Pacific Symposium on Biocomputing 12:281-291(2007)

EVALUATING THE AUTOMATIC MAPPING OF HUMAN GENE AND PROTEIN MENTIONS TO UNIQUE IDENTIFIERS

ALEXANDER A. MORGAN[1], BENJAMIN WELLNER[2], JEFFREY B. COLOMBE, ROBERT ARENS[3], MARC E. COLOSIMO, LYNETTE HIRSCHMAN

MITRE Corporation, 202 Burlington Road
Bedford, MA, 01730, USA
Email: alexmo@stanford.edu; lynette@mitre.org

We have developed a challenge task for the second BioCreAtIvE (Critical Assessment of Information Extraction in Biology) that requires participating systems to provide lists of the EntrezGene (formerly LocusLink) identifiers for all human genes and proteins mentioned in a MEDLINE abstract. We are distributing 281 annotated abstracts and another 5,000 noisily annotated abstracts along with a gene name lexicon to participants. We have performed a series of baseline experiments to better characterize this dataset and form a foundation for participant exploration.

1. Background

The first Critical Assessment of Information Extraction in Biology's (BioCreAtIvE) Task 1B involved linking mentions of model organism genes and proteins in MEDLINE abstracts to their corresponding identifiers in three different model organism databases (MGD, SGD, and FlyBase). The task is described in some detail in [1], along with descriptions of many different approaches to the task in the same journal issue. There has been quite a bit of past work associating text mentions of human genes and proteins with unique identifiers including the early work by Cohen et al. [2] and the AZURE system [3]. Very recently, Fang et al. [4] reported excellent results on a data set they created using one hundred MEDLINE abstracts. This widespread community interest in the issue and our experience with the first BioCreAtIvE motivated us to prepare another evaluation task for inclusion in the second BioCreAtIvE [5]. This task will require systems to link mentions of human genes and proteins with their corresponding EntrezGene (LocusLink) identifiers. We hope that researchers in this area can use this data set to compare techniques and

[1] Currently at Stanford Biomedical Informatics, Stanford University
[2] Also, the Department of Computer Science, Brandeis University
[3] Currently at the Department of Computer Science, University of Iowa

gauge performance gains. It can also be used to address issues in the general portability of normalization techniques and to investigate the relationships between co-mentioned genes and proteins.

2. Task Definition

The most important part of evaluating system performance is, of course, a very careful definition of the task. The original Task 1B required each system to provide a list of all the model organism database identifiers for the species-specific (mouse, fly or yeast) genes and gene products mentioned in a MEDLINE abstract. There are a number of possible uses for such a system, such as improved document retrieval for specific genes, data mining over gene/protein co-mentions, or direct support of relation extraction (e.g., protein-protein interaction) and/or attribute assignment (e.g., assignment of Gene Ontology annotations). The latter might be immediately useful to researchers attempting to analyze high throughput experiments, performing whole genome or comparative genomics analyses, or data-mining for relationship discovery, all of which require links to the unique identifiers.

Our initial investigations into a human gene/protein task suggested that UniProt identifiers [6] might be a good target to which we might normalize mentions of human proteins and their coding genes, and we hoped that this might bring the task into closer alignment with other efforts such as BioCreAtIvE I Task 2 [7] which required associating GO codes with human proteins identified through protein identifiers. UniProt provides a unified set of protein identifiers and represents a great leap forward for bioinformatics research, but it contains many redundancies: different fragments of the same polypeptide, polypeptide sequences derived from the same gene that differ in non-synonymous polymorphisms, and alternate transcripts from the same gene all may have separate entries and unique identifiers. We eventually settled on EntrezGene identifiers as unique target identifiers, despite incomplete mappings of UniProt to EntrezGene identifiers and what can be a complex many-to-many (e.g. alternate transcripts and gene duplications) relationship between genes and proteins. As described in [8], our annotation viewed genes and their products as equivalent because experience has found their typical usage interchangeable and/or indistinguishable. This is, of course, a simplification for purposes of evaluation; we recognize that this distinction is important in other cases.

A significant difference between the normalized gene list task (BioCreAtIvE Task 1B) and general entity normalization/grounding is that each gene list is associated with the abstract as a whole, whereas general entity grounding requires

the annotation of each mention in the text. The advantage of the "gene list" approach is that it avoids the issue of how to delimit the boundaries when annotating gene and protein mentions [9]. This becomes more of a problem in normalization when mentions are elided under various forms of conjunction. For example, it is difficult to identify the boundaries for the names of the different forms of PKC in "PKC isoforms alpha, delta, epsilon and zeta". Then there is the more difficult example of ellipsis: "AKR1C1-AKR1C4". Clearly AKR1C2 and AKR1C3 are being included in this mention, and functional information extracted about that group should include them. Fang et al. [4] excluded these cases from consideration, but we feel that these are important instances that need to be annotated and normalized. Equally difficult is the large gray area in gene and protein nomenclature between a description and a name and the related question of what should be tagged. The text "Among the various proteins which are induced when human cells are treated with interferon, a predominant protein of unknown function, with molecular mass 56 kDa, has been observed" mentions the protein also known as "interferon-induced protein 56", but the text describes the entity rather than using the listed name derived from this description. Our compromise was to keep the gene list task, but to provide a richer data set that associates at least one text string with each entry in the gene list, a significant addition over the first BioCreAtIvE Task 1B.

Polysemy in gene and protein names creates additional complexity, both within and between organisms [10]. Determination of the gene or protein being described may require the interpretation of the whole abstract – or several genes may be described with one "family name" term (see the **Discussion** section for further exploration of this issue). The particular species can be intentionally under-specified when the text is meant to refer to all the orthologues in relevant species, but in other cases, a name is meant to be highly species specific. For example: "Anoxia activates AMP-activated protein kinase (AMPK), resulting in the inhibition of biosynthetic pathways to conserve ATP. In anoxic rat hepatocytes or in hepatocytes treated with 5-aminoimidazole-4-carboxamide (AICA) riboside, AMPK was activated and protein synthesis was inhibited." The mention of the properties of AMPK in the first sentence is meant to be general and to include activity in humans, but the subsequent experimental evidence is, of course, in rats.

3. Corpus Construction

3.1. *Abstract Collection*

To identify a collection of abstracts with a high likelihood of mentions of human genes and proteins, we obtained the *gene_association.goa_human* file [11] on 10 October 2005. This provided us with 11,073 PubMed identifiers for journal articles likely to have mentions of human genes and proteins. We obtained abstracts for 10,730 of these. The file *gene2pubmed* obtained from NCBI [12] on 21 October 2005 was used, along with the GO annotations, to create the automatic/noisy annotations in the 5,000 abstracts set aside as a noisy training set as described in [8]. This is further described in the **Evaluation of Noisy Training Data** section. We selected our abstracts for hand annotation from the 5,730 remaining abstracts.

3.2. *Lexicon Creation*

The basic gene symbol and gene name information corresponding to each human EntrezGene identifier was taken from the *gene_info* file from NCBI [12]. This was merged with **name**, **gene** and **synonym** entries taken from UniProt [6]. Suffixes containing "_HUMAN", "1_HUMAN", "H_HUMAN", "protein", "precursor", "antigen" were stripped from the terms and added to the lexicon as separate terms in addition to the original term. HGNC [13] **symbol**, **name**, and **alias** entries were also added. We identified the phrases most repeated across identifiers and those that had numerous matches in the 5000 abstracts of noisy training data; we then used these to create a short (381 term) list to remove the most common terms that were unlikely to be gene or protein names but which had entered the lexicon as full synonyms. Examples of entries in this list are "recessive", "neural", "Zeta", "liver", "glycine", and "mediator". This list is available from the CVS archive [5]. This left us with a lexicon of 32,975 distinct EntrezGene identifiers linked to a total of 163,478 unique terms. The majority of identifiers have more than one term attached (average 5.5), although 8,385 had only one. For example, identifier 1001 has the following synonyms: "PCAD; CDHP; CDH3; cadherin 3, type 1, P-cadherin (placental); HJMD". It is important to note that many of these terms are unlikely to be used as mentions in abstracts for the given proteins and genes.

Many of the terms/synonyms were not unique among the identifiers, with the terms often being shared across a handful of identifiers (Table 1). Sometimes this reflects noise inherited from the source databases; the most egregious example is "hypothetical" which shows up as a name for 89 genes. Similarly, "human" (alone)

shows up 15 times, "g protein coupled receptor" 12 times, and "seven transmembrane helix receptor" 30 times. Each normalized (Section 4) phrase included as a synonym in this relatively noisy lexicon is linked to an average of 1.1 different unique identifiers, although 80% of phrases link to only one identifier. These synonyms average 16.5 characters in length if whitespace is removed.

Table 1. Lexicon statistics

Unique Gene ID's	32,975	Avg Term Length (Characters)	16.51
Unique Un-Normalized Terms	177,200	Avg Gene Identifiers per Term	1.12
Unique Normalized Terms	163,478	Avg Term Length (Words)	2.17
		Avg Terms per Identifier	5.55

3.3. *Annotation Tool and Annotation Process*

We developed a simple annotation tool using dynamic webpages with PHP and MySQL to support the creation of the normalized gene lists and extraction of the associated mention excerpts from the text. Annotators could annotate via their own web browsers. We could also make rapid changes to the interface as soon as they were requested without needing to update anything but the scripts on the server.

The simple annotation guidelines and the PHP scripts used for the annotation are available for download from the Sourceforge CVS archive [5]. The interface presented the plain text of the title and abstract to the annotators, along with suggested annotations (based on the automatic/noisy process). Using these resources, annotators had to provide the EntrezGene identifiers and supporting text for all mentions of human genes and proteins. All annotations then went through a review process to examine abstracts marked with comments and to merge the differences between annotators before inclusion in the gold standard set.

A total of 300 abstracts were annotated for the freely distributed training set, although 19 were removed for a variety of reasons, such as, having mentions which could not be normalized to EntrezGene, leaving 281 for distribution. The annotators found of an average of 2.27 different human genes mentioned per abstract. We have annotated another ~263 for use as an evaluation set. We plan to correct errors in these annotations based on pooling of the participants' submissions, as was done in the previous BioCreAtIvE [8]. The Sourceforge CVS archive will allow us to track corrections to these datasets [5].

3.4. *Inter-annotator Agreement*

We studied the agreement between different annotators on the same abstracts. The annotation was done by three annotators (two with PhD's in biological sciences, one with an MS; none are specialists in human biology, but all had previous experience in annotation). There was one annotator (primary) who did annotations for all abstracts. Our first pass of agreement studies was done on the first abstracts in the training set and was done mostly to check our annotation guidelines. Two annotators annotated the same 30 abstracts. There were 71 annotations (same EntrezGene identifiers for the abstract) in common and 7 differences (91% agreement). A second agreement experiment was performed with 26 new abstracts. There was only 87% agreement, but all disagreements were missed mentions or incorrect normalizations by the non-primary annotator. Unfortunately, these small sample sizes can only be suggestive of the overall level of agreement.

4. Characterizing the Data

In order to better characterize the properties of this dataset and task, we performed some baseline experiments, described below, to generate the list of EntrezGene identifiers for each abstract using the lexicon. We evaluated this using simple match against the gold standard annotations. For matching the terms from the lexicon, we ignored case and any punctuation or internal whitespace in the terms matched to the lexicon, but required match of start and end token boundaries as described in [14].

Table 2. Properties of the Data

Experiment	True Positive	False Positive	False Negative	Precision	Recall
Noisy Training Data Quality	348	49	292	0.877	0.544
Coverage of Lexicon	530	7941	110	0.063	0.828

4.1. *Evaluation of Noisy (Automatically Generated) Training Data*

We wanted to estimate the quality of the noisy training data and to evaluate our assumption that the document level annotations from the *gene2pubmed* file were indicative of a high likelihood of the mention of those genes in the abstract. To do this, we evaluated the gene lists derived from the *gene2pubmed* file (automatic/noisy data process) against those derived from human annotation (see Table 2). However, many genes may be mentioned in the abstract and paper but may not included in the *gene2pubmed* file causing our noisy training data to systematically underreport

genes mentioned, and we estimate from this result that only half of all genes mentioned are included in the automatic/noisy data annotations (recall 0.544).

4.2. *Evaluating the Coverage of the Lexicon*

We also evaluated the coverage of the lexicon by using it to do simple pattern matching. This mirrors some of our early experiments in developing normalized gene lists for *Drosophila melanogaster* [15]. Our goal was to estimate a recall ceiling on performance for systems requiring exact match to the lexicon. The recall of 0.828 clearly shows the limits of the simple lexicon (Table 2). This demonstrates the need to extend exact lexical match beyond such simple rules as ignoring case, punctuation and white space. In some cases, very small affixes (e.g. h-, -p, -like), either in the lexicon or the text, caused a failure to match. There were numerous cases of acronyms, often embedded in longer terms, which caused problems ("actinin-1" vs. "ACTN1" or "GlyR alpha 1" vs. "Glycine receptor alpha-1 chain precursor" or "GLRA1"). The various modifiers indicating subtypes were a serious problem, e.g. "collagen, type V, alpha 1"; modifiers such as "class II", "beta subtype", "type 1", and "mu 1" varied in orthography and placement, and the modifier "1" is often optional. Conjunctions such as "freac1-freac7" are particularly costly from an evaluation perspective since it can count as several false negatives at once. There was a considerable amount of name paraphrase (see **Discussion** section), involving word ordering and term substitutions or insertions and deletions. This arises because the long phrases in the lexicon are often more descriptive than nominal, although the associated acronyms can give some indication as to how a mention might actually occur in text. For example, the text contains "kappa opioid receptor", whereas the lexicon contains "KOR" and "opioid receptor, kappa 1"). Lan Aronson has investigated these issues in term variation while mapping concepts to text extensively [16]. Interestingly, self-embedded terms (e.g. "insulin-like growth factor-1 (IGF-I) receptor") seem to be a relatively rare problem at the level of the whole abstract. As expected, the precision based on lexical pattern matching (Table 2, row 2) was very low due to false positive matches of terms in the lexicon against common English terms, ambiguous acronyms, and so forth.

4.3. *Biological Context of Co-Mentioned Genes and Proteins*

As an example of how this dataset might be used outside of the evaluation, we looked at the biological relationships between genes and proteins which are mentioned together in the same abstracts. Our experience annotating the abstracts

indicated that genes or proteins are typically co-mentioned because of sequence homology and/or some functional relationship (e.g., interaction), although cell markers (e.g., CD4) may be mentioned in a variety of contexts. Many sophisticated techniques have arisen for comparing genes based on functional annotations and sequence, but for this initial analysis we intentionally used something naïve and simple. We computed two different similarity measurements for each pair of genes mentioned together in our dataset. For a sequence similarity computation, we used BioPython's pairwise2 function [17]:

```
pairwise2.align.globalxs(seq1,seq2,-1,-.1,penalize_end_gaps=0,score_only=1).
```

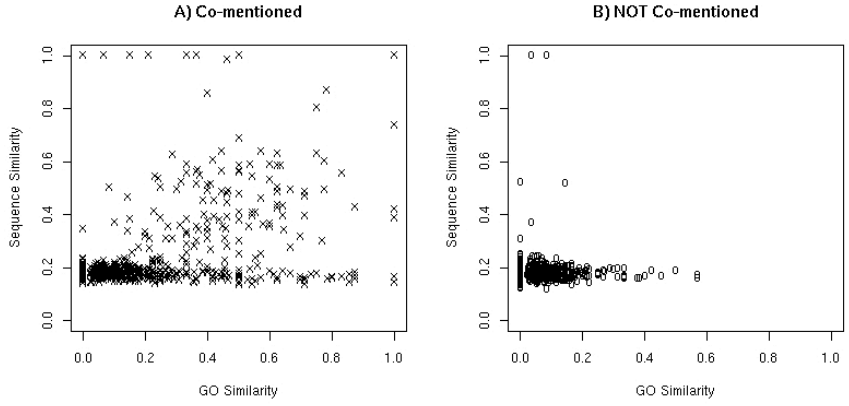
For the sequence, we used the longest protein RefSeq for each gene. For a measure based on functional annotations, we computed the Jacquard set similarity (1-Tanimoto distance) for the set of all GO annotations for each gene:

$$\text{Set Similarity} = \frac{|S_1 \cap S_2|}{|S_1| + |S_2| - |S_1 \cap S_2|}$$

We excluded all GO codes that had an accompanying qualifier, which for human genes, is restricted to "contributes_to", "colocalizes_with", and "NOT". This GO-derived similarity measure is a poor one for many reasons, including mixing experimental and homology based GO codes, ignoring the structure of GO, and ignoring the fact that the three main hierarchies are very different.

Figure 1 shows the result of computing these similarity measures for the 737 pairs of genes that are co-mentioned in our hand annotated training set and for 1,630 pairs of randomly selected genes which are explicitly *not* co-mentioned. Of the 737 co-mentioned pairs, 100 have both similarity measures above 0.3, while none of the 1,630 non co-mentioned pairs do. This suggests that in the context of the evaluation, even simple biological knowledge may be helpful in such tasks as disambiguation (dealing with polysemy) for normalization or in ascertaining if co-mention suggests functional and/or physical interaction or simply homology. It is hoped that this dataset can encourage the use of greater exploration into the use of biological knowledge to improve text mining.

Figure 1: Biological similarity between co-mentioned genes vs. not co-mentioned genes



5. Discussion

It is interesting to compare this new corpus with Task 1B of BioCreAtIvE 1 for insights into portability of normalization techniques. One set of measures in Table 3 seems to indicate that human may be easier than mouse; it has over twice the number of terms for each identifier, it has many fewer unique identifier targets, and

Table 3: A comparison of gene mention normalization

	Noisy Data Recall	Noisy Data Precision	Max Recall Approach Recall	Max Recall Approach Precision	Average Synonym Length in Words	Number of Unique ID's	Average # Synonyms/ Identifier	Average # Identifiers/ Synonym (ambiguity)	BioCreAtIvE 1 Max Submitted F-measure
Human	0.54	0.86	0.83	0.06	2.17	32,975	5.55	1.12	
Mouse	0.55	0.99	0.83	0.19	2.77	52,494	2.48	1.02	0.79
Yeast	0.86	0.99	0.93	0.33	1.00	7,928	1.86	1.01	0.92
Fly	0.81	0.86	0.85	0.07	1.47	27,749	2.94	1.09	0.82

only slightly more ambiguity. However, this does not really represent how the terms in the lexicon map to the text. The synonyms in the model organism databases are drawn from text, whereas the lexicon that we created for human genes includes database identifiers or descriptive forms that have very little overlap with actual text mentions. This overestimates the number of useful term variants in the lexicon and probably underestimates ambiguity in practice. The affects of polysemy/ambiguity in gene/protein mention identification is discussed in detail in [10].

An important contrast between human and mouse nomenclature on the one hand, and yeast and fly on the other, is that the nomenclature is often much more descriptive than nominal as mentioned in the **Task Definition** section. In *Drosophila*, the gene rather whimsically named "Son of sevenless" ("Sos") is named just that. It would never be called "child of sevenless" or "Sevenless' son". However, the names of human genes may vary quite a bit. The Alzheimer's disease related "APP" gene is generally known as "beta-amyloid precursor protein", although "beta-amyloid precursor polypeptide" may be used as well. Many other equivalent transformations are also acceptable, such as "amyloid beta-protein precursor", and "betaAPP". In general, any semantically equivalent description of the gene or protein may be used as a name. However, the regularity of the allowed transformations suggests that it might be possible to design or automatically learn transformation rules to permit better matching, something investigated by past researchers [18].

As Vlachos et al. observed [19], in biomedical text there is a high occurrence of families of genes and proteins being mentioned by a single term such as: "Mxi1

belongs to the Mad (Mxi1) family of proteins, which function as potent antagonists of Myc oncoproteins". In future work in biomedical entity normalization, we suggest that normalizing entity mentions to family mentions may be an effective way to support other biomedical text mining tasks. Possibly the protein families in InterPro [6] could be used as normalization targets for mentions of families. For example, the mention of "Myc oncoproteins" could link to InterPro:IPR002418. This would enable information extraction systems that extract facts (relations, attributes) on gene families to attach those properties to all family members.

6. Conclusion

In summary, we have described the motivation and development of a dataset for evaluating the automatic mapping of the mention of human genes/proteins to unique identifiers, which will be used as part of the second BioCreAtIvE. We have elucidated some of the properties of this data set, and made some suggestions about how it may be used in conjunction with biological knowledge to investigate the properties of co-mentioned genes and proteins. Anonymized submissions by evaluation participants along with the evaluation set gold standard annotations will be made publicly available [5] after the workshop, tentatively scheduled for the spring of 2007.

7. References

1. Hirschman, L., et al., *Overview of BioCreAtIvE task 1B: normalized gene lists*. BMC Bioinformatics, 2005. **6 Suppl 1**: p. S11.
2. Cohen, K.B., et al. *Contrast and variability in gene names*. in *Proceedings of the workshop on natural language processing in the biomedical domain, pp. 14-20. Association for Computational Linguistics*. 2002.
3. Podowski, R.M., et al., *AZuRE, a scalable system for automated term disambiguation of gene and protein names*. Proc IEEE Comput Syst Bioinform Conf, 2004: p. 415-24.
4. Fang, H., et al., *Human Gene Name Normalization using Text Matching with Automatically Extracted Synonym Dictionaries*, in *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*. 2006, Association for Computational Linguistics: New York, New York. p. 41--48.

5. http://biocreative.sourceforge.net/, *BioCreAtIvE 2 Homepage.*
6. Wu, C.H., et al., *The Universal Protein Resource (UniProt): an expanding universe of protein information.* Nucleic Acids Res, 2006. **34**(Database issue): p. D187-91.
7. Blaschke, C., et al., *Evaluation of BioCreAtIvE assessment of task 2.* BMC Bioinformatics, 2005. **6 Suppl 1**: p. S16.
8. Colosimo, M.E., et al., *Data preparation and interannotator agreement: BioCreAtIvE Task 1B.* BMC Bioinformatics, 2005. **6 Suppl 1**: p. S12.
9. Tsai, R.T., et al., *Various criteria in the evaluation of biomedical named entity recognition.* BMC Bioinformatics, 2006. **7**: p. 92.
10. Tuason, O., et al., *Biological nomenclatures: a source of lexical knowledge and ambiguity.* Pac Symp Biocomput, 2004: p. 238-49.
11. http://www.geneontology.org/, *The Gene Ontology.*
12. ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/, *NCBI Gene FTP site.*
13. Wain, H.M., et al., *Genew: the Human Gene Nomenclature Database, 2004 updates.* Nucleic Acids Res, 2004. **32**(Database issue): p. D255-7.
14. Wellner, B., *Weakly Supervised Learning Methods for Improving the Quality of Gene Name Normalization Data*, in *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*. 2005, Association for Computational Linguistics: Detroit. p. 1--8.
15. Morgan, A.A., et al., *Gene name identification and normalization using a model organism database.* J Biomed Inform, 2004. **37**(6): p. 396-410.
16. Aronson, A.R., *The effect of textual variation on concept based information retrieval.* Proc AMIA Annu Fall Symp, 1996: p. 373-7.
17. http://biopython.org, *BioPython Website.*
18. Hanisch, D., et al., *Playing biology's name game: identifying protein names in scientific text.* Pac Symp Biocomput, 2003: p. 403-14.
19. Vlachos, A., et al., *Bootstrapping the Recognition and Anaphoric Linking of Named Entities in Drosophila Articles.* Pac Symp Biocomput, 2006. **11**: p. 100-111.