Human Gene Name Normalization using Text Matching with Automatically Extracted Synonym Dictionaries

Haw-ren Fang
Department of Computer Science, University of Maryland
College Park, MD 20742, USA
hrfang@cs.umd.edu

Kevin Murphy and **Yang Jin** and **Jessica S. Kim** and **Peter S. White***
Division of Oncology, Children's Hospital of Philadelphia
Philadelphia, PA 19104, USA
{murphy,jin,kim,white}@genome.chop.edu

Abstract

The identification of genes in biomedical text typically consists of two stages: identifying gene mentions and normalization of gene names. We have created an automated process that takes the output of named entity recognition (NER) systems designed to identify genes and normalizes them to standard referents. The system identifies human gene synonyms from online databases to generate an extensive synonym lexicon. The lexicon is then compared to a list of candidate gene mentions using various string transformations that can be applied and chained in a flexible order, followed by exact string matching or approximate string matching.

Using a gold standard of MEDLINE abstracts manually tagged and normalized for mentions of human genes, a combined tagging and normalization system achieved 0.669 F-measure (0.718 precision and 0.626 recall) at the mention level, and 0.901 F-measure (0.957 precision and 0.857 recall) at the document level for documents used for tagger training.

1 Introduction

Gene and protein name identification and recognition in biomedical text are challenging problems. A recent competition, BioCreAtIvE, highlighted the two tasks inherent in gene recognition: identifying gene mentions in text (task 1A) (Yeh et al., 2005) and normalizing an identified gene list (task 1B) (Hirschman et al., 2005). This competition resulted in many novel and useful approaches, but the results clearly identified that more important work is necessary, especially for normalization, the subject of the current work.

Compared with gene NER, gene normalization is syntactically easier because identification of the textual boundaries of each mention is not required. However, gene normalization poses significant semantic challenges, as it requires detection of the actual gene intended, along with reporting of the gene in a standardized form (Crim et al., 2005). Several approaches have been proposed for gene normalization, including classification techniques (Crim et al., 2005; McDonald et al., 2004), rule-based systems (Hanisch et al., 2005), text matching with dictionaries (Cohen, 2005), and combinations of these approaches. Integrated systems for gene identification typically have three stages: identifying candidate mentions in text, identifying the semantic intent of each mention, and normalizing mentions by associating each mention with a unique gene identifier (Morgan et al., 2004). In our current work, we focus upon normalization, which is currently underexplored for human gene names. Our objective is to create systems for automatically identifying human gene mentions with high accuracy that can be used for practical tasks in biomedical literature retrieval and extraction. Our current approach relies on a manually created and tuned set of rules.

* To whom correspondence should be addressed.

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 41–48,
New York City, June 2006.

2 Automatically Extracted Synonym Dictionaries

Even when restricted to human genes, biomedical researchers mention genes in a highly variable manner, with a minimum of adherence to the gene naming standard provided by the Human Gene Nomenclature Committee (HGNC). In addition, frequent variations in spelling and punctuation generate additional non-standard forms. Extracting gene synonyms automatically from online databases has several benefits (Cohen, 2005). First, online databases contain highly accurate annotations from expert curators, and thus serve as excellent information sources. Second, refreshing of specialized lexicons from online sources provides a means to obtain new information automatically and with no human intervention. We thus sought a way to rapidly collect as many human gene identifiers as possible. All the statistics used in this section are from online database holdings last extracted on February 20, 2006.

2.1 Building the Initial Dictionaries

Nineteen online websites and databases were initially surveyed to identify a set of resources that collectively contain a large proportion of all known human gene identifiers. After examination of the 19 resources with a limited but representative set of gene names, we determined that only four databases together contained all identifiers (excluding resource-specific identifiers used for internal tracking purposes) used by the 19 resources. We then built an automated retrieval agent to extract gene synonyms from these four online databases: The HGNC Genew database, Entrez Gene, Swiss-Prot, and Stanford SOURCE. The results were collected into a single dictionary. Each entry in the dictionary consists of a gene identifier and a corresponding official HGNC symbol. For data from HGNC, withdrawn entries were excluded. Retrieving gene synonyms from SOURCE required a list of gene identifiers to query SOURCE, which was compiled by the retrieval agent from the other sources (i.e., HGNC, Entrez Gene and Swiss-Prot). In total, there were 333,297 entries in the combined dictionary.

2.2 Rule-Based Filter for Purification

Examination of the initial dictionary showed that some entries did not fit our definition of a gene identifier, usually because they were peripheral (e.g., a GenBank sequence identifier) or were describing a gene class (e.g., an Enzyme Commission identifier or a term such as “tyrosine kinase”). A rule-based filter was imposed to prune these uninformative synonyms. The rules include removing identifiers under these conditions:

1. Follows the form of a GenBank or EC accession ID (e.g., 1-2 letters followed by 5-6 digits).
2. Contains at most 2 characters and 1 letter but not an official HGNC symbol (e.g., P1).
3. Matches a description in the OMIM morbid list[1] (e.g., Tangier disease).
4. Is a gene EC number.[2]
5. Ends with ‘, family ?’, where ? is a capital letter or a digit.
6. Follows the form of a DNA clone (e.g., 1-4 digits followed by a single letter, followed by 1-2 digits).
7. Starts with ‘similar to’ (e.g., similar to zinc finger protein 533).

Our filter pruned 9,384 entries (2.82%).

2.3 Internal Update Across the Dictionaries

We used HGNC-designated human gene symbols as the unique identifiers. However, we found that certain gene symbols listed as “official” in the non-HGNC sources were not always current, and that other assigned symbols were not officially designated as such by HGNC. To remedy these issues, we treated HGNC as the most reliable source and Entrez Gene as the next most reliable, and then updated our dictionary as follows:

[1] ftp://ftp.ncbi.nih.gov/repository/OMIM/morbidmap

[2] EC numbers are removed because they often represent gene classes rather than specifi c instances.

- In the initial dictionary, some synonyms are associated with symbols that were later withdrawn by HGNC. Our retrieval agent extracted a list of 5,048 withdrawn symbols from HGNC, and then replaced any outdated symbols in the dictionary with the official ones. Sixty withdrawn symbols were found to be ambiguous, but we found none of them appearing as symbols in our dictionary.
- If a symbol used by Swiss-Prot or SOURCE was not found as a symbol in HGNC or Entrez Gene, but was a non-ambiguous synonym in HGNC or Entrez Gene, then we replaced it by the corresponding symbol of the non-ambiguous synonym.

Among the 323,913 remaining entries, 801 entries (0.25%) had symbols updated. After removing duplicate entries (42.19%), 187,267 distinct symbol-synonym pairs representing 33,463 unique genes were present. All tasks addressed in this section were performed automatically by the retrieval agent.

3 Exact String Matching

We initially invoked several string transformations for gene normalization, including:

1. Normalization of case.
2. Replacement of hyphens with spaces.
3. Removal of punctuation.
4. Removal of parenthesized materials.
5. Removal of stop words[3].
6. Stemming, where the Porter stemmer was employed (Porter, 1980).
7. Removal of all spaces.

The first four transformations are derived from (Cohen et al., 2002). Not all the rules we experimented with demonstrated good results for human gene name normalization. For example, we found that stemming is inappropriate for this task. To amend potential boundary errors of tagged mentions, or to match the variants of the synonyms, four mention reductions (Cohen et al., 2002) were also applied to the mentions or synonyms:

1. Removal of the first character.
2. Removal of the first word.
3. Removal of the last character.
4. Removal of the last word.

To provide utility, a system was built to allow for transformations and reductions to be invoked flexibly, including chaining of rules in various sequences, grouping of rules for simultaneous invocation, and application of transformations to either or both the candidate mention input and the dictionary. For example, the mention "alpha2C-adrenergic receptor" in PMID 8967963 matches synonym "Alpha-2C adrenergic receptor" of gene ADRA2C after normalizing case, replacing hyphens by spaces, and removing spaces. Each rule can be built into an invoked sequence deemed by evaluation to be optimal for a given application domain.

A *normalization step* is defined here as the process of finding string matches after a sequence of chained transformations, with optional reductions of the mentions or synonyms. We call a normalization step *safe* if it generally makes only minor changes to mentions. On the contrary, a normalization step is called *aggressive* if it often makes substantial changes. However, a normalization step safe for long mentions may not be safe for short ones. Hence, our system was designed to allow a user to set optional parameters factoring the minimal mention length and/or the minimal normalized mention length required to invoke a match.

A *normalization system* consists of multiple normalization steps in sequence. Transformations are applied sequentially and a match searched for; if no match is identified for a particular step, the algorithm proceeds to the next transformation. The normalization steps and the optional conditions are well-encoded in our program, which allows for a flexible system specified by the sequences of the step codes. Our general principle is to design a normalization system that invokes safe normalization steps first, and then gradually moves to more aggressive

[3] ftp://ftp.cs.cornell.edu/pub/smart/English.stop

ones. As the process lengthens, the precision decreases while the recall increases. The balance between precision and recall desired for a particular application can be defined by the user.

Specifically, given string s, we use $\mathcal{T}(s)$ to denote the transformed string. All the 7 transformation rules listed at the beginning of this subsection are *idempotent*, since $\mathcal{T}(\mathcal{T}(s)) = \mathcal{T}(s)$. Two transformations, denoted by $\mathcal{T}_1$ and $\mathcal{T}_2$, are called *commutative*, if $\mathcal{T}_1(\mathcal{T}_2(s)) = \mathcal{T}_2(\mathcal{T}_1(s))$. The first four transformations listed form a set of commutative rules. Knowledge of these properties helps design a normalization system.

Recall that NER systems, such as those required for BioCreAtIvE task 1B, consist of two stages. For our applications of interest, the normalization input is generated by a gene tagger (McDonald and Pereira, 2005), followed by the normalization system described here as the second stage. In the second stage, more synonyms do not necessarily imply better performance, because less frequently used or less informative synonyms may result in ambiguous matches, where a match is called *ambiguous* if it associates a mention with multiple gene identifiers. For example, from the Swiss-Prot dictionary we know the gene mention 'MDR1' in PMID 8880878 is a synonym uniquely representing the ABCB1 gene. However, if we include synonyms from HGNC, it results in an ambiguous match because the TBC1D9 gene also uses the synonym 'MDR1'.

We investigated the rules separately, designed the initial normalization procedure, and tuned our system at the end. To evaluate the efficacy of our compiled dictionary and its sources, we determined the accuracy of our system with all transformations and reductions invoked sequentially, and without any efforts to optimize the sequence (see section 6 for evaluation details). The goal in this experiment was to evaluate the effectiveness of each vocabulary source alone and in combination. Our experimental results at the mention level are summarized in Table 1. The best two-staged system achieved a precision of 0.725 and recall of 0.704 with an F-measure of 0.714, by using only HGNC and Swiss-Prot entries.

As errors can be derived from the tagger or the normalization alone or in combination, we also assessed the performance of our normalization program alone by directly normalizing the mentions in the gold standard file used for evaluation (i.e., assuming the tagger is perfect). Our normalization system achieved 0.824 F-measure (0.958 precision and 0.723 recall) in this evaluation.

Table 1: Results of Gene Normalization Using Exact String Matching

Steps		Recall	Precision	F-measure
(1)	HGNC	0.762	0.511	0.611
(2)	Entrez Gene	0.686	0.559	0.616
(3)	Swiss-Prot	0.722	0.622	0.669
(4)	SOURCE	0.743	0.431	0.545
(1)+(2)		0.684	0.564	0.618
(1)+(3)		0.725	0.704	0.714
(2)+(3)		0.665	0.697	0.681
(1)+(2)+(3)		0.667	0.702	0.684
(1)+(2)+(3)+(4)		0.646	0.707	0.675

4 Approximate String Matching

Approximate string matching techniques have been well-developed for entity identification. Given two strings, a *distance metric* generates a score that reflects their similarity. Various string distance metrics have been developed based upon edit-distance, string tokenization, or a hybrid of the two approaches (Cohen et al., 2003). Given a gene mention, we consider the synonym(s) with the highest score to be a match if the score is higher than a defined threshold. Our program also allows optional string transformations and provides a user-defined parameter for determining the minimal mention length for approximate string matching. The decision on the method chosen may be affected by several factors, such as the application domain, features of the strings representing the entity class, and the particular data sets used. For gene NER, various scoring methods have been favored (Crim et al., 2005; Cohen et al., 2003; Wellner et al., 2005).

Approximate string matching is usually considered more aggressive than exact string matching with transformations; hence, we applied it as the last step of our normalization sequence. To assess the usefulness of approximate string matching, we began with our best dictionary subset in Subsection 3

(i.e., using HGNC and SwissProt), and applied approximate string matching as an additional normalization step.

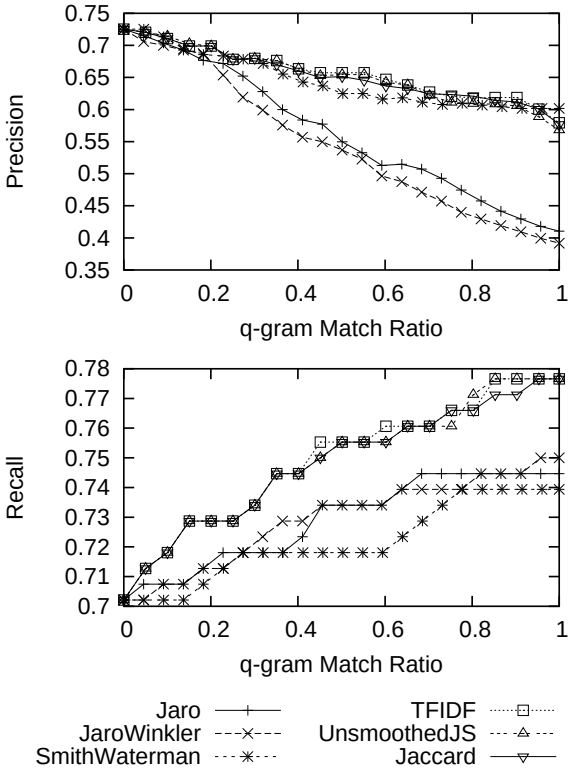

Figure 1: Performance of Approximate String Matching for Gene Normalization.

We selected six existing distance metrics that appeared to be useful for human gene normalization: Jaro, JaroWinkler, SmithWaterman, TFIDF, UnsmoothedJS, and Jaccard. Our experiment showed that TFIDF, UnsmoothedJS and Jaccard outperformed the others for human gene normalization in our system, as shown in Figure 1. By incorporating approximate string matching using either of these metrics into our system, overall performance was slightly improved to 0.718 F-measure (0.724 precision and 0.713 recall) when employing a high threshold (0.95). However, in most scenarios, approximate matching did not considerably improve recall and had a non-trivial detrimental effect upon precision.

5 Ambiguation Analysis

Gene identifier ambiguity is inherent in synonym dictionaries as well as being generated during normalization steps that transform mention strings.

5.1 Ambiguity in Synonym Dictionaries

If multiple gene identifiers share the same synonym, it results in ambiguity. Table 2 shows the level of ambiguity between and among the four sources of gene identifiers used by our dictionary. The rate of ambiguity ranges from 0.89% to 2.83%, which is a rate comparable with that of mouse (1.5%) and *Drosophila* (3.6%) identifiers (Hirschman et al., 2005).

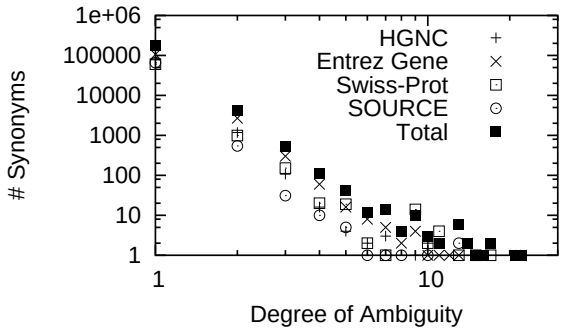

Figure 2: Distribution of ambiguous synonyms in the human gene dictionary.

Figure 2 is a log-log plot showing the distribution of ambiguous synonyms, where the degree is the number of gene identifiers that a synonym is associated with. Comparing Figure 2 with (Hirschman et al., 2005, Figure 3), we noted that on average, human gene synonyms are less ambiguous than those of the three model organisms.

Another type of ambiguity is caused by gene symbols or synonyms being common English words or other biological terms. Our dictionary contains 11 gene symbols identical to common stop words[4]: T, AS, DO, ET, IF, RD, TH, ASK, ITS, SHE and WAS.

5.2 Ambiguous Matches in Gene Normalization

We call a match *ambiguous* if it associates a mention with multiple gene identifiers. Although the

[4] ftp://ftp.cs.cornell.edu/pub/smart/English.stop

Table 2: Statistics for Dictionary Sources

Dictionary	# Symbols	# Synonyms	Ratio	Max. # of Synonyms per Gene	# with One Definition	Ambiguity Rate
HGNC	22,838	78,706	3.446	10	77,389	1.67%
Entrez Gene	33,007	109,127	3.306	22	106,034	2.83%
Swiss-Prot	12,470	61,743	4.951	17	60,536	1.95%
SOURCE	17,130	66,682	3.893	13	66,086	0.89%
Total	33,469	181,061	5.410	22	176,157	2.71%

normalization procedure may create ambiguity, if a mention matches multiple synonyms, it may not be strictly ambiguous. For example, the gene mention "M creatine kinase" in PMID 1690725 matches the synonyms "Creatine kinase M-type" and "Creatine kinase, M chain" in our dictionary using the TFIDF scoring method (with score 0.866). In this case, both synonyms are associated with the CKM gene, so the match is not ambiguous. However, even if a mention matches only one synonym, it can be ambiguous, because the synonym is possibly ambiguous.

Figure 3 shows the result of an experiment conducted upon 200,000 MEDLINE abstracts, where the degree of ambiguity is the number of gene identifiers that a mention is associated with. The maximum, average, and standard deviation of the ambiguity degrees are 20, 1.129 and 0.550, respectively. The overall ambiguity rate of all matched mentions was 8.16%, and the rate of ambiguity is less than 10% at each step. Successful disambiguation can increase the true positive match rate and therefore improve performance but is beyond the scope of the current investigation.

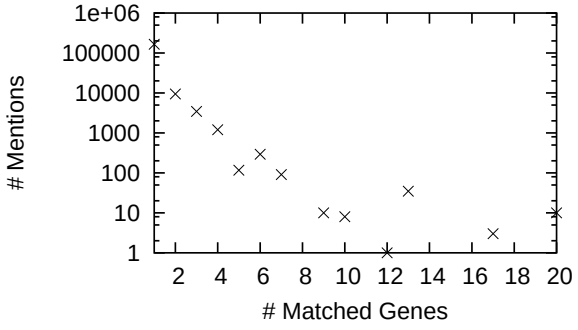

Figure 3: Distribution of Ambiguous Genes in 200,000 MEDLINE Abstracts.

6 Application and Evaluation of an Optimized Normalizer

Finally, we were interested in determining the effectiveness of an optimized system based upon the gene normalization system described above, and also coupled with a state-of-the-art gene tagger. To determine the optimal results of such a system, we created a corpus of 100 MEDLINE abstracts that together contained 1,094 gene mentions for 170 unique genes (also used in the evaluations above). These documents were a subset of those used to train the tagger, and thus measure optimal, rather than typical MEDLINE, performance (data for a generalized evaluation is forthcoming). This corpus was manually annotated to identify human genes, according to a precise definition of gene mentions that an NER gene system would be reasonably expected to tag and normalize correctly. Briefly, the definition included only human genes, excluded multi-protein complexes and antibodies, excluded chained mentions of genes (e.g., "HDAC1- and -2 genes"), and excluded gene classes that were not normalizable to a specific symbol (e.g., tyrosine kinase). Documents were dual-pass annotated in full and then adjudicated by a 3rd expert. Adjudication revealed a very high level of agreement between annotators.

To optimize the rule set for human gene normalization, we evaluated up to 200 cases randomly chosen from all MEDLINE files for each rule, where invocation of that specific rule alone resulted in a match. Most of the transformations worked perfectly or very well. Stemming and removal of the first or last word or character each demonstrated poor performance, as genes and gene classes were often incorrectly converted to other gene instances (e.g., "CAP" and "CAPS" are distinct genes). Re-

moval of stop words generated a high rate of false positives. Rules were ranked according to their precision when invoked separately. A high-performing sequence was "0 01 02 03 06 016 026 036", with 0 referring to case-insensitivity, 1 being replacement of hyphens with spaces, 2 being removal of punctuation, 3 being removal of parenthesized materials, and 6 being removal of spaces; grouped digits indicate simultaneous invocation of each specified rule in the group. Table 3 indicates the cumulative accuracy achieved at each step[5]. A formalized determination of an optimal sequence is in progress. Approximate matching did not considerably improve recall and had a non-trivial detrimental effect upon precision.

Table 3: Results of Gene Normalization after Each Step of Exact String Matching

Steps	Recall	Precision	F-measure
0	0.628	0.698	0.661
01	0.649	0.701	0.674
02	0.654	0.699	0.676
03	0.665	0.702	0.683
06	0.665	0.702	0.683
016	0.718	0.685	0.701
026	0.718	0.685	0.701
036	0.718	0.685	0.701

The normalization sequence "0 01 02 03 06 016 026 036" was then utilized for two separate evaluations. First, we used the actual textual mentions of each gene from the gold standard files as input into our optimized normalization sequence, in order to determine the accuracy of the normalization process alone. We also used a previously developed CRF gene tagger (McDonald and Pereira, 2005) to tag the gold standard files, and then used the tagger's output as input for our normalization sequence. This second evaluation determined the accuracy of a combined NER system for human gene identification.

Depending upon the application, evaluation can be determined more significant at either at the mention level (redundantly), where each individual mention is evaluated independently for accuracy, or as in the case of BioCreAtIvE task 1B, at the document level (non-redundantly), where all mentions within a document are considered to be equivalent. For pure information extraction tasks, mention level accuracy is a relevant performance indicator. However, for applications such as information extraction-based information retrieval (e.g., the identification of documents mentioning a specific gene), document-level accuracy is a relevant gauge of system performance.

For normalization alone, at the mention level our optimized normalization system achieved 0.882 precision, 0.704 recall, and 0.783 F-measure. At the document level, the normalization results were 1.000 precision, 0.994 recall, and 0.997 F-measure.

For the combined NER system, the performance was 0.718 precision, 0.626 recall, and 0.669 F-measure at the mention level. At the document level, the NER system results were 0.957 precision, 0.857 recall, and 0.901 F-measure. The lower accuracy of the combined system was due to the fact that both the tagger and the normalizer introduce error rates that are multiplicative in combination.

7 Conclusions and Future Work

In this article we present a gene normalization system that is intended for use in human gene NER, but that can also be readily adapted to other biomedical normalization tasks. When optimized for human gene normalization, our system achieved 0.783 F-measure at the mention level.

Choosing the proper normalization steps depends on several factors, such as (for genes) the organism of interest, the entity class, the accuracy of identifying gene mentions, and the reliability of the underlying dictionary. While the results of our normalizer compare favorably with previous efforts, much future work can be done to further improve the performance of our system, including:

1. Performance of identifying gene mentions. Only approximately 50 percent of gene mentions identified by our tagger were normalizable. While this is mostly due to the fact that the tagger identifies gene classes that cannot be normalized to a gene instance, a significant subset of gene instance mentions are not being normalized.

2. Reliability of the dictionary. Though we have

[5]The last two steps did not generate new matches using our gold standard file and therefore the scores were unchanged. These rule sets may improve performance in other cases.

investigated a sizable number of gene identifier sources, the four representative sources used for compiling our gene dictionary are incomplete and often not precise for individual terms. Some text mentions were not normalizable due the the incompleteness of our dictionary, which limited the recall.

3. Disambiguation. A small portion (typically 7%-10%) of the matches were ambiguous. Successful development of disambiguation tools can improve the performance.

4. Machine-learning. It is likely possible that optimized rules can be used as probabilistic features for a machine-learning-based version of our normalizer.

Gene normalization has several potential applications, such as for biomedical information extraction, database curation, and as a prerequisite for relation extraction. Providing a proper synonym dictionary, our normalization program is amenable to generalizing to other organisms, and has already proven successful in our group for other entity normalization tasks. An interesting future study would be to determine accuracy for BioCreAtIvE data once mouse, Drosophila, and yeast vocabularies are incorporated into our system.

Acknowledgment

This work was supported in part by NSF grant EIA-0205448, funds from the David Lawrence Altschuler Chair in Genomics and Computational Biology, and the Penn Genomics Institute. The authors acknowledge Shannon Davis and Jeremy Lautman for gene dictionary assessment, Steven Carroll for gene tagger implementation and results, Penn BioIE annotators for annotation of the gold standard, and Monica D'arcy and members of the Penn BioIE team for helpful comments.

References

K. B. Cohen, A. E. Dolbey, G. K. Acquaah-Mensah, and L. Hunter. 2002. Contrast and variability in gene names. In *ACL Workshop on Natural Language Processing in the Biomedical Domain*, pages 14–20.

W. W. Cohen, P. Ravikumar, and S. E. Fienberg. 2003. A comparison of string distance metrices for name-matching tasks. In *Proceedings of IIWeb Workshop*.

A. M. Cohen. 2005. Unsupervised gene/protein entity normalization using automatically extracted dictionaries. In *Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, Proceedings of the BioLINK2005 Workshop*, pages 17–24. MI: Association for Computational Linguistics, Detroit.

J. Crim, R. McDonald, and F. Pereira. 2005. Automatically annotating documents with normalized gene lists. *BMC Bioinformatics*, 6(Suppl 1)(S13).

D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck. 2005. Prominer: Rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6(Suppl 1)(S14).

L. Hirschman, M. Colosimo, A. Morgan, and A. Yeh. 2005. Overview of biocreative task 1b: Normalized gene lists. *BMC Bioinformatics*, 6(Suppl 1)(S11).

R. McDonald and F. Pereira. 2005. Identifying gene and protein mentions in text using conditional random fi elds. *BMC Bioinformatics*, 6(Suppl 1)(S6).

R. McDonald, R. S. Winters, M. Mandel, Y. Jin, P. S. White, and F. Pereira. 2004. An entity tagger for recognizing acquired genomic variations in cancer literature. *Journal of Bioinformatics*, 20(17):3249–3251.

A. A. Morgan, L. Hirschman, M. Colosimo, A. S. Yeh, and J. B. Colombe. 2004. Gene name identifi cation and normalization using a model organism database. *Journal of Biomedical Informatics*, 37(6):396–410.

M. F. Porter. 1980. An algorithm for suffi x stripping. *Program*, 14(3).

B Wellner, J. Castaño, and J. Pustejovsky. 2005. Adaptive string similarity metrics for biomedical reference resolution. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pages 9–16, Detroit. Association for Computational Linguistics.

A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman. 2005. Biocreative task 1a: Gene mention fi nding evaluation. *BMC Bioinformatics*, 6(Suppl 1)(S2).