The Gene Indexing Assistant

A Prototype Application for Partially-Automated Gene Indexing

NLM Associate Fellow: J. Caitlin Sticco

Project Sponsors: Lou Knecht and Deborah Ozga

Project Collaborators: Lan Aronson, Jim Mork, and Antonio Jimeno Yepes

Project Interns: Amy Dunaway-Knight and Anderson Wright

August 2011

Table of Contents

Abstract

Background

A Gene Reference Into Function (GeneRIFs) is a concise entry in an Entrez Gene record that summarizes novel information about the gene function or structure from scientific literature. Every year, the National Library of Medicine Index Section manually creates upwards of 80,000 hyperlinked geneRIFs from articles indexed for MEDLINE. The creation of these hyperlinked geneRIFs is known as gene indexing.

Objective

To create a prototype application for assisting in gene indexing by identifying genes in citations, suggesting a link to the appropriate Entrez Gene record, and suggesting several candidate sentences from which to derive geneRIFs.

Methods

The scope of this project was limited to human genes, and the citations used in our development were extracted from 43 journals of human genetics indexed for MEDLINE between 2002 and 2011. Rapid-prototyping was used to create an initial modular end-to-end application. Each module was then iteratively revised. Citations were filtered for gene indexing based on rules used by the Index Section and the presence of gene mentions in the citation. To avoid confounding non-human genes, abstracts that mentioned any non-human species were also filtered out for this prototype stage. For gene mention identification, we developed a dictionary-based approach after lackluster performance from existing conditional random fields (CRF)-based software. Candidate sentences for geneRIFs were identified with a classifier that we trained on a manually-annotated corpus of 1,987 sentences.

Results

We currently identify explicit mentions of human genes in citations with 88% recall and 84% precision. We are able to identify and normalize them to the correct Entrez Gene record with 86% recall and 82% precision. We are able to identify candidate geneRIF sentences with 76% recall and 64% precision.

Conclusions

Accurate identification and normalization of gene mentions in citations will allow automatic links to the correct records in Entrez Gene, and should represent a significant time-savings and potential cost-savings in gene indexing. This goal is well within reach for most human genes and human-only citations. We plan to improve gene mention identification and normalization by expanding our dictionary and combining it with previously developed CRF-based applications. Our work on gene identification and normalization has benefitted greatly from the availability of related research on this topic, much of it from NLM-sponsored competitions. While we are having moderate success identifying geneRIF candidate sentences, our slow progress reflects the

greater complexity of the task and the comparative lack of previous research in similar applications. Our contributions in this area represent a novel approach, and our manually annotated dataset will be a valuable resource to offer the research community.

Introduction

A Gene Reference Into Function (geneRIF) is a concise entry in an Entrez Gene record that summarizes novel information about the gene function or structure from scientific literature. Every year, the National Library of Medicine (NLM) manually creates upwards of 80,000 hyperlinked geneRIFs from articles indexed for MEDLINE. The creation of these hyperlinked geneRIFs is known as gene indexing, and it is performed for most articles that focus on the basic biology of genes or gene products. Gene indexing has been performed by the Index Section since 2002.

The manual process involved in gene indexing is time consuming and detailed. Once an indexer has determined that an article is suitable, they use a special interface in the Data Creation Management System (DCMS) to perform the gene indexing. From here, they follow links to an external Entrez Gene interface, to manually search for each gene and species, and then import links for the relevant Entrez Gene records back into DCMS. There, they manually create the geneRIF annotation, usually using information directly from the abstract. This process is outlined in length in four Technical Memoranda from the Indexing Manual [1]-[4].

Our research group recognized many steps of this process as candidates for automation. A great deal of successful previous research has been done on automatically identifying gene and protein names in biomedical text [5]-[20], and normalizing those names to unique identifiers such as Entrez Gene IDs [6],[7],[17]-[20]. Some researchers have even attempted to automatically identify geneRIFs from MEDLINE abstracts, albeit with more limited success [26]-[28]. Building on this previous research, we sought to create a prototype program for assisting in gene indexing. We call this program the Gene Indexing Assistant (GIA).

The GIA performs many tasks that will hopefully increase the speed and comprehensiveness of gene indexing. The prototype is built with a modular design, with each of four modules fulfilling certain gene indexing functions.

1. **Citation Filter**
   This module evaluates MEDLINE citations with abstracts to determine whether or not each article is suitable for gene indexing.
2. **Gene Mention Identification**
   This module finds all mentions of genes in the citations. Gene naming is highly variable, with many different names and abbreviations for each gene in use, and many genes with identical names or abbreviations.
3. **Gene Mention Normalization**

This module attempts to normalize each gene mention to the correct Entrez Gene ID. It suggests links to the appropriate Entrez Gene records. This step is also affected by the ambiguity of gene names, as more ambiguous names as more difficult to normalize.

4. **GeneRIF Extraction**
   This module identifies and suggests the most likely candidate sentences from which to derive geneRIFs.

Accurate identification and normalization of gene mentions in citations will allow automatic links to the correct records in Entrez Gene. This alone should represent a significant time-savings and potential cost-savings in gene indexing. Our work shows this goal is well within reach for most human genes and human-only citations, while previous work indicates that success can be expected with additional species, as well. While we are having moderate success identifying geneRIF candidate sentences, our slower progress here reflects the greater complexity of the task and the comparative dearth of previous research in similar applications. Our contributions in this area represent a novel approach.

Methods

A rapid-prototyping philosophy guided the project. A modular end-to-end application was quickly created by first employing simple methods and existing resources. Each module was then iteratively revised. The process of creating each module is described below. A general overview of the system architecture can be found in Figure 1.

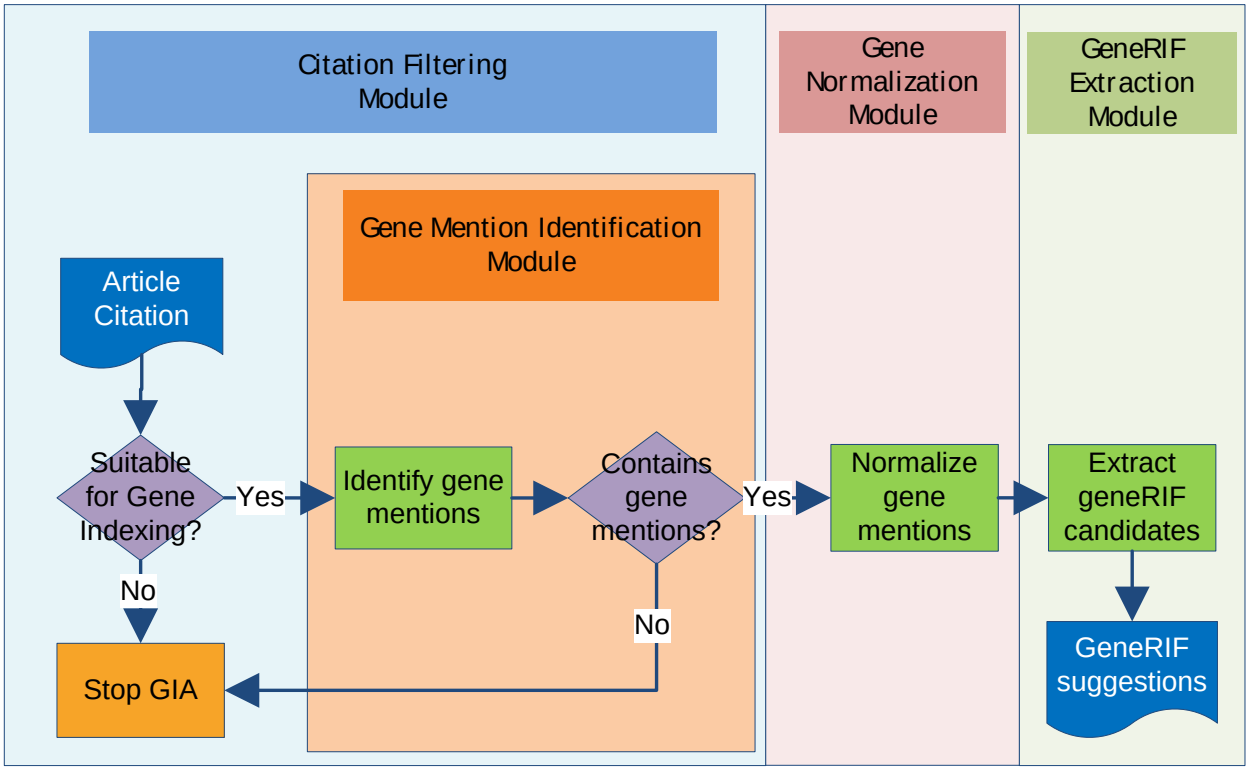

Figure 1: Gene Indexing Assistant Structure

Citation Filter

Citations are filtered for gene indexing based on the presence of gene mentions in the citation, rules used by the Index Section for which articles are suitable, and the requirements of this project.

Index Section Rules

The following list of relevant rules is adapted from Technical Memoranda [1]-[4]. The details of our implementation for each are noted where applicable.

- Link articles in which the basic biology of a gene from an in scope organism is the primary point of the article. Do not create links for articles where the focus is genetic engineering, genetic databases, sample banks, population genetics unrelated to disease or function, and any topics other than the basic biology of the gene. In clinical articles, do not create links unless the focus of that article is on some new aspect of that gene.
    - We are currently unable to implement this set of rules effectively. We do discard citations that do not contain gene mentions recognized by the Gene Mention Identification Module, but we are not able to automatically detect the main focus of articles in an abstract way.
- Do not create links for case reports with only a single patient
    - We are currently unable to implement this rule, but expect to be able to in the future. Case reports with only a single patient are considered insufficient evidence to create a genotype-phenotype link.
- Review articles may be included only if they focus on a particular gene
    - We filter out any review articles that mention more than three genes. We may be able to refine this in the future. Obviously this filtering depends on the success of the Gene Mention Identification module.
- Do not link news items, editorials or letters commenting on genes or proteins in another article.
    - We filter out all of these publication types. Although a very, very small percentage of these items may comment on a gene or protein not discussed in another article, they generally do not have abstracts from which we can work.
- Restrict to organisms that are in the taxonomy list for Entrez Gene.
    - For our prototype, we are limited to humans only. The NCBI Taxonomy is used as a dictionary to filter out any citations that mention other species. These citations may be suitable for gene indexing, but our current program is not equipped to distinguish between genes from different species.
- Restrict to abstracts with three or fewer genes, unless they are mentioned in the title
    - We do not currently implement this rule. It is designed to arbitrarily offer a limit on the number of geneRIFs created for a single article.

Additional Rules

Additional rules were created for the purposes of this project.

- Limit to citations with an abstract
    - Titles only do not offer sufficient text for processing
- Limit to articles that do not contain any non-human species mentions
    - As mentioned above, all citations with non-human species mentioned are discarded.
- Limit to articles in human-genetics journals from 2002-2011
    - The scope of this project was limited to human genes, and we required a test bed of citations about human genes exclusively. We began by selecting all the journals indexed for MEDLINE with the journal subject *Human genetics.* From these, we eliminated journals devoted to non-human species such as *Journal of experimental zoology* and those primarily concerning gene therapy, genetic counseling, or ethics such as *Law and the human genome review.* We queried the citations of the remaining journals in PubMed to determine the prevalence of confounding species, using the following 11 words for common model species as a heuristic: mouse, murine, yeast, fly, drosophila, cow, cattle, bovine, worm, elegans, and plant. Journals with fewer than 40% of their citations confounded by representative non-human species were selected to form our test bed. A total of 105,255 citations were available from these 43 journals in the selected time period. From these, we randomly selected our training and testing materials.

Gene Mention Identification

Gene Mention Identification is a well-studied task in Named Entity Recognition (NER). NER seeks to extract from text any words or phrases from predefined categories such as genes, proteins, or diseases. Sponsored contests specifically targeting Gene Mention Identification, such as BioCreative 2 and 3, have been successful not only at attracting participants, but at spurring subsequent research [5],[7]. There are a number of barriers to building on previous research from this area, however.

1. Many of these programs are not made public. Only a small fraction of the entries from the BioCreative challenges were made available
2. Many of those made public are not open source, or are available only as web services
3. Those made public may be difficult to implement due to lack of documentation and user support
4. Those made public may not be updated regularly

Since these competitions are intended to advance the state of the art, the resources derived from this kind of sponsored research is something that should be carefully considered in designing the competitions.

We identified a number of high-performing applications for Gene Mention Identification resulting from competitions and other research. Our list is likely not comprehensive. What we did find is represented in Table 1.

Table 1: Gene Mention Identification Tools

Tool	Description	Access	Reference
Tools for Gene Mention Identification Only			
ABNER	Problems with documentation	Downloadable	[8]
AIIA-GMT	Uses genia , Mallet/CRF++	Web service only	[9]
ando	Unusual semi-supervised learning method	Cannot locate	[10]
BANNER	Problems with documentation	Downloadable	[11]
Biotagger (Penn)	not recently updated	Downloadable	[12]
BioTagger-GM	Combination of four systems. Uses Genia, Mallet, SRDEF	Build your own	[13]
lingpipe	Lower-performing general system that can be trained to genes	Downloadable	[14]
NERbio	Cannot locate	Cannot locate	[15]
WCL BioNER	A method for combining a variety of tools	Build your own	[16]
Tools for Gene Mention Identification and Normalization			
GeNo	Remote UIMA-AnalysisEngine--full gene normalization suite. Entity tagger is JNET.	Web service only	[17]
GNAT	Uses BANNER and LINNAEUS; cross species full suite	Downloadable or web service	[18]
moara	Full suite, freely available, now integrated in U-Compare	Downloadable	[19]

We originally attempted to simply use an existing tool for gene mention identification. We first selected BANNER, because it reported high scores compared to other tools on a variety of test collections. BANNER is also a relatively well-known application, due to its success and availability. We assumed that this would also make it easier to implement. Instead, we found that documentation on its functions was so insufficient as to render it virtually impossible to implement without assistance from its developers. We encountered the same problem with ABNER.

We turned then to AIIA-GMT, which also scored well on test collections. AIIA is available as a simple web interface. It was always clear we would eventually need access to a complete

program, not just an interface, but this allowed us to experiment until another solution could be devised. We were unsatisfied with the performance of AIIA, specifically, with inexplicably missed identical mentions within the same abstract.

In keeping with a rapid prototyping philosophy, we decided not to let the project become mired in the details of implementing and troubleshooting a previously created gene mention identification application. To ensure that we were intimately familiar with the workings of our program and had control over its shortcomings and development, we chose to create our own gene mention identification module. We developed a simple, dictionary-based approach which we eventually supplemented with additional rules and preprocessing.

Dictionary-Based Gene Identification

In our approach, words in a citation are simply matched to a dictionary of gene names. Our dictionary terms are based on gene designations from the following sources:

- Human Gene Names from NCBI's Homo_sapiens.gene_info.gz [21]file which provides *Homo sapiens*-specific gene information. We are using the following fields: tax_id, GeneID, Symbol, Synonyms, description, and Other designations.

- Discontinued Gene Names are from NCBI's gene_history file [22]. This file covers all species, but we have retrieved just the human-specific genes using the tax_id field (9606). We then used the first four fields (tax_id, GeneID, Discontinued GeneID, and Discontinued Symbol) to identify the discontinued genes and link them to their current replacements using the "GeneID" field. When available, this field links to the current gene that replaced the discontinued gene.

- Online Mendelian Inheritance in Man (OMIM) numbers are from NCBI's mim2gene file [23]. This file provides links between the OMIM number and their corresponding gene identifier. For creating the dictionary, we ignored the "phenotype" entries and focused solely on the "gene" entries. Using the list of human genes identified above, we filtered this file to identify only the human-related OMIM numbers.

Once we had built a comprehensive list of gene names from the NCBI resources, we removed the duplicates and then filtered out some misleading or ambiguous gene names. We removed any gene name that ended with any of the following terms: 'disease', 'syndrome', or 'susceptibility'. For example: *619509|Wittwer syndrome*, *7439|Best disease*, and *220296|cancer susceptibility*. In text, our program cannot differentiate between mentions of these phenotypes and the genes with identical names. However, in most cases it is the disease and not the gene, so performance is currently improved by ignoring these mentions.

The dictionary was then expanded with variants of each term to account for author preferences in punctuation and spacing. Variants were created for gene names that have a single dash in them. The variant generation algorithm creates a version replacing the dash with a space, and another

variant with the dash simply removed. For example, “cortexin-2” would generate “cortexin 2” and “cortexin2” as variants. Variant generation will be improved in the next phase. One possibility will be to use one of the existing tools with a more advanced variant generation or identification algorithm.

Each resulting entry in the dictionary is linked to its originating Entrez Gene ID. In the case of an entry that is common to multiple Entrez Gene records, all relevant IDs are associated with the dictionary entry. The final list was than sorted into longest to shortest gene name order to facilitate identifying the longest possible matches in the text before identifying a component of the gene name.

Sentences in an abstract are detected and tokenized using MetaMap [24]. We initially wrote our own simple sentence detector and tokenizer, but implemented MetaMap on a subsequent revision of the module. MetaMap is used for sentence detection, tokenization, acronym and abbreviation identification, as well as during geneRIF extraction, benefitting the entire GIA application.

Gene Mention Normalization

Gene mention normalization links each gene mention to the appropriate Entrez Gene ID as a unique identifier. Normalization was a low priority for this prototype and did not receive significant revision from the first iteration of this module. Thus, our current normalization step is very simple. Since our dictionary is primarily built from Entrez Gene, all of our dictionary terms are linked back to their originating Entrez Gene records. Therefore, if we find a mention at all, we have also found at least one Entrez Gene ID. If only one Entrez Gene ID is returned, the mention is normalized to that as the only possible option. In the case of ambiguous terms, a list of all matching Entrez Gene IDs is returned, and we disambiguate between them.

When we have multiple possible IDs, we currently use a very simple strategy for disambiguating: We determine whether the match is an official name or official abbreviation, versus a lesser-used type of synonym. The decision process follows this sequence of preferences:

1. Prefer an official name or abbreviation
2. Prefer a mention that corresponds to an official name identified elsewhere in the article.
3. Prefer a designation from the “Synonyms” field
4. Prefer a designation from the “Description” field
5. Prefer a designation from the “Others” field
6. Otherwise select the first mention in the list

GeneRIF Extraction

GeneRIF extraction is the process of finding the sentences in the abstract most likely to be selected by the indexer as geneRIFs for each gene mention. At the moment, our project is focused on simply identifying all the candidate sentences, and has not yet attempted to rank

them or link them to specific gene mentions. This module will suggest the best sentences to the indexer.

Original Methodology

We initially hoped that existing geneRIFs and their originating abstracts could be used as a gold standard to train a geneRIF/non-geneRIF classifier. We excluded geneRIFs that did not originate from NLM indexers, as they tend not to meet the Index Section guidelines. Existing geneRIFs that originate from indexers were obtained from a file kept by DCMS. These geneRIFs were compared to their originating abstracts with a string matching algorithm, to identify the sentences from which the geneRIFs were derived. Thus, a corpus of "geneRIF sentences" and "non geneRIF sentences" was formed from these abstracts.

These sentences were analyzed for their position in the abstract, as well as the frequency of unigrams, bigrams, trigrams, and quadgrams within them.Very few predictive features were identified. The results of these analyses are presented and discussed in more length in the Results section. The factors that led to the failure of this tack were found to be inherent to using existing geneRIFs as a gold standard. Our insights from this setback shaped the strategy of our subsequent approach.

Current Methodology

Classes

Three types of classes were devised, as shown in Table 1. GeneRIF classes are based on the Gene Indexing guidelines as outlined in Technical Bulletins [1]-[4] and our observations from existing text. Discourse Classes represent the rhetorical part of the abstract the sentence is in, like the background or the results, and are based on the Structured Abstract components used by NLM [25]. Claims Classes represent whether or not the sentence makes a scientific claim, and if so, whether that claim is an established fact or something proved by the author. Our hypothesis was that Discourse and Claims Classes could be used to enhance the prediction of geneRIF Classes during classification. The full guidelines used to define classes and determine tagging are available in Appendix A: Guidelines for Class Assignment.

Table 2: Classes for Annotation

GeneRIF Classes	**Discourse Classes**	**Claims Classes**
• Expression • Function • Isolation • Reference • Structure • Non-geneRIF • Other	• Title • Background • Purpose • Methods • Results • Conclusions	• Established Claim • Putative Claim • Non-Claim

Dataset Creation

The abstracts in our training and test sets were selected randomly from the test bed described under the Citation Filter's Additional Rules. From here, 373 abstracts were selected for the training set, and 151 for the test set. All sentences were processed by the Gene Mention Identification module to tag gene mentions. JCS then manually tagged for the incorrect and missed gene mentions, and marked each article for suitability for gene indexing.

For abstracts considered suitable for gene indexing, JCS and ADK independently coded all sentences with at least one geneRIF Class, at least one Discourse Class, and one Claims Class. A total of 2986 sentences were coded with these classes. After calculating the intercoder reliability of the two coders, the discrepancies were reconciled by discussion to gain consensus. Both the original data and reconciled data are available as supplementary data.

Feature Evaluation

The features selected for analysis reflect both standard practice for text classification and literature and observations specific to geneRIFs.

Unigrams and bigrams

The occurrence of N-grams is a standard feature for text classification.

Sentence Position

Sentence position is well-known to influence the likelihood of a sentence being a geneRIF. Previous analysis has shown a high percentage of geneRIFs come from the title or the final sentences of an abstract [26], [27]. Our own analysis of human-related geneRIFs in MEDLINE confirmed this, as shown in Table 3.

Table 3: GeneRIF Origin Sentences

	Number of geneRIFs	Percent of geneRIFs
Total geneRIFs Matched*	248,371	100%
Exact Matches Found*	173,586	70%
Total Found in Last Sentence of AB	110,333	44%
Total Found in Middle Sentences of AB	56,641	23%
Total Found in Title	39,732	16%
Total Found in Penultimate Sentence of AB	37,300	15%
Total Found in First Sentence of AB	4,365	2%

*Matches are based on string similarity. A geneRIF was considered matched to a sentence at a threshold of 60% similarity. Exact matches are based on a complete match between the geneRIF and a sentence or part of a sentence.

Gene and Disease Names

We suspected that the presence of disease and gene names in a sentence would be predictive, because our casual observation was that sentences summarizing novel clinical findings often

contained both. We used MetaMap to identify disease names in the sentence, and allowed the classifiers to attempt to construct rules taking them into account. Analysis was performed using both the actual names present in the text, and also with the actual text replaced by the generic representative strings “xxxgenexxx” or “xxxdiseasexxx.” The latter was attempted to see if generalized mentions of genes and diseases created more effective rules for the classifier than the specific names.

Discourse Classes and Claims Classes

Gobeill et al achieved some success using Discourse Classes to identify geneRIFs [27]. This is expected, because non-referential geneRIFs must be based on novel findings, and most novel findings will be located in the title, results, and conclusions. Gobeill et al trained their Discourse Classifier using existing categories in structured abstracts, and tested its performance on structured abstracts with the labels removed. They suggested that in addition to Discourse Classes, that tags distinguishing putative from established claims might also be helpful. Because we used a hand-tagged dataset of both structured and unstructured abstracts for training and testing, we were able to incorporate both Discourse Classes and Claims Classes into our analysis.

Classifiers

A set of open-source standard classifiers from Weka [29] were trained and tested on our data, including Naïve Bayes, Support Vector Machine, and AdaBoost (for binary classification only). From the complete set, 1,987 sentences were used for training, and 999 were used for testing.

Results and Analysis

Citation Filter

We have found it problematic to evaluate the performance of the filter. The filter we are currently employing for the purpose of the project is different in important ways from the filter that would eventually be implemented in production. For example, rules like filtering out all non-human species mentions are implemented just for the development of this prototype. The current filter is also far less sophisticated and sensitive than what we would expect to implement in a production environment. A thorough evaluation of the current filter would be premature and time-consuming, so we have presented only some preliminary results here.

Additionally, there is difficulty with what results to include in such an analysis. Strictly speaking, we are “filtering” out the majority of MEDLINE, but this insight does not give us meaningful numbers on our performance. Similarly, evaluating the performance of the filter where it is based on hard-and-fast rules like the acceptable publication types does not offer meaningful insight. Therefore, the analysis that we choose to present here handles only those citations where the filter could reasonably be expected to make a “mistake.” The analyzed citations for this evaluation therefore include just the citations included by the filter in our

dataset, plus a proportional number of citations filtered out for criteria based on gene mentions (approximately an extra 20%).

Our ability to evaluate even the current filter is hampered by the necessity of a non-expert guessing what an indexer would deem appropriate for gene indexing. The coder marking citations as suitable or non-suitable for gene indexing is not an indexer. JW did graciously review the citation coding from the test set for our accuracy on this count. We disagreed on 12 of the 151 citations, putting our agreement at 79%. However, our disagreements did not significantly alter the ratio of suitable to unsuitable citations in the set, as shown in Table 4.

Table 4: Citations from the Test Set Evaluated for Gene Indexing by Two Coders

	Suitable	Unsuitable
JCS	100	51
JW	98	53

In addition to the test set, the original coder looked at an additional representative set of citations rejected by the filter for absence of gene mentions. Of these 31, four contained missed gene mentions and were suitable for gene indexing. Thus, based on this small analysis set of 182, the current filter has fairly low precision (around 65%) and high recall (around 96%).

We also used these results to estimate what percent of articles are correctly receiving gene indexing. Note that two citations were eliminated from this set because they contained non-human species mentions that the filter failed to detect. The filter is currently intended to filter these out, but only for the purposes of this project. Therefore, including these in this analysis would have misrepresented the efforts of the Index Section.

Table 5: Gene Indexing on a Sample of Filtered Abstracts

Citations	Instances	Received geneRIFs
Unsuitable for Gene Indexing	53	4
Suitable for Gene Indexing	96	60

Only 60 of the 96 suitable citations have actual geneRIFs associated with them. (See Table 5) This suggests that the remaining 36 were not correctly processed for gene indexing. Six of those 36 were from 2002, the first year that gene indexing was practiced, and that year is likely incomplete. However, the GIA might encourage comprehensive treatment by suggesting those other 30 that are being missed right now.

Four of the 53 unsuitable citations had geneRIFs associated with them. JW reviewed these and verified that three of those four should not have been gene indexed. The remaining article would not have been gene indexed based only on the abstract, but qualified when examined in full text. It is possible that better filtering could also prevent some undesirable entries from being created, though this is a very small problem in comparison.

However, reaping the benefits of more comprehensive gene indexing will depend on improving the filter specificity. Of the original 151 citations, 52 were ultimately unsuitable for gene indexing. If more than a third of the GIA suggestions are bad, indexers will probably ignore them. Worse, indexers might be slowed by stopping to evaluate so many bad suggestions.

Gene Mention Identification

The Gene Mention Identification module currently performs at 88% recall, 84% precision, and an 86% F_1-measure. Please refer to Table 6.

Based on data available from the literature, we estimate our current results to be similar to other high-performing gene identification resources, but we have not yet run direct comparisons of different tools on identical test sets. Since we plan to continue to develop this module, we will have additional opportunities to run such comparisons.

The current results represent significant improvement from the original incarnation of this module. The dictionary alone had mediocre performance on the original training set, with precision originally lower than 60%. Performance was improved through analysis and refinements discussed below.

Table 6: Identification Results

Mention Type	Instances
Correct Mentions	1,014
Missed Mentions (False Negatives)	143
Bad Mentions (False Positives)	198

Bad Mentions

We focused on reducing the bad mentions, since our precision was quite low. We reviewed the bad mentions in the training set, attempting to categorize them according to the origin of the errors. A few of the highlights of this analysis are below.

Identical Abbreviations

A large percentage of the errors stemmed from abbreviations for things like diseases and experimental techniques that were identical to abbreviations for gene names. Implementing MetaMap alleviated much of this problem. MetaMap contains an algorithm for identifying author defined acronyms or abbreviations where the short form is found within parentheses immediately following the long form. Using MetaMap, we are able to match most abbreviations in the abstracts to their long forms. We then expand any subsequent use of that abbreviation in the text, to replace it with the long form. This technique alone eliminated about 70% of the bad mentions during development with the training set. This represented a large gain in performance, despite introducing a few additional missed mentions. These came from cases where we had recognized the original abbreviation, but did not recognize the long form.

Common Words

We created a list of ordinary words that are also gene designations, like “great,” “simple,” and “sky.” The identification of these words triggers a case-sensitive check that most of the text is not subject to. Only if the word matches the case-specific use for the gene designation does it get marked as a mention. We do not use case-sensitive criteria for all gene designations indiscriminately, since minor author variations would cause us to miss too many genes. Although this manual-identification approach was effective, it is not likely scalable for use with additional species.

Domain Specific Rules

We devised a number of additional rules to handle specific mentions. For example, CGH is a gene, but is also a commonly used microarray. If the words “array” or “microarray” appear on either side of CGH, we do not consider it a gene mention. Similarly, we do not count “insulin” when it appears in “insulin-resistant” or “insulin-resistance.” Manual analysis yielded many such rules. Although we currently have these rules implemented, this approach is again not scalable to additional species, because of the extensive manual review it requires.

Disease Names

One of the most intransigent identification problems is caused by diseases that share a name with a gene. A few examples are *neurofibromatosis type 1*, *multiple sclerosis*, and *Rett syndrome*. These names are identical between the disease and the gene, and are therefore impossible to resolve with a simple dictionary match. Like the common words and domain-specific rules above, these disease names point to the need for a context-sensitive identification method. This could theoretically be provided with a statistical tool, albeit imperfectly.

Missed Mentions

We also reviewed the missed mentions, to identify additional possible improvements. The bulk of these error types are covered in the list below. We were able to address some of these with changes to our application and dictionary, and others may be addressed by future development or by the use of statistical methods.

- Lexical variants
    - Punctuation or spacing is different than that in the dictionary. Most of these errors have been corrected by adding additional variants to the dictionary.
- Symbolic variation
    - Different abbreviations or numerals are used, such as “a” for alpha or “B” for beta, or different Roman and Arabic numerals.
- Partial Tagging
    - Only part of the full gene name was recognized. Ex: “alcohol dehydrogenase” mentioned and official name “alcohol dehydrogenase IB” missed, even though it appears in the text. This problem has been mostly corrected using a length-based matching system, where longer matches are preferred.
- Name is too general
    - The name found is not specific enough to match an Entrez Gene record. For example, “hemoglobin” when there are various types such as beta, gamma B, and gamma A.
- Rearranged name components
    - A similar synonym appears in the Entrez Gene record, but it is rearranged compared to the one we missed. For example, the official name: hydroxysteroid (11-beta) dehydrogenase 1 was mentioned as: 11beta-hydroxysteroid dehydrogenase type 1.
- Synonym dissimilar

    - A synonym appears in Entrez Gene, but it was too dissimilar for us to recognize due to the addition of extra words or other changes.
- Name unlisted
    - The gene mention was totally unrecognizable in Entrez Gene.
- Sequences
    - Gene mentions were missed because multiple genes were listed in sequence. Ex: Cox 1-4, with only Cox 1 identified. Or Cox 1, 2, and 4, with only Cox 1 identified. We may be able to identify and expand these kinds of sequences during preprocessing, in a manner similar to the abbreviation matching and expansion we currently employ.

Gene Mention Normalization

We are able to identify and normalize gene mentions in our test set with 86% recall, 82% precision, and an 84% F_1-measure. To provide some sense of the performance of this module excluding the errors that originate with identification, we also looked at the performance for just the correct mentions—that is, just the mentions that refer to a real gene of some kind, excluding the bad mentions and the missed mentions. For correct mentions only, we are able to normalize 98% correctly. Our high overall performance reflects two things: the simple effectiveness of the dictionary approach in the majority of cases, and the low incidence of ambiguity in our test set: only 4% of the mentions are ambiguous, having more than one possible Entrez Gene ID.

Table 7: Disambiguation Results

Mention Type	Instances
Overall Correct Mentions	998
Overall Missed Mentions (False Negatives)	159
Overall Bad Mentions (False Positives/Bad Normalization)	214
Ambiguous Correct Mentions	41
Ambiguous Bad Mentions (False Positives/Bad Normalization)	16

For the ambiguous mentions, we correctly normalize only 61%. This low performance on ambiguous mentions reflects the relative lack of priority given to developing this module. Considering only a few days were spent developing our rather unsophisticated approach, and the many options available for future improvement, we are pleased with these results.

GeneRIF Extraction

Original Methodology

As reported above in Methods, we originally planned to use existing geneRIFs to train a classifier. However, this approach did not succeed, as the features we examined were not sufficiently predictive. Closer manual examination of the geneRIFs and their originating articles revealed some important insights.

Multiple sentences in an abstract could be geneRIFs

Contrary to our expectations, geneRIFs are not necessarily limited to a single sentence. A significant percentage of geneRIFs contained multiple full sentences or parts from multiple sentences combined by the indexer. Thus, multiple sentences in a citation can contain information worthy of geneRIFs. In fact, examining the abstracts, we realized that sometimes different sentences would paraphrase the same summarizing information, such as the title and conclusions. If one of those sentences is a geneRIF, surely the other could also have met the criteria for a good geneRIF. These insights led to the realization that our non-geneRIF corpus was likely filled with examples of sentences that could be geneRIFs or contribute to geneRIFs. This helps explain why it is difficult to find predictive features from the datasets we were using.

Types of geneRIFs have different features

Because we had strongly expected to find predictive n-grams, but found almost nothing, we examined the text of many geneRIFs manually to redefine our expectations for distinctive linguistic features geneRIFs might have. This examination revealed that geneRIFs have multiple latent types, based on the information they contain. These are also implied by the Index Section technical bulletins [1]-[4]. These types appear to have different features, including the n-grams they are likely to contain. Attempting to train a classifier to identify them as a single class (geneRIFs), probably forced the classifier to accept too many combinations of these features to make effective distinctions.

Selected geneRIFs may be suboptimal or inconsistent

During this analysis, we also noticed some geneRIFs that were not derived from the sentences we would have selected as those containing the most vital information. With many possible types of information possible in a geneRIF, there may be many sentences in an abstract that are in scope and unrelated to one another, detailing different findings on the same gene. We realized that the indexing guidelines for creating geneRIFs do not outline clear criteria for prioritizing the information that should be preferred for a geneRIF. This likely leads to significant variability between indexers, who may use different mental concepts of priorities to create their geneRIF annotations.

Current Methodology

The insights above made clear that all possible geneRIF sentences in an abstract needed to be identified for us to study them successfully, not just the one or two sentences ultimately selected. Thus, we decided to create a manually annotated dataset as described above in Methods, in order to train more sophisticated classifiers. Presented below are the most relevant high-performance data from the classification experiments. More complete results are available in the supplemental data for this project.

GeneRIF Classifier

The ultimate goal of this module is to distinguish between the geneRIF and non-geneRIF sentences. For this type of binary classification, our best results are an F_1 measure of .70, which

can be derived from Naïve Bayes (precision of .64 and recall of .76) or SVM (precision of .70 and recall of .70).

Table 8 shows the cumulative contributions of each feature currently being used by the classifier, while Table 9 gives more detail on the features tested. In Table 9, "Position" refers to the position of the sentence in the abstract. "Unigrams" and "Bigrams" are simply n-grams the classifier has identified as relevant. "Renamed Genes/Diseases" refers to sentences where genes and disease names have been replaced with generic identifier, as described in the methods. "Discourse" and "Claims" refers to using the results from the Discourse and Claims Classifiers as features. The final two rows of the table show the results using the manually annotated Discourse and Claims Classes, instead of those derived from the classifiers.

Table 8: Classifier Improvement from Additional Features

Features Used	**Recall**	**Precision**	F_1
Position in abstract	.33	.61	.42
+unigrams	.61	.59	.60
+bigrams	.65	.60	.62
+disease and gene names	.78	.64	**.70**

Table 9: Binary GeneRIF Classification

	Naïve Bayes			SVM		
Features	**Precision**	**Recall**	F_1	**Precision**	**Recall**	F_1
Position (P)	0.61	0.33	0.42	0.61	0.33	0.42
Unigrams (U)	0.59	0.61	0.60	0.65	0.52	0.58
P+U	0.60	0.65	0.62	0.65	0.61	0.62
Bigrams(B) + P+U	0.63	0.69	0.66	0.68	0.698	0.688
Renamed Genes/Diseases (D) +P+U	0.61	0.63	0.62	0.66	0.58	0.62
Renamed Genes/Diseases(D) +B+P+U	0.64	0.76	**0.70**	0.70	0.70	**0.70**
Discourse +P+U	0.62	0.65	0.64	0.67	0.61	0.64
Discourse +B+P+U+D	0.65	0.72	0.68	0.66	0.67	0.67
Claims +P+U	0.70	0.56	0.62	0.75	0.61	0.67
Claims +B+P+U+D	0.69	0.63	0.66	0.73	0.58	0.64
Using Hand-Annotated Discourse and Claims Data						
Discourse +B+P+U+D	0.74	0.87	0.80	0.78	0.92	0.85
Claims +B+P+U+D	0.82	0.78	0.80	0.90	0.80	0.84

As shown above, using classifier-generated Discourse Classes or Claims Classes as features to identify geneRIFs resulted in a drop in performance, due to the mediocre performance of both the discourse and Claims Classifiers. We also tested these features with the manually annotated classes instead of the classifier-derived classes. These results, shown in the final two rows of table 9, suggest that improving the classifiers for these classes could allow us to reap a benefit from both features.

Discourse Classifier and Claims Classifier

The Discourse and Claims Classifiers were created to support the geneRIF classifier, though they may prove useful for other research, as well.

Interestingly, the best results are obtained from both classifiers when individual classes were given binary classification, rather than all classes being simultaneously assigned in a multi-way classification. For discourse, the best results from the multi-way classification are available in table 10, while the best results from individual binary classifications are in tables 11-16. For claims, the best results from the multi-way classification are available in table 17, while the best results from individual binary classifications are in tables 18-20.

Table 10: Discourse All Classes Classification Using position + bigrams in Naïve Bayes.

	Precision	Recall	F_1
Results	0.64	0.63	0.63
Title	0.99	0.80	0.89
Background	0.77	0.70	0.73
Purpose	0.38	0.67	0.48
Methods	0.54	0.59	0.56
Conclusions	0.63	0.65	0.64

Table 11: Discourse "Results" Classification Using position + bigrams in SVM.

	Precision	Recall	F_1
Non-results	0.96	0.80	0.87
Results	0.67	0.91	**0.77**

Table 12: Discourse "Title" Classification Using position + unigrams in SVM

	Precision	Recall	F_1
Non-title	1.00	1.00	1.00
Title	1.00	1.00	1.00

Table 13: Discourse "Background" Classification using position + unigrams in AdaBoost

	Precision	Recall	F_1

Non-background	0.88	0.95	0.91
Background	0.82	0.62	**0.71**

Table 14: Discourse "Purpose" Classification Using position + bigrams in Naïve Bayes.

	Precision	Recall	F_1
Non-purpose	0.99	0.92	0.95
Purpose	0.31	0.76	**0.44**

Table 15: Discourse "Methods" Classification Using position + bigrams in Naïve Bayes

	Precision	Recall	F_1
Non-methods	0.91	0.85	0.88
Methods	0.58	0.70	**0.63**

Table 16: Discourse "Conclusion" Classification Using position + bigrams in Naïve Bayes.

	Precision	Recall	F_1
Non-Conclusion	0.92	0.90	0.91
Conclusion	0.58	0.65	**0.62**

Table 17: Claims All Classes Classification

	Precision	Recall	F_1
Established	0.57	0.81	**0.67**
Putative	0.90	0.58	**0.71**
Non-Claim	0.62	0.78	**0.69**

Table 18: Claims "Established" Classification

AdaBoostM1	Precision	Recall	F_1
Non-Established	0.89	0.94	0.91
Established	0.78	0.63	**0.70**

Table 19: Claims "Putative" Classification

	Precision	Recall	F_1
Non-putative	0.81	0.84	0.82
Putative	0.82	0.78	**0.80**

Table 20: Claims "Non-Claim" Classification Using position + bigrams in SVM.

	Precision	Recall	F_1
No Non-Claim	0.96	0.66	0.78
Non-Claim	0.51	0.92	**0.65**

Recommendations

Improve Current Application

Citation Filter

Improve Gene Mention Identification

Our primary strategy to improve the filter will actually be to continue to improve our ability to identify gene mentions accurately. Since some citations contain mentions that we miss, and some contain spurious mentions that we pick up incorrectly, improving gene mention identification will cut down on both false positives and false negatives.

Implement Additional Filter Criteria

However, we also hope to develop more sophisticated filtering based on some of the more subtle indexing criteria. Specifically, we think we may be able to filter out case studies with only one patient, and also articles about gene databases instead of actual genes, neither of which receive gene indexing under the guidelines of the Indexing section, and both of which seem to have fairly clear indicators in the text.

Test and Analyze an Expanded Test Set

Filter performance may vary dramatically when additional species and journals are introduced in the next year. The filter errors may be different than those that we currently observe in our limited subset. The filter performance will need additional testing and analysis to determine additional appropriate strategies for improvement in a more general MEDLINE setting.

Gene Mention Identification

Refine the Dictionary

Expanding our dictionary is an obvious strategy for improving our ability to find terms. This will improve recall. The current dictionary is built only from Entrez Gene entries and variants created through regular expressions. Additional resources for gene names must be identified and incorporated into the dictionary. Additional variants may also be created, based on our further analysis of missed mentions.

Incorporate Additional Existing Software

With additional time now allotted to this project, we plan to revisit the available identification tools that use statistical machine learning methods. These approaches could improve both

precision and recall. It is possible that such a tool will offer robustness that a rigid dictionary does not for unfamiliar names and variants, as well as help eliminate some of the bad mentions by recognizing their unlikely context. Our dictionary and related rules may improve the accuracy of a combined approach. Our annotated dataset could also be used to improve the training of an existing tool. Exploring ways to combine the information streams from a machine-learning tool and our dictionary matching program is a priority.

Use Parts of Speech

Identifying the part of speech that a word belongs to may clarify whether it is a gene mention or not. Verbs will never be genes, for example. It is likely that many additional types of words, and possibly even words appearing in certain kinds of phrases, will not be gene mentions. Since MetaMap runs as part of our program already, using its integrated Part Of Speech tagger would be relatively simple. This could improve precision.

Gene Mention Normalization

Refine the Dictionary

Because our normalization strategy rests heavily on our dictionary, expanding the dictionary as discussed above will also benefit this module.

Implement Gene Profiles

We are currently developing a method for comparing the similarity between the information in a candidate Entrez Gene record and the citation in question. We create a "gene profile" by extracting a collection of keywords from an Entrez Gene record. We can compare those keywords to a citation, and score the degree of matching.

Use Additional Verification Cues

Additional information in the abstract or metadata, or possibly even the article full text, could be used to disambiguate gene mentions. Cues that might be helpful and could be verified against the candidate Entrez Gene records include:

- gene location or chromosome number
- gene sequences
- number of exons or introns
- the position of amino acids in the wildtype (often mentioned in the context of variants or mutations)

Using any of these additional cues requires developing the means to identify them first, so this strategy will require significant time investment. However, it is likely that identifying these additional features will be useful for other types of information extraction, so the time investment is likely to pay off in expanded applicability for the program.

Try Fuzzy Matching

If we are able to successfully integrate a machine-learning based tool into our application, we will thereby generate some mentions that do not appear in our dictionary. We will also need to develop a method for making fuzzy matches to the dictionary entries when an exact match is not available.

GeneRIF Extraction

Test Additional Features

The classifiers may perform better with additional features. The density of terms from the Gene Ontology in sentences has been found to be predictive for geneRIFs, so we would like to try a similar analysis [22], [23]. Additionally, verb tenses may be predictive, especially if combined with sentence position or Claims Class. A variety of other linguistic patterns could also be investigated, to find dependencies that are predictive, like subject-object relationships with gene names.

Enhance Dataset Tagging

We plan to continue to enhance the tagging on the dataset. Although virtually any sentence with novel genetic information is marked as a potential geneRIF, some sentences are obviously better candidates than others. For example, geneRIFs tend to be sentences that summarize overall findings and make sense independent of surrounding context. We plan to develop some criteria for identifying high-quality geneRIFs or otherwise ranking geneRIF quality, and tagging the dataset with this additional information. That information could then be used to weight the training data, or to developing ranking for candidate sentences, depending on which approach works better.

Develop Simple Anaphora Resolution

Anaphora means referring to a concept by only using part of its name or an indirect expression such as a pronoun. Resolving some of this kind of abbreviation as it relates to biological concepts would probably help us recognize geneRIF sentences. Through the course of reading many geneRIF sentences for this project, two areas present themselves as obvious candidates for simple anaphora resolution: mutations referring to genes and patients referring to diseases. Resolving anaphora of these types would make obvious instances where a gene and disease are both mentioned in a sentence. It would also give us better sentences to suggest to the indexers, since the sentences would contain more complete ideas.

Below are several examples of sentences with anaphora resolution. The corrective anaphora expansion is given in brackets.

Examples of mutations referring to genes. This type of anaphora is likely to be resolveable.

> *The B1B1 homozygote [of cholesterol ester transfer protein] was associated with significantly lower HDL-C levels in females and non-smoking males.*

Significant differences in survival were detected among Rett syndrome cases grouped for the eight most frequent mutations [of MECP2].

Example of a general protein type referring to a specific protein. This type of anaphora is more difficult to resolve.

Our study shows that this factor [SOX10], in synergy with EGR2, strongly activates Cx32 expression in vitro by directly binding to its promoter.

Examples of patients referring to disease. This is another case that is relatively straightforward.

SCA 3 was identified in 31 (53.4%) patients [with ataxia] from 15 families.

Of the 239 patients [with acute myeloblastic leukemias], 30 (12.6%) showed MLL abnormalities under FISH analysis.

Haplotype analysis revealed that affected individuals [with autosomal dominant hearing loss] were heterozygous for one core SNP CAGTC haplotype, confirming location and autosomal dominant inheritance of the DFNA41 locus.

An example of anaphora for symptoms, instead of a disease. This type of anaphora is again more problematic.

Direct sequence analysis revealed a deletion of 108 bp of factor V in eight out of 20 individuals in this family [with normal factor V coagulant and anticoagulant properties].

Many additional type of anaphora exist, many of which are likely to be virtually impossible to resolve. We have not performed a full analysis of anaphora occurring in our dataset in order to design a detailed approach. However, our general strategy will be, as elsewhere, to target the easiest problems for the greatest gain, and then revisit the approach as needed for additional potential benefit.

Expand to Additional Species

This application must eventually address the whole of MEDLINE. The next step to expand this project is to develop the ability to identify and normalize genes across multiple species. This will entail the creation of several additional modules for the application, or incorporation of existing programs that perform similar functions. For example, at least two programs are available for download that already perform cross-species identification and normalization [17],[19]. If we can implement one of these, we may be able to accelerate the development of cross-species identification by building on existing research.

Develop a Species Identification Module

Species identification will be similar to any other NER, like gene mention identification. Some research on this has been conducted for the BioCreative 3 challenge, and additional research is very likely to exist, as well. The favored approach from preliminary reading appears to be dictionary based. A full literature review must be performed, and a dictionary of species names constructed. A suitable pre-annotated dataset for this is likely already available.

Develop Species-Specific Gene Dictionaries

Dictionaries of gene names will need to be developed for each species, similar to the one constructed for humans in this prototype. As before, Entrez Gene can be used as the primary resource, and additional synonyms can be harvested from other databases. Many model organisms have dedicated gene databases. These must be identified and incorporated.

Develop Species-to-Gene Assignment Module

In cases where multiple species are mentioned or implied in a paper, any gene mentions must be assigned to the appropriate species. Methods from BioCreative 3 for disambiguating species-to-gene assignments generally followed a proximity model, using various measurements of closeness of the species name to the gene mention to make the assignment [7]. Additional disambiguation features like chromosome location and sequences discussed above could also be used to enhance this module.

Design New Indexing Interface and Test Protocols

To put the system into production, the interface that gene indexers currently use in DCMS will need to be redesigned to allow them to evaluate and modify suggestions, to manually perform tasks when the GIA fails, and ideally to give instant feedback that can improve the functioning of the tool, for example by adding terms to the dictionary and marking bad mentions. The design process is likely to be conducted through focus grouping. After an interface is designed, indexers will need to run a trial with it to evaluate how the new design and features affect speed and to give feedback.

Certain gene indexing guidelines should be revisited at this point, as well. For example, the prohibition against using the title as a geneRIF is generally ignored by indexers, and may be misguided. Additionally, the arbitrary maximum geneRIF length of 255 characters is unnecessary and may waste expensive indexer time, by forcing them to shuffle and trim words simply to accommodate the limit. Finally, the indexing “rule of three” is another limit that may be rendered obsolete by the GIA. Indexers are instructed to ignore gene indexing for articles with more than three genes if they do not appear in the title. The intent of this rule is to prevent indexers from becoming mired in excessive linking for single articles. However, if linking is automated, the time spent making additional links may become negligible. In this case, the “rule of three” should be abandoned in favor of comprehensive gene indexing.

Experiment with Full Text

We suspect, based on conversations with indexers and other researchers, that as much as half of the information we will require on genes and species may be available only in the full text of an article. However, analyzing the full text may prove more difficult, due to the amount of noise a full text article is likely to have relative to an abstract. Additionally, gaining access to full text articles during the overnight pre-processing period may prove impossible. Publishers have so far been resistant to allowing the NLM this kind of access for use with the MTI, but the need for full text analysis is only likely to grow as the number of applications for information extraction expands. Although getting access and analysis of full text is problematic, it may prove vital. The magnitude of these problems and options for addressing them need to be addressed.

Release Resources for the Greater Research Community

Release Annotated Corpus for Other Researcher

The annotated corpus produced in this project is a valuable research product to make available to the research public. Even with only two coders, few hand annotated datasets of this size are available publically. It will be made publically available and updated as we develop additional tagging and possibly discover and correct mistakes.

Develop Modules into Standalone, User-Friendly NLP tools

Kabiljo *et al* have found that despite the availability of many supposedly turn-key named entity recognition (NER) tools, most such tools are in fact outside the ability of average biology researchers to implement [30]. If this is indeed the case, NLM is in a position to provide a set of user-friendly information-mining tools for which there exists a great demand and present void. As our prototype evolves into a full-fledged application for use within DCMS, we strongly recommend that it also be developed for public release into a set of standalone, user-friendly tools for NER, normalization, and possibly auto-summarization or relationship extraction.

Investigate Additional Areas for Automation

Investigate Additional Indexing Functions

Additional indexing functions might also be automated or partially automated with similar techniques to those we have used here. Furthermore, the modules developed here might also be used to enhance the MTI, improving the selection of genes as subject headings, for example. Similar entity recognition techniques might be applied to chemicals and chemical families that receive chemical flagging during indexing, also improving the functions of the MTI. Furthermore, it may be possible to automatically select the articles from journals that receive selective indexing, and a brief review of the relevant guidelines does suggest that we could effectively eliminate many irrelevant articles before any human time is devoted to examining them. A full review of the indexing workflow might reveal additional targets for automation.

Investigate Semantic Relationship Extraction

A current focus of biomedical data mining is the extraction of semantic relationships from literature, including previous and ongoing research at both Lister Hill Center and NCBI. These relationships include gene-gene, gene-protein, and protein-protein interactions. Additional biochemicals such as flavonoids and cholesterol are also of interest, as are gene-disease or gene-phenotype relationships. The desire is to distill complex sentences containing information about any of these named entities into simple triplet relationships of the basic structure: entity-relationship-entity. Such relationships can be represented computationally, and so can be sorted, filtered, queries, searched, combined, represented and otherwise manipulated in a myriad of ways that free text does not accommodate. There would be obvious benefits to capturing information in this computable form, if possible, instead of the free text of the current geneRIFs. Our efforts to identify sentences with novel genetic information represents an excellent starting place from which to better extract genetic relationships.

Efforts are ongoing to build so-called "interactomes," grand databases of such relationships through which metabolic pathways and links between chemicals and diseases may be revealed and represented. However, the scope of such projects is enormous, and no existing resources approach comprehensiveness. Due to the labor-intensive nature of manual relationship extraction from literature, the landscape of interactomes is currently a patchwork of free and subscription resources covering various organisms to various degrees for only certain types of relationships. NLM should carefully consider the role it wishes to have in contributing to, curating, and hosting interactome resources in the future.

Encourage the Adoption of Reporting Standards

An unavoidable conclusion of this type of research is that machine processing of biomedical text would be made substantially easier if the output of authors could be controlled in key ways. If authors could be compelled to submit metadata with their articles that named relevant entities such as genes and proteins with standard identifiers, the ambiguity that is a source of so much struggle here could be eliminated. This approach would not require altering the paper itself, but rather would provide a key to analysis of the named entities as they appear in the paper. Other named entities that could be coded this way include diseases, drugs, species and strains, cell lines, and biologically active compounds such as cholesterol. Although this would obviously leave open the problem of past text, the most current research could be processed with much greater efficiency and accuracy.

As demand grows for detailed curation of experimental efforts, including details on how data has been acquired, well-designed reporting standards will be essential to facilitate searching and mining of data. For example, the BioGrid interaction database contains interactions among genes and proteins [31]. However, each interaction extracted from the literature is additionally annotated with the methods by which the original data was gathered. A number of more

extensive reporting standards for specific types of experiments are currently being developed by discipline-specific committees. Many of these efforts are loosely organized under an umbrella effort known as Minimum Information for Biological and Biomedical Investigations (MIBBI), but progress advancing them is patchy and slow [32].

Although a detailed examination of the needs and options for data standardization in biomedical research is outside the scope of this report, we do advise that the NLM should consider its role in facilitating standards development for data reporting as a means to organize and disseminate biomedical information.

References

[1] Janice Ward. Technical Memorandum 448: Further Notes on Gene Indexing. October 25, 2002

[2] Janice Ward. Technical Memorandum 466: Gene Indexing with Entrez Gene. November 5, 2004

[3] Janice Ward. Technical Memorandum 469: Additional Criteria for Gene Indexing. Computational Analysis and Specific Gene Databases). July 22, 2005

[4] Janice Ward. Technical Memorandum 482: Gene Indexing Selections. July 23, 2007

[5] L. Smith et al., "Overview of BioCreative II gene mention recognition," *Genome Biology*, vol. 9 Suppl 2, p. S2, 2008.

[6] A. Morgan et al., "Overview of BioCreative II gene normalization," *Genome Biology*, vol. 9 Suppl 2, p. S3, 2008.

[7] Z. Lu et al., "The gene normalization task in BioCreative III," *BMC Bioinformatics*, vol. 12, no. 8, p. S9, 2011.

[8] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, p. 3191, 2005.

[9] C. J. Kuo et al., "Rich feature set, unification of bidirectional parsing and dictionary filtering for high F-score gene mention tagging," in *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, 2007, pp. 105–107.

[10] R. K. Ando, "BioCreative II gene mention tagging system at IBM Watson," in *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, 2007, pp. 101–103.

[11] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 652-663, 2008.

[12] R. McDonald and F. Pereira, "Identifying gene and protein mentions in text using conditional random fields," *BMC Bioinformatics*, vol. 6, no. 1, p. S6, 2005.

[13] M. Torii, Z. Hu, C. H. Wu, and H. Liu, "BioTagger-GM: a gene/protein name recognition system," *Journal of the American Medical Informatics Association: JAMIA*, vol. 16, no. 2, pp. 247-255, Apr. 2009.

[14] Alias-I. *LingPipe 4.1.0.* Available: http://alias-i.com/lingpipe,
[15] R. Tsai, C.-L. Sung, H.-J. Dai, H.-C. Hung, T.-Y. Sung, and W.-L. Hsu, "NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition," *BMC Bioinformatics*, vol. 7, no. 5, p. S11, 2006.
[16] P. P. Kuksa and Y. Qi, "Semi-Supervised Bio-Named Entity Recognition with Word-Codebook Learning."
[17] J. Wermter, K. Tomanek, and U. Hahn, "High-performance gene name normalization with GENO," *Bioinformatics*, vol. 25, no. 6, pp. 815-821, Feb. 2009.
[18] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez, "Inter-species normalization of gene mentions with GNAT," *Bioinformatics*, vol. 24, no. 16, p. i126-i132, Aug. 2008.
[19] M. L. Neves, J.-M. Carazo, and A. Pascual-Montano, "Moara: a Java library for extracting and normalizing gene and protein mentions," *BMC Bioinformatics*, vol. 11, p. 157, 2010.
[20] Y. Kano et al., "U-Compare: share and compare text mining tools with UIMA," *Bioinformatics (Oxford, England)*, vol. 25, no. 15, pp. 1997-1998, Aug. 2009.
[21] NCBI. August 2011. *GENE_INFO Directory.* Available: ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz
[22] NCBI. August 2011. *Gene_history file.* Available: ftp://ftp.ncbi.nih.gov/gene/DATA/gene_history
[23] NCBI. August 2011. *Mim2gene file.* Available: ftp://ftp.ncbi.nih.gov/gene/DATA/mim2gene
[24] Aronson AR, Lang FM. *An Overview of MetaMap: Historical Perspective and Recent Advances*. J Am Med Inform Assoc. 2010 May 1;17(3):229-36
[25] Structured abstracts reference
[26] Z. Lu, K. B. Cohen, and L. Hunter, "Finding GeneRIFs via gene ontology annotations," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 52-63, 2006.
[27] J. Gobeill, I. Tbahriti, F. Ehrler, A. Mottaz, A.-L. Veuthey, and P. Ruch, "Gene Ontology density estimation and discourse analysis for automatic GeneRiF extraction," *BMC Bioinformatics*, vol. 9 Suppl 3, p. S9, 2008.
[28] Z. Lu, K. B. Cohen, and L. Hunter, "GeneRIF quality assurance as summary revision," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 269-280, 2007.
[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
[30] R. Kabiljo, A. B. Clegg, and A. J. Shepherd, "A realistic assessment of methods for extracting gene/protein interactions from free text," BMC Bioinformatics, vol. 10, no. 1, p. 233, 2009.

[31] C. Stark et al., “The BioGRID Interaction Database: 2011 update,” *Nucleic Acids Research*, vol. 39, no. Database issue, pp. D698-704, Jan. 2011.
[32] F. Taylor et al., “Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project,” *Nature biotechnology*, vol. 26, no. 8, pp. 889–896, 2008.

Appendix: Guidelines for Class Assignment

Discourse Blocks

- Title: the title of the citation
- Background: Background information defines current challenges or questions in the field, reiterates established findings, or describes previous earlier work related to the current experiments.
- Objective: Objective describes the hypothesis or research challenge the experiment was intended to address.
- Methods: Methods describe the research techniques used to collect the research data. This includes the descriptions of a sample population. In a case report, this includes just the initial summarization of the presentation of the case “a family of four presenting with etc…” Methods include methods of analysis that are not “wet lab” based, such as computer modeling, statistical analysis, and meta-analysis.
- Results: Results summarize the data reported.
- Conclusions: Conclusions report the status of the research hypotheses, and discuss the additional significance or implications of the findings. Pay attention to the difference between reporting the actual data (Results) and drawing conclusions from the data (Conclusions). For example: “Mutations were strongly *correlated to* disease phenotypes.” (Results) versus “This mutation *causes* heart attacks.” (Conclusion) Conclusions should also include non-putative discussion of the significance of the work (“This research adds to our knowledge of heart disease”) and implications for further research (“Other mutations may still be implicated in heart attacks”).
- Not all types of discourse need appear in a given abstract.
- All sentences should be tagged.
- Categories are not mutually exclusive, and multiple categories may be needed to describe a sentence. For example, it is not uncommon to see a sentences that contains both methods and results, as in, “Sequencing of the exons revealed 17 new mutations.” Another common combination is objective and methods, as in, “In order to investigate the connection between COX1 and mitral valve disease, we sequenced chromosome 7 in 30 patients with prolapsed valve disorder.”

Scientific Claims

- Established: established claims are statements of relevant scientific information that was known or assumed before the publication of the paper. These are generally provided as background introductory information, or offered as contrast to the new findings in the paper, and include negative claims (non-association of risk factors, for example). Factual sentences that do not make scientific claims, like those provided for historical character or narrative continuity, should be marked as non-claims.
- Putative: putative claims are factual statements either reporting or interpreting the data collected or analyzed in the paper, including negative findings. Claims include those that are qualified or "cautious" findings like "may indicate", "suggest", or "predict." Putative claims also include those made in the title that may not be full sentences, but indicate a new scientific finding, such as "Association between NC1 polymorphisms and schizophrenia."
- Non-claims: Trivial facts that are tangential to the scientific claims of the paper, or non-factual statements. Speculative consequences of further research and references to the significance of the research should be considered non-claims. Recommendations are non-claims.
- All sentences should be tagged.
- Categories are mutually exclusive. Sentences to which multiple categories could be applied should be given their category in this order of preference: Putative, Established, Non-Claim.

GeneRIF types

- Non-geneRIF: Any sentence that does not contain GeneRIF information. A geneRIF must contain new scientific information about a gene or gene product such as protein or RNA. geneRIFs may not be the results of gene therapy trials.
- Referential: A sentence that summarizes the nature of the paper or findings, without giving actual findings. These include sentences summarizing the scope of a review or population study. Titles and methods often contain referential geneRIFs. Referential geneRIFs should be tagged even when "better" geneRIF candidates exist elsewhere in the abstract.
- Isolation: A sentence that indicates the first isolation and description of a new gene or gene product(s). This does not include the isolation of new mutations, SNPs, or other polymorphisms in known genes. New mutations should be described as function if they are associated with a particular disease or other phenotype and as structure if they are given in the context of specific regions or structural effect on a protein or RNA product.
- Structure: A sentence describing the structure of a gene or gene product. This includes sequence, folding and other conformational changes, composition and formation of multiunit products, and modifications or variations to a primary structure. Identification of new attributes in the gene or protein itself, such as exons, promoter regions, binding sites, repeats, and so on are considered structure geneRIFs. Identification of variant

proteins should be considered structure. Not all descriptions of mutations are structural geneRIFs.
    - Mutations resulting implicitly in a major structural change, such as a large deletion or a frameshift, should be marked as structural.
    - Smaller mutations are also structural geneRIFs if they are described in terms of the region of a gene or protein where the mutation occurs or is near, like “the hydrophilic binding region” or “the promoter region.”
    - Single nucleotide or amino acid substitutions should not be described as structural geneRIFs UNLESS an explicit structural change resulting from them is given (folding error, etc.) or a region is described (as above). Such single substitutions that are linked to a conformational change are structure geneRIFs. Substitutions that describe a functional result without describing any other structural result are function geneRIFs. If the sentence describes a substitution occurring in a particular region with a known function, and thereby effecting a functional change, they should be marked as both structure and function. Single substitutions without additional information given are non-geneRIFs.
- Function: A sentence describing the role of a gene or gene product. This can include normal or pathological functions, interactions with other molecules and genes outside of complexing, and association to disease or other phenotypes. Disease related mutations are functions. The prevalence of mutations or variants in a disease should be labeled as functions, whereas the amount of expression in a disease should be labeled as expression.
- Expression: A sentence describing differential expression of a gene or gene product, including colocalization with other molecules, expression profiles in tissues or organs, expression and occurrence in specific populations, differential expression like up- or down-regulation or over-transmission in response to processes like heat shock, disease, and embryonic development.
- Other: Anything that is a geneRIF candidate (carries significant new information about a gene or gene product), but doesn’t fall into any previous categories. This category should not be used unless absolutely unavoidable.
- All sentences should be tagged.
- Categories are not mutually exclusive, and multiple categories should be assigned to sentences that contain multiple kinds of geneRIF information. For example, it is especially common for structural information to be accompanied by function, as in, “This mutation leads to an improperly folded protein that is enzymatically inactive.”