figure 3: Simplified ER-Diagram of the database feeding *gemuline*.

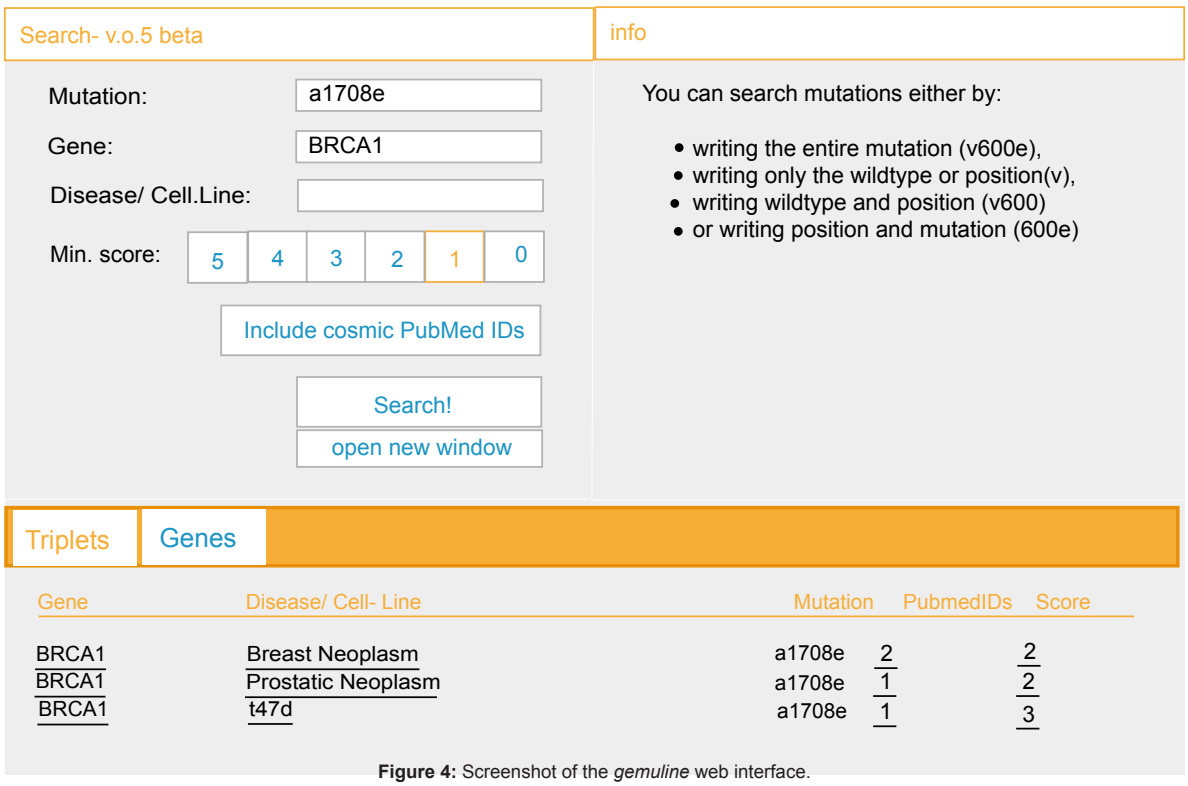

Figure 4: Screenshot of the *gemuline* web interface.

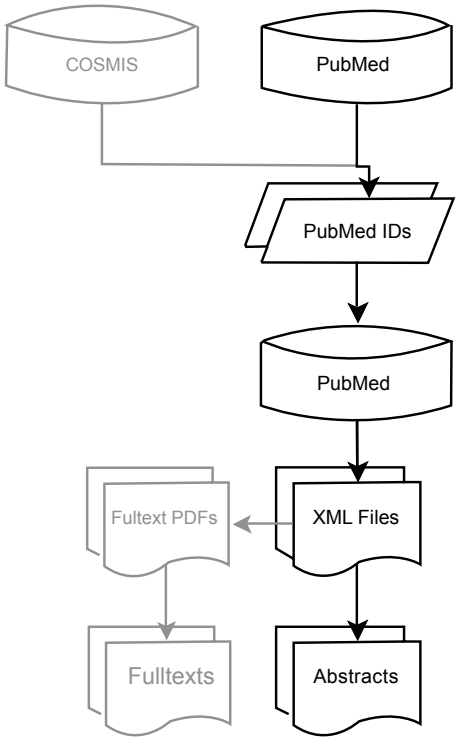

Figure 5: Text retrieval workflow in *gemuline*. Typically, PubMed IDs result from a PubMed query. Optionally, the tool can receive any type of abstract. The tool retrieves XML files from PubMed and extracts the abstract from the XML files. Optionally, the tool can download full text versions of publications.

chosen. Both result in a mapping of Entrez Gene IDs to positions in the text.

The association of genes extracted by GNAT or the NCIBI Name Tagger to mutations extracted by Mutation Finder is done next. A relation among an individual gene and an individual mutation is considered valid, if the Uniprot database has an entry of the given mutation associated with the gene. For validation, *gemuline* queries Uniprot for the specific gene, retrieves a list of registered mutations and checks for the mutation extracted from the abstract. If the mutation is recorded within Uniprot, the association is marked valid. In case Uniprot has no relationship in its database, the gene's sequence in FASTA format is investigated. *gemuline* compares the wild type and the position of the extracted entity to the gene's wild type at that location. A mutation can be validly connected to a gene if wild type of extracted entity and genomic sequence concur. Should multiple possible associations result from this procedure, the valid gene with the shortest distance to the mutation is chosen.

After associating genes and mutations, the next step is to assign the associated pair to disease models in form of cell lines or to disease terms. For this purpose, we have developed a dictionary-based disease extractor. Its dictionary was assembled from MeSH disease terms and all cell line names within COSMIC. It contains 6,009 disease terms and cell line names, and an additional 15,164 aliases. Retrieval of terms from text was done after removal of non-alphanumeric characters, utilizing case-ignorant matching with the following regular expression:

(^|[\\s\\(\\[\\'\"/,\\-])(ALIAS)([.,\\s)\\]\\'\":;\\-?!/]|$)

First, the tool searches for cell lines. Only if no cell line name can be extracted, it searches for MeSH neoplasm terms. Furthermore, a

file with stop words (http://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words) prohibits the false recognition of diseases, which are equal to common English words. If the algorithm extracts only one cell line or disease name from the text, associating the mutation-gene-pair to that one disease is reasonable. However, in case the algorithm finds multiple diseases, it is important to distinguish between right and wrong associations. Altered gene expression in cell lines is a strong indicator for the existence of a mutation [17-19]. It is therefore most likely that the gene and its mutation are associated to the cell line with the most explicit change in gene expression. Therefore, *gemuline* compares the gene expression values in cell lines to the genes and cell lines extracted from text; thus finding the match with the most explicit expression. Did *gemuline* not extract any cell lines, but only disease terms, the tool chooses the term with the shortest distance to the gene.

The resulting triplets of associated gene, mutation and disease are scored based on how they were matched. The overall score is the sum of the individual scores, thus resulting in a score between one and four, where one is the worst rating and four is the best. The score depends on mutation-gene association and on gene-disease association. For mutation-gene association, the score improves if Uniprot already includes a reference to the extracted mutation. For gene-disease association, a cell line scores higher than a MeSH disease term, because it is less likely to be extracted falsely. Furthermore, the score improves, if the gene has a prominent expression in the cell line. The score is the sum of the mutation-gene rating and the gene-disease rating. Table 1 summarizes scores according to the type of match.

Implementation rationale

Gemuline was implemented in Java. Java was chosen because most of the existing tools that can be incorporated also have a Java version available (Mutation Finder, GNAT, Moara). The architecture of the tool was divided logically. The main package contains only the main class to handle the workflow and input arguments. In object Mapping are classes representing database objects. Xml Parsing contains Simple API for XML (SAX) parsers and handlers for the various types of XML files. Finally, utils consists of all classes executing parts of the tool. For example, cell line extraction or retrieving a gene's sequence in FASTA format.

The tool must be able to run on different machines with different databases and most importantly with different usernames. For the purpose of easily changing these values, *gemuline* supplies a configuration file from which necessary information is read on start-up. The configuration file contains amongst others different database connection strings (e.g. jdbc:oracle:thin:@example.com:1521:exampledatabase), usernames and passwords as well as directory locations and proxy configuration. Furthermore, it stores the base URLs for web accesses to PDF full texts.

Implementation

The package object Mapping has a class for each entity and a class *GemulineDatabase*, which allows for saving and reading of all entities to and from *gemuline's* database. utils holds a general database interface (*IDatabase*) and its implementation together with different parts of *gemuline*, i.e. *CellLineExtractor* and *GnatLoader*. Finally, the package xml Parsing has classes to read different kinds of XML files, e.g. from MeSH, Medline and Array Express. Additionally, the following libraries are imported: bc3 (GNAT), jakarta-oro (Perl5 regular expression usage for disease / cell line extraction), mutationFinder (to extract mutations from text), ojdbc (Oracle database driver), pdfbox (to extract text from pdf files).

	Match	Score
Mutation, Gene	Gene's sequence	0
	Uniprot Reference	1
Gene, Disease	MeSH term	1
	Cell Line name	2
	Cell Line name and gene expression	3

Table 1: Scoring of results.

Usage

To run the tool, the system must provide the following requirements:

a) JAR file of *gemuline*,

b) Internet access for the NCBI Name Tagger or access to the GNAT database,

c) Access to the tool's own database,

d) Access to the COSMIC, Uniprot and ArrayExpress databases,

e) All of the following required JAR libraries must be in a lib/ directory or elsewhere accessible via classpath:

1. Mutation Finder
2. GNAT
3. OJDBC
4. PDFBox (if desired)

Then, the tool can be run from the command line using

java –jar Gemuline.jar [args]

Omitting the arguments or using -h or --help prints the help, listing all available arguments (see the Supplementary File, Readme and source code for further details).

Corpus preparation for precision and recall evaluation

For evaluation, a manually generated annotated corpus was used. To build this corpus, 150 abstracts were randomly selected from COSMIC, in order to have a set of literature with high occurrence of mutations, diseases and genes derived from the Gold Standard. They were then filtered using Mutation Finder leaving 111 abstracts in which at least one mutation was identified. All of these abstracts were independently curated by two domain experts. This corpus was then analyzed by two runs of our workflow, utilizing one of the two incorporated gene recognition tools (i.e. GNAT and the NCIBI Name Tagger) each time.

Results

Experimental evaluation

In order to determine precision and recall of each of the tools within our workflow, evaluation runs were conducted on the prepared corpus. Results were compared to the manual annotations. Utilizing GNAT, our workflow achieved a precision of 86.8% and recall of 30.3% (F1-Score: 0.449). It yielded a precision and recall of 70.3% and 23.9% (F1-Score: 0.356) with the NCIBI Name Tagger. Almost all of the false positives (i.e. 90%) derive from abstracts explicitly stating that a mutation, a gene and a disease are not related. For example: "We did not observe any correlation between the Ser1245Cys polymorphism of

the hOGG1 gene and gastric cancer, including subjects with impaired DNA repair and/or high levels of endogenous oxidative DNA lesions" [20].

Benchmarking COSMIC

In order to benchmark our workflow to the Gold Standard COSMIC, we applied *gemuline* to extract mutations, genes, cell lines (or disease names) and their relationships from roughly 127,000 PubMed abstracts not listed in COSMIC version 49 using the following PubMed query:

(Cancer OR carcinoma OR tumor OR tumour OR carcinoid OR adenocarcino-ma OR neoplasm) AND (cell OR cells OR cellline OR celllines OR "cell line" OR "cell lines" OR cell-line OR cell-lines) AND (mutations OR mutation OR SNP OR SNPs)

The workflow extracted 1,978 distinct combinations of abstract, mutation, gene and disease not listed in COSMIC. Results were found in 1,420 texts listing 264 distinct genes associated with 202 distinct disease terms and 594 distinct mutations. Thus, a total of 1,258 unique combinations of gene, mutation and disease were found, as opposed to COSMIC listing 143,716 of such distinct combinations. We found that for 61% of the genes to which a mutation was associated by using *gemuline*, fewer mutations are associated in COSMIC. Among others, we found several mutations annotated to BRCA1 that were not listed in COSMIC.

Discussion and Conclusion

The main goal of our study was the automated extraction of genes, mutations, diseases and their relation among one another within one workflow. Furthermore, a normalization task was necessary to map the results to established databases containing genes, mutations or diseases. To the best of our knowledge, *gemuline* delivers associated genes, mutations and diseases from literature with the highest precision and competitive recall to date. Freely available tools were integrated into a new workflow extracting these relations with a precision of 87% and a recall of 30%. Compared to the results from previous works, the overall precision of the workflow ranges in the intermediate field.

Related methods

MutationFinder finds and normalizes point mutation mentions in free text [10]. Regular expressions provide the ability to find certain matching patterns in the text. Mutation Finder utilizes these regular expressions for recognizing the specific notations of point mutations. It builds upon a baseline system, which is a partial reimplementation of MuteXt [21]. Mutation Finder achieves a precision of 97.5% and a recall of 80.7% regarding the extraction of normalized mutations. It extracts mutations only, not genes, diseases or disease models.

Yip et al. tried another approach to extract mutation mentions using regular expressions [22]. They generated the regular expressions manually by reviewing several hundred abstracts. Their approach yields a precision of 89.3% and a recall of 84%. Since precision is lower using this method, we selected Mutation Finder to be part of our workflow in this study.

GNAT [8] searches text for mentions of genes and normalizes each gene to an Entrez Gene ID. GNAT consists of a multi-step procedure of refining an initial set of predictions until a final conclusion is reached. Its main steps are named entity recognition (NER), validation, correlation and disambiguation. On a test set with human genes, GNAT achieved a precision of 90.1% and a recall of 81.1%. It extracts genes only, but neither mutations nor diseases or disease models.

The National Centre for Integrative Biomedical Informatics (NCIBI) provides an online name tagger, which tags genes in PubMed abstracts [9]. It is not stated how the algorithm actually works, but it provides online access to pre-processed PubMed and PubMed central articles tagged with genes. Thus, it offers very quick access to gene mentions in text. It only extracts genes but neither mutations nor diseases or disease models. Its exact precision and recall values are unknown, but benchmarks performed in the course of this work have shown that it performs worse than GNAT which is why GNAT was chosen here.

In their study, Chun et al. first use a dictionary technique to find entities (genes and diseases) and then filter results by machine learning [23]. For extraction, they select sentences with at least one gene and one disease in it. Then, they associate the entities. They compile dictionaries from public biomedical databases, also extracting the primary symbols and unique IDs for all entities. They achieve a precision of 89.0% for genes, 90.0% for diseases and a recall of 90.9% for genes, 96.6% for diseases. Without filtering, their relation extraction has a precision of 51.8%. With filtering, it has a precision of 78.5%, but the recall drops to 87.1% of the unfiltered recall. For filtering, they use a maximum entropy model.

A very recent study by Doughty et al. presents an approach utilizing the "EMU method" (Method for extracting mutations from the biomedical literature) [4]. The EMU algorithm extracts mutation information as well as gene-related data. EMU mutation extraction utilizes regular expressions to find mutations, where the input is plain text and the output is a list of mutation terms. The algorithm uses two sets of regular expressions; one to identify possible mutations and another one to deselect wrong hits. The false patterns include, among others, 6,541 cell line names. For identifying gene information, a dictionary containing all gene names from the Human Genome Organization (HUGO) and from the National Center for Biotechnology Information (NCBI) except those identical to codon names is generated. To associate genes and mutations they compare the mutation wild type and its position to the gene's protein sequence in RefSeq [24]. This step, named "SEQ Filter", flags falsely identified gene-mutation relations and removes them from the results. So far, the method is restricted to prostate and breast cancer, but it is claimed that additional diseases could easily be incorporated. From 1,721 abstracts relating to prostate cancer and 5,967 relating to breast cancer, 179 manually verified mutations that are not currently annotated in the Online Mendelian Inheritance in Man (OMIM) or SwissProt databases were identified. Thus, the method yields an average precision of 60% on a SEQ-filtered corpus.

Table 2 gives an overview over many of the considered methods, their precision and recall values and the types of entities and relations they extract.

Precision and recall

Most of the tools reviewed in the course of this work do not associate entities, but only recognize them. Precision and recall of *gemuline* consider both, recognition and association and are counted towards precision and recall only when both steps are successful. The EMU method is most similar to *gemuline* in terms of functionality. In direct comparison, however, *gemuline's* precision of 87% is superior. Doughty et al. give no statement regarding recall. Of two gene taggers that were tested, GNAT achieved better scores, but the NCIBI Name

Tool	Mutation	Gene	Disease	Precision	Recall	Ref
MutationFinder	+	-	-	97.5%	80.7%	[10]
BANNER	+	+	+	85.09%	79.06%	[25]
ABNER	+	+	+	Depends	Depends	[26]
ProMiner	-	+	-	82.53%	82.93%	[27]
GNAT	-	+	-	90.1%	81.1%	[8]
Moara	-	+	-	GNAT	GNAT	[28]
NameTagger	-	+	-	Unknown	Unknown	[9]
OSIRISv1.2	+	+	-	99%	82%	[29]
Extraction of Gene-Disease relations	-	+	+	78.5%	Unknown	[23]
SNPshot	+	+	+	96.0%	88.7%	[30]
EMU	+	+	(+)	60%	Unknown	[4]

The table shows the name of each tool or algorithm, which of the three types of named entities can be extracted, their respective precision and recall values (if mentioned by the authors) and the reference to the original publication.

Table 2: Overview over text-mining algorithms reviewed for this study.

Tagger was much faster. GNAT took several days to process roughly 127,000 abstracts in an unparallelized manner. The NCIBI Name Tagger's speed was only limited by internet access and connection speed, as PubMed abstracts are pre-annotated.

A low recall (i.e. occurrence of false negatives) is solely owed to the failed attempt of extracting all of the three required entities mutation, gene and disease, even if all three entities should have been found in the text. Depending on the algorithm, for example regular expressions did either not contain the right pattern, or dictionaries did not contain the term in question. In other algorithms (i.e. GNAT), the heuristic failed at some point to assign an Entrez Gene ID to a gene mention in the text.

GNAT and the Mutation Finder both have a recall of 81%. While Mutation Finder found a mutation in every abstract, GNAT found a gene in all but 6 of them. Thus, disease and cell line extraction failed in at least 71 abstracts resulting in a recall of 35.69% at best. As described above and summarized in Table 3, such dictionary-based approaches score much lower, since the dictionaries do not contain all synonyms typically used in the texts. This is most likely an artifact due to the use of domain-specific language in many biomedical publications. This has been noted in previous text-mining studies [31]. Specifically, the notations are not always consistent with MeSH. For example, the abstract from PubMed ID 9827921 explicitly states "acute lymphoblastic leukaemia (ALL)". However, the dictionary covers only the term "acute lymphoblastic leukemia" because it is the official MeSH term without having the above term denoted as a synonym. Another example is in the abstract from PubMed ID 18544621. The term "adrenocortical tumors" is not in the dictionary, only "adrenocortical cancers".

gemuline found cell lines in 21 of 30 abstracts containing a confirmed cell line mention. For example, *gemuline* did not find the cell line named "NIH3T3" in the abstract from PubMed ID 19802009, because that cell line was not listed in the COSMIC version 49 used to generate the dictionary.

Outlook

To date, *gemuline* only extracts point mutations. An improvement could be the inclusion of algorithms that do not only extract point mutations, but also linguistically more complex genomic alterations. We expect an increase in recall, if the tool extracted other types of mutations as well considering the many publications that include other types of mutations apart from point mutations (e.g. fusions, chromosomal rearrangements, etc.). In order to do so, other tools besides Mutation Finder need to be incorporated or Mutation Finder's list of regular expressions must be expanded to fit these notations. However, this would most likely lead to a decrease in precision as it will become more complicated to determine possible occurrences of more mutation types in natural language. Thus, mutation extraction with high recall and high precision remains a challenge.

In a next step, the addition of more journals with full text access should be considered, especially when they have a focus on cancer. To achieve this, different URLs for different journals and their respective login credentials need to be incorporated within the tool. For many journals, PubMed XML provides journal title and access link to the article via its PubMed ID. However, other journals are not accessible as easily. Nature, for example, does not provide means to access a full text article through this method. To download such articles, the PubMed website source code needs to be parsed for the appropriate link. Usually, the availability of the full text PDF is not certain before the query, leading to a computational overhead.

In a similar way, the analysis of supplementary material could increase recall significantly. However, currently it is not obvious how to gain automated access to supplementary material. A feasible but tedious approach could be the manual acquisition and automated extraction of entities and relations from such files.

The extraction of results from articles stating that a relation is not given could be prevented, if the tool scanned for appearances of, for example, "not" within the sentence containing the derived entities, also increasing performance.

The possibility of manually annotating *gemuline's* results by scientists can be used to enhance the extraction process further. It could be analyzed, for example, if patterns within frequent false

Reason for false positive	Percentage of FP	Reason for false negative	Percentage of FN
Negation	90%	Term not described in dictionary/regular expression (i.e. MutationFinder)	95%
ambiguous terminology	8%	Term not recognized by heuristic (i.e. GNAT)	5%
Lack of association between terms	2%		

Table 3: Reasons and distribution of false positive and false negative extractions.

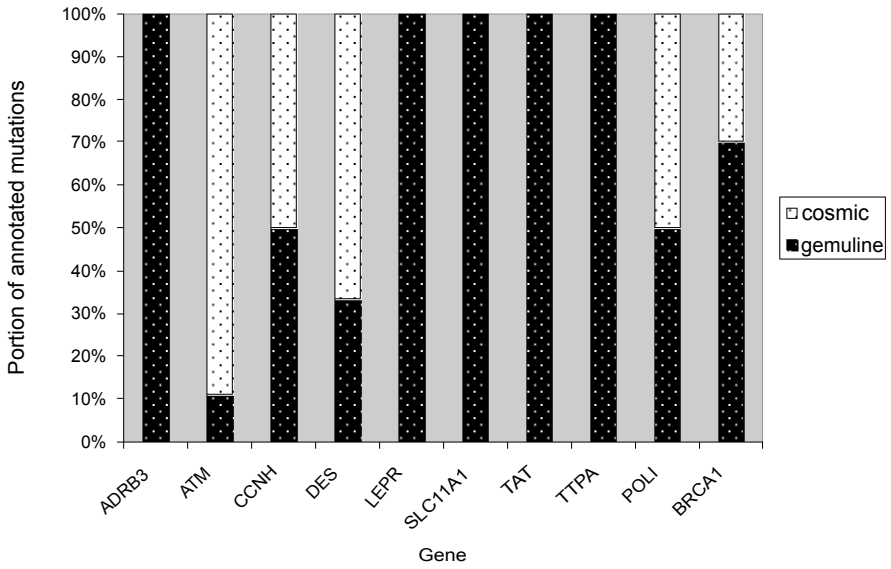

Figure 6: This figure depicts the amount of annotated mutations in COSMIC (white) and as a result from our workflow (black) for a random selection of nine genes and BRCA1. For example, the exact same amount of mutations (not necessarily the same ones) is annotated to CCNH in COSMIC and gemuline. For BRCA1, 70% of all mutations from either COSMIC or gemuline derive from our workflow.

positives emerge. In such cases and the algorithm could be adapted to such patterns, resulting in a higher precision.

Regarding the entire workflow, it seems clear that the highest impact on improved precision and recall could be achieved by enhancing the disease extractor. Adding additional terms and enriching the list of synonyms for each term will greatly improve the extractor's recall which is currently the lowest within the workflow. Another approach to the problem could be pre-annotating the biomedical texts with terms from more general ontologies in order to diminish the impact of domain-specific vocabulary. However, this step would require tedious, manual work from experts in the field and a very good ontology covering all disease terms and popular synonyms.

Still, *gemuline* offers a good basis to improve productivity within the drug discovery process. Already, it enables scientific researchers to go beyond COSMIC and investigate new potential oncogenes and their mutations in cancer. The tumor suppressor genes BRCA1 and BRCA2 are associated with high risks of breast, ovarian and contralateral breast cancer [PMID: 22144499]. The lifetime risk of breast cancer in women with a *BRCA1* or *BRCA2* mutation is approximately 75%. For *BRCA1*, there is little evidence that the risk varies for different mutations [PMID: 22127115]. BRCA1 is a checkpoint and DNA damage repair gene that secures genome integrity [32]. A mutation in this gene may lead to genomic instability, resulting in the accumulation of mutations and eventual cancer development [32]. However, not all mutations of BRCA1 play such an important role in developing cancer.

As an example, an application of *gemuline* has extracted the BRCA1 mutation A1708E which is yet unreported in COSMIC but associated with breast cancer. Lovelock et al. have shown a severe functional abrogation of BRCA1 proteins carrying the A1708E mutation [33]. They further demonstrate that the histopathology of A1708E-associated tumors have a typical BRCA1-like phenotype. This example shows the importance of finding additional mutations not yet in COSMIC from literature and their possible relevance for cancer research. In Figure 6 we summarize for a random selection of nine genes and BRCA1, how application of our workflow increases the amount of mutation annotations for genes. For example, BRCA1 is annotated with at least twice as many mutations after application of our workflow.

Certainly, for genes of high interest like the Androgen Receptor in Prostate Cancer or BRCA1 in Breast Cancer, there are specialized mutation databases (see e.g. the Breast Cancer Information Core database http://research.nhgri.nih.gov/bic/) annotating many more mutations than COSMIC and gemuline. However, we extract mutations from the scientific literature without focus on any gene or sub-type of cancer. We gain a reasonable benchmark with this approach only when comparing our results to the largest public repository on mutations across indications (i.e. COSMIC). This is a use case, showing that the results of *gemuline* can improve the data availability for oncology research. It can easily be adapted to a more focused approach, e.g. when a limited number of genes or only a sub-type of cancer is of interest. We expect higher precision and recall in such a case.

We have shown in thus study using most of the relevant literature from PubMed that *gemuline* is detecting new mutations unreported in COSMIC with reasonable quality, ultimately enabling new discoveries.

Acknowledgement

The authors would like to acknowledge funding from the BMBF (Federal Ministry of Education and Research), MedSys projects PREDICT (0315428F) and OncoPred (0315416B).

References

1. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, et al. (2002) Mutations of the BRAF gene in human cancer. Nature 417: 949-954.
2. Harris T (2010) Does large scale DNA sequencing of patient and tumor DNA yet provide clinically actionable information? Discov Med 10: 144-150.
3. Sharma SV, Haber DA, Settleman J (2010) Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. Nat Rev Cancer 10: 241-253.
4. Doughty E, Kertesz-Farkas A, Bodenreider O, Thompson G, Adadey A, et al. (2011)Toward an automatic method for extracting cancer- and other disease-

related point mutations from the biomedical literature. Bioinformatics 27: 408-415.

5. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2006) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 34: 173-180.
6. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, et al. (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet 10: 10.11.
7. Hakenberg J, Plake C, Royer L, Strobelt H, Leser U, et al (2008) Gene mention normalization and interaction extraction with context models and sentence motifs. Genome Biol 2: S14.
8. Hakenberg J, Plake C, Royer L, Strobelt H, Leser U, et al. (2008) Inter-species normalization of gene mentions with GNAT. Bioinformatics 24: 126-132.
9. Ade A, Wright Z, Jagadish H (2009) The NLP web service.
10. Caporaso JG, Baumgartner WA Jr, Randolph DA, Cohen KB, Hunter L (2007) MutationFinder: a high-performance system for extracting point mutation mentions from text. Bioinformatics 23: 1862-1865.
11. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, et al. (2007) Array Express--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 35 : 747-750.
12. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, et al. (2004) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34: 187-191.
13. Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 35: D26-31.
14. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2011) Ensembl 2011. Nucleic Acids Res 39: 800-806.
15. Lipscomb CE (2000) Medical Subject Headings (MeSH). Bull Med Libr Assoc 88: 265-266.
16. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2011) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 39: 38-51.
17. Achan V (1999) An introduction to molecular biology: gene structure, expression, and mutation. Pediatr Cardiol 20: 94-96.
18. Jiang Y, Zhou XD, Liu YK, Wu X, Huang XW (2002)Association of hTcf-4 gene expression and mutation with clinicopathological characteristics of hepatocellular carcinoma. World J Gastroenterol 8: 804-807.
19. Langerød A, Zhao H, Borgan Ø, Nesland JM, Bukholm IR, et al. (2007) TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. Breast Cancer Res 9: 30.
20. Poplawski T, Arabski M, Kozirowska D, Blasinska-Morawiec M, Morawiec Z, et al.(2006) DNA damage and repair in gastric cancer--a correlation with the hOGG1 and RAD51 genes polymorphisms. Mutat Res 601: 83-91.
21. Horn F, Lau AL, Cohen FE (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.
22. Yip YL, Lachenal N, Pillet V, Veuthey AL (2007) Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase. J Bioinform Comput Biol 5: 1215-1231.
23. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, et al.(2006) Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Pac Symp Biocomput: 4-15.
24. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35: 61-65.
25. Leaman R, Gonzalez G (2008) BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. 652-663.
26. Settles B (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 21: 3191-3192.
27. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J (2005) ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics 6: 14.
28. Neves ML, Carazo JM, Pascual-Montano A (2010) Moara: a Java library for extracting and normalizing gene and protein mentions. BMC Bioinformatics 11: 157.
29. Furlong LI, Dach H, Hofmann-Apitius M, Sanz F (2008) OSIRISv1.2: a named entity recognition system for sequence variants of genes in biomedical literature. BMC Bioinformatics 9: 84.
30. Hakenberg J, Voronov D, Nguyen VH, Liang S, Lumpkin B, et al. (2010) Taking a SNPshot of PubMed - a repository of genetic variants and their drug response phenotypes. in GPD-Rxn Workshop: Genotype-Phenotype-Drug Relationship Extraction from Text. Pac Symp Biocomput (PSB), Big Island of Hawaii, USA.
31. Groth P, Weiss B, Pohlenz HD, Leser U (2008) Mining phenotypes for gene function prediction. BMC Bioinformatics 9: 136.
32. Cao L, Kim S, Xiao C, Wang RH, Coumoul X, et al. (2006) ATM-Chk2-p53 activation prevents tumorigenesis at an expense of organ homeostasis upon Brca1 deficiency. EMBO J 25: 2167-2177.
33. Lovelock PK, Healey S, Au W, Sum EY, Tesoriero A, et al. (2006) Genetic, functional, and histopathological evaluation of two C-terminal BRCA1 missense variants. J Med Genet 43: 74-83.

This article was originally published in a special issue, **Bioinformatics** handled by Editor(s). Dr. Yixuan Wang, Albany State University, USA