Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Philadelphia, July 2002, pp. 14-20. Association for Computational Linguistics.

Contrast And Variability In Gene Names

K. Bretonnel Cohen
Center for Computational Pharmacology
University of Colorado Health Sciences Center
kevin.cohen@uchsc.edu

Andrew E. Dolbey
Center for Computational Pharmacology
University of Colorado Health Sciences Center
andrew.dolbey@uchsc.edu

George K. Acquaah-Mensah
Center for Computational Pharmacology
University of Colorado Health Sciences Center
george.acquaah-mensah@uchsc.edu

Lawrence Hunter
Center for Computational Pharmacology
University of Colorado Health Sciences Center
larry.hunter@uchsc.edu

Abstract

We studied contrast and variability in a corpus of gene names to identify potential heuristics for use in performing entity identification in the molecular biology domain. Based on our findings, we developed heuristics for mapping weakly matching gene names to their official gene names. We then tested these heuristics against a large body of Medline abstracts, and found that using these heuristics can increase recall, with varying levels of precision. Our findings also underscored the importance of good information retrieval and of the ability to disambiguate between genes, proteins, RNA, and a variety of other referents for performing entity identification with high precision.

1 Introduction

Almost all current approaches to entity identification are actually not tackling the identification per se, but rather merely the (still difficult) location of named entities in text. The difference between these is that entity location consists of the (difficult enough) task of demarcation of the boundaries of names in text, whereas entity identification consists of the same thing, plus mapping the located names to the canonical entities that they refer to. In this paper we present data on variability in the orthographic representation of gene names, and then show how knowledge about that variability can be used for heuristics that increase recall in the entity identification task. (We use the term "gene name" as shorthand for "gene, protein, or RNA name.")

To understand why it is important to be able to map located names to the canonical entities that they refer to, consider the outcome of running an information extraction routine with access only to entity location against a hypothetical document about rat somatotropin. It contains the synonymous names *rat somatotropin, somatotropin,* and *growth hormone,* all of which refer to the same biomolecule, whose canonical name we will assume to be *somatotropin.* (The document is hypothetical; somatotropin and its synonyms are not.) Suppose that the paper includes three separate assertions, of the form *somatotropin is upregulated by X, transcription of rat somatotropin is blocked by Y,* and *growth hormone is expressed by cells of type Z.* The system correctly extracts three assertions, but incorrectly attributes them to three separate biomolecules, only one of which is the canonical form. Now consider the outcome of running an information extraction routine with access to entity identification against the same document. Again, the system extracts three assertions, but this time all three assertions are correctly attributed to the same biomolecule, i.e. somatotropin. Krauthammer et al. (2000), who are arguably the only researchers who have attempted to do actual identification as we

define it, have noted that while it is possible to recognize gene names in the face of variability, it remains difficult to map the recognized names to their canonical referents. They point out that heuristics might be helpful in doing this.

We studied variability in gene names with an eye toward finding such heuristics. Our goal was to differentiate between kinds of variability that tend to differentiate between names with different referents, e.g. *aha* vs. *aho* or *ACE1* vs. *ACE2,* as opposed to kinds of variability that only differentiate between synonyms that share a referent, such as *tumour protein homologue* and *tumor protein homolog* or *ACE* and *ACE1*. We then use data on contrast and variability to suggest our heuristics. The idea behind using such heuristics is that an identified entity in some text that differs minimally from the canonical name for some entity can be mapped to that canonically-labelled entity if such mapping is allowed by some heuristic.

We use the term *contrast* to describe or refer to dimensions or features which can be used to distinguish between two samples of natural language with different meaning. Issues of contrast versus variability can be discussed with reference to individual characters or sequences of characters, or with reference to more abstract features, such as orthographic case. In the molecular biology domain, we will say that some feature is contrastive if it encodes the difference between the names of two different genes. In other words, contrasts occur inter-entity. We will say that some feature is (noncontrastively) variable if it differs merely between synonyms; in other words, variability occurs within members of a synonym set.

We can trivially identify the contrast between *BRCA1* vs. *MHC class I polypeptide-related sequence C,* or between *sonic hedgehog* vs. *eyeless.* What we are really interested in is minimally different tuples—sets that differ with respect to only one feature. For instance, we would want to look at *BRCA1* and *BRCA2,* which differ with respect to whether the character at the right edge is *1* or *2,* or *estrogen receptor beta* and *oestrogen receptor beta,* which differ with respect to the presence or absence of an *o* at the left edge. Ideally, then, we are looking for sets of names that differ only by a single unit. However, the size and scope of the unit needs further discussion. When dealing with written language, the unit of concern will usually be the grapheme. A grapheme may be as small as a single character, but may also be considerably longer, e.g. the sequence *ough* in *dough* or *through.* In this study, we considered graphemes longer than a single character only in the case of vowels. Sometimes we will want to consider tuples that differ with respect to strings that are considerably larger than a grapheme, such as a word, or a string of parenthesized material; this will be discussed further in the first *Methods* section. The issue of tuple size will be discussed there, as well.

2 Methods I: Investigating dimensions of contrast and variability

2.1 Corpus

We examined a large corpus of gene names and of synonyms for those gene names to determine what sorts of features are contrastive in gene names, and what sorts of features can vary without affecting the referential status of a gene name. The corpus was derived from the LocusLink `LL_tmpl` file (the version on the LocusLink download site at 2:32 p.m. on Sept. 13, 2001), available by ftp from `ftp://ncbi.nlm.nih.gov`. This is an easily readable dump of LocusLink, which "provides a single query interface to curated sequence and descriptive information about genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers, UniGene clusters, homology, map locations, and related web sites" (`www.ncbi.nlm.nih.gov/locuslink`). We then pulled out the names and synonyms for all LocusLink entries for the species Mus musculus, Rattus norvegicus, and Homo sapiens. We took the fields labelled as OFFICIAL GENE NAME, PREFERRED GENE NAME, OFFICIAL SYMBOL, PREFERRED SYMBOL, PRODUCT, PREFERRED PRODUCT, ALIAS SYMBOL, and ALIAS PROT. Some genes were unnamed, and these were excluded from the analysis. To our surprise, we also found that some gene names were duplicated within the same genome--e.g., in the M. musculus genome, there are two genes named *reciprocal translocation, Ch4 6*

and 7, Adler 17; we filtered out duplicate names and excluded them, as well. This left 42,608 genes for the mouse, 4457 for the rat, and 25,915 for the human. For each organism, we created one file containing just gene names, and for each organism we created a set of files containing all gene names and their synonyms. For the gene name file, we used just those fields labelled OFFICIAL GENE NAME or PREFERRED GENE NAME; for the combined name/synonym files, we used all of the fields given above.

2.2 Finding contrasts in the corpus

For each species, we pulled out a list of all names that were indicated as OFFICIAL GENE NAME or PREFERRED GENE NAME in the `LL_tmpl` file. Each name in this file represents a different gene. We examined the names in this single large file for contrastive differences.

2.3 Finding noncontrastive variability in the corpus

For each species, for each gene, we pulled out the list of all names that were indicated by any of the set of labels listed above, and stored them separately. With each of the many resulting files (one per gene), we examined the small set of synonymous names for noncontrastive variability.

2.4 Finding minimal tuples

The most obvious way to find minimal tuples would be to first determine the minimum edit distance between all pairs of gene names, and then select all pairs with minimum edit distance below some cutoff value. However, this approach would suffer from two obvious flaws. The first flaw is that it is computationally expensive, since it is a $O(n^2)$-complex problem. The second flaw is that it is ineffective. It only yields tuples of size 2, but in fact sets of minimally differing gene names occur in sets of size 3, 4, 5, and even considerably larger, e.g. the three-member set *conserved sequence block I, conserved sequence block II,* and *conserved sequence block III.* We chose an alternative approach to the problem of finding minimal tuples. It consists of the following steps:

For each gene name
- transform the gene name to some reduced form
- using the reduced form as the key in a hash of keys → lists, add the full form to a list of full forms from which that reduced form was derived

For each key in the hash
- retrieve the list of names that is mapped to by that key
- if the list of names pointed to by that key has more than one element, report the list

For example, if the input is the list of gene names *gamma-glutamyltransferase 1, gamma-glutamyltransferase 2, gamma-glutamyltransferase 3, matrix metalloproteinase 23A, matrix metalloproteinase 23B,* and *acrosin,* and the transformation that is being applied to each name consists of deletion of the last character, then the output will be two lists of > 1 element pointed to by *gamma-glutamyltransferase* and *matrix metalloproteinase 23,* and one list of a single element, *acrosin.* The two lists with > 1 element would be reported as minimal tuples.

2.5 Transformations

We applied four transformations designed to investigate syntagmatic, or positional, effects. These consisted of removing the first character, the first word, the last character, and the last word.

We applied four transformations designed to investigate paradigmatic, or content-based, effects. These consisted of mapping vowel sequences to a constant string (the purpose of this being to look at American vs. British dialectal differences in gene names); replacement of hyphens with spaces; removal of parenthesized material; and normalization of case. These relatively simple transformations miss a number of categories of differences between gene names, e.g. single-character differences in non-edge positions, such as *0 BETA-1 GLOBIN* vs. *0 BETA-2 GLOBIN;* single-word differences in non-edge positions, such as *DOPAMINE D1A RECEPTOR* vs. *DOPAMINE D2 RECEPTOR;* proper substring relationships, such as *EYE* vs. *EYE2;* and interactions between the features that we did examine, such as *calsequestrin 1 (fast-twitch, skeletal muscle)* vs. *calsequestrin 2 (cardiac*

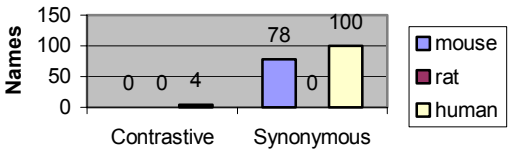
muscle), which is not found by any one heuristic but would be found by the combination of the parenthesized-material transformation followed by either of the right-edge transformations. Nonetheless, they seem like a reasonable starting point.

3 Results I

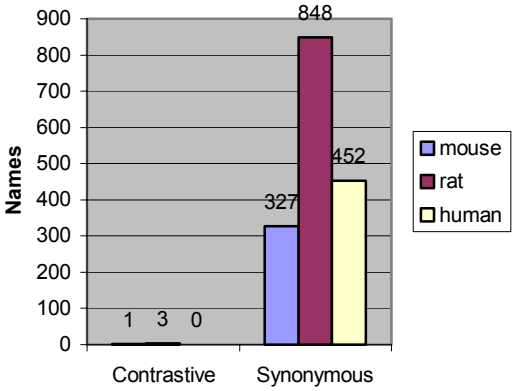
Table 1 and Graphs 1, 2, 3, and 4 summarize our findings on contrast and variability in gene names.

One surprising finding was that every paradigmatic dimension of contrast that we examined turned out to be contrastive in at least some very small number of cases. We did not expect hyphenation to ever be contrastive, but found that within the H. sapiens genome, the two genes at LocusLink ID's 51086 and 112858 differ in just that feature, having the names *putative protein-tyrosine kinase* and *putative protein tyrosine kinase,* respectively. The two genes at LocusLink ID's 51251 and 90859 differ in the same way, being named *uridine 5′-monophosphate hydrolase 1* and *uridine 5′ monophosphate hydrolase 1,* respectively.

Graph 1. Hyphenation: contrast and variability



Graph 2. Case: contrast and variability



DOC	*S*	*Contrastive (i.e., names)*	*%N*	*Variable (i.e., synonyms)*
LMC	M	4	.009	72
	R	4	0.090	801
	H	2	0.008	123
LMW	M	2556	5.999	135
	R	836	18.757	759
	H	4013	15.485	258
RMC	M	15540	36.472	360
	R	687	15.414	49
	H	8684	33.510	921
RMW	M	22290	52.314	191
	R	940	21.090	25
	H	11627	44.866	675
VS	M	7	0.016	37
	R	0	0.000	4
	H	4	0.015	30
HYPH	M	0	0.000	78
	R	0	0.000	0
	H	4	0.015	100
CASE	M	1	0.002	327
	R	3	0.067	848
	H	0	0.000	452
PM	M	14	0.033	102
	R	13	0.292	25
	H	51	0.197	526

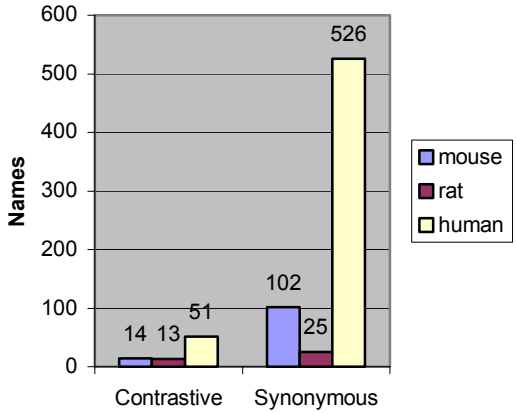
Table 1. Contrastive and noncontrastive variability in gene names. "Percentage" columns give percentage of total names considered for that species, rounded to three decimal places. DOC = dimension of contrast, L/RMC = left/right-most char, L/RMW = left/right-most word, VS = vowel sequences, HYPH = hyphenation, CASE = case, PM = parenthesized material. S = species, M = mouse, R = rat, H = human. %N = contrastive names as percentage of total names for that species.

We did not expect case ever to be contrastive, but found that within the R. norvegicus genome, the two genes at LocusLink ID's 24969 and 83789 differ with respect to just that feature, having the names *Ribosomal protein S2* and *ribosomal protein S2,* respectively. The two genes at LocusLink ID's 56764 and 65028 differ in the same way, having the names *dnaj-like protein* and *DnaJ-like protein.* As Graphs 1 and

2 show, these contrasts were not common, but we were surprised to observe them at all.

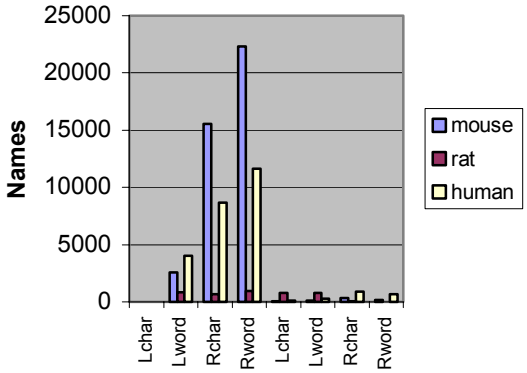
(In considering these findings, it should be noted that these results are specific to a particular version of LocusLink. We were interested in the extent to which these unexpected minimal pairs might be erroneous, so we examined the corresponding LOCUSID's in a subsequent revision of the file from several months later (May 1, 2002, 10:21 a.m.). We found that some of these entries had been combined, and some had been assigned an OFFICIAL_GENE_NAME, but others were unchanged, and so while we cannot eliminate the possibility that they are in error and have just managed to elude the editing process thus far, it is certainly the case that these anomalous contrasts continue to exist in the database, and we have no reason to assume that such names will not continue to be entered into the database, erroneously or otherwise, and therefore it behooves us to consider their implications for entity identification.)

Graph 3. Parenthesized material: contrast and variability



We found marked edge effects. Contrasts are much more likely to be marked at the name boundary than are noncontrastive differences. There is a marked asymmetry in the directionality of the location of contrastive differences: they are much more likely to appear at the right edge of the word than at the left edge of the word. There are also marked intra-species differences. For example, although large edge effects are obvious for names (as opposed to synonyms) in the mouse and human genomes, they are not in the rat genome. In interpreting variability, it will likely be helpful to have some awareness of what species is being discussed.

Graph 4. Edge effects: contrastive on left, synonymous on right



4 Methods II: Testing the heuristics

These findings suggested a set of heuristics for allowing weakened pattern matches on gene names. The heuristics are stated as transformations applied to regular expressions representing gene names to generate new regular expressions for the same gene names, but the heuristics can be applied in other ways as well, e.g. by grammar-based generation of alternate forms. The heuristics are listed below:

1. Equivalence of vowel sequences: for any regular expression representing a gene name, substitute the regular expression formed by replacing all vowel sequences with one or more of any vowel.

2. Optionality of hyphens: for any regular expression representing a gene name, substitute the regular expression formed by replacing every hyphen with the disjunction of a hyphen or a space.

3. Optionality of parenthesized material: for any regular expression representing a gene name, substitute the regular expression formed by making any paired parentheses and the material they enclose (and surrounding whitespace, as appropriate) optional.

4. Case insensitivity: for any regular expression, apply it case-insensitively.

To evaluate the extent to which each of these heuristics led to increased entity recognition, we ran our heuristics against a large body of Medline abstracts. We counted the number of entities that were found by an exact pattern match to a LocusLink name, and counted the number of additional names that were found by each heuristic. Although none of our heuristics specifically addressed morphologically-induced variability, we also added a search for pluralized gene names, so that we could compare the extent to which recognition of plurals improved recall to the extent to which our heuristics improved recall.

5 Results II

Table 2 shows the results. As intuition would suggest, all heuristics were effective in locating more names than strict pattern matches alone. For example, optional hyphenation heuristic allowed the official gene name *alpha-2-macroglobulin* to find a match in *Moreover, C5a also enhanced transcription of the gene for the type-2 acute phase protein alpha 2-macroglobulin n HC indirectly by increasing LPS-dependent IL-6 release from KC.*

names located by strict pattern matching	1846
Additional names located by vowel sequence heuristic matches	586
Additional names located by optional hyphen heuristic matches	37
Additional names located by case insensitive heuristic matches	864
Additional names located by optional parentheses heuristic matches	432
Additional names located by plural matches	87

Table 2. Names found by strict pattern match, heuristics, and plurals.

However, we were concerned about the possibility of poor precision, i.e. false-positives. For this reason, we ran our heuristics against the same body of Medline abstracts, then randomly selected up to 100 tokens of gene names suggested by each heuristic (some found less than 100 tokens in our corpus--see Table 2 above). We labelled each putative gene name with the canonical gene name that we believed it to refer to, and then asked a subject matter expert to evaluate whether the gene names that we had identified were or were not the gene names that we believed them to be. The expert was presented with a three-way forced-choice paradigm, the options being *yes, no,* and *can't tell.* It seemed useful to be able to compare the precision of our technique with the incidence of false positives from strict pattern matches, so the expert was also presented with a number of strict matches (i.e., not identified by our heuristics) to evaluate, in a quantity roughly equivalent to the number of heuristically-suggested names that they were asked to evaluate. Table 3 shows the results.

condition	*total tokens*	*marked yes*	*marked no*	*marked "can't tell"*	*percentage false positive ((can't tell + no) ÷ total tokens) x 100*
Vowel seq.	100	15	81	4	85.0
Hyph.	37	34	0	3	8.1
Case insens.	97	72	20	5	25.8
Paren. Material	99	93	0	6	6.1
Plurals	86	75	8	3	12.8
Strict pattern match	500	425	40	35	15.0

Table 3. False positives.

We note the following:

1. Even strict pattern matches and forms that vary only with respect to inflectional morphology (i.e., the plurals) yield a nontrivial percentage of false positives—a percentage which is actually higher than two of our heuristics (optionality of hyphenation and optionality of parenthesized material).

2. Two of our heuristics (equivalence of vowel sequences and case insensitivity) yielded unexpectedly high rates of false positives. The vowel sequence heuristic can probably be made to yield a lower rate by fine-tuning it. For example, false positives from this heuristic can be reduced by disregarding any weak matches

that come from one or more name-final upper-case *I*'s, since these are commonly used in gene names to form Roman numerals. The high false positive rate of the case insensitivity heuristic is unexpected, and will be investigated further.

6 Conclusions

Entity identification is a difficult task whose success is partly dependent on performance in other tasks, including disambiguation and information retrieval. Disambiguation of the actual referent of an apparent gene or protein name is even more important than one might expect. Hatzivassiloglou et al. (2001) points out the benefits and the difficulties inherent in distinguishing between genes, proteins, and RNA; we found that it was also important to differentiate between genes, proteins, RNA, and receptors, promoters, antagonists, domains, and binding sites, as well as diseases, syndromes, conditions, phenotypes, and mutants, as all of these were noted by our subject-matter expert as sources of false positives. Good information retrieval is clearly also a prerequisite for high-precision entity identification. In some cases, false positives arose when (abstracts of) irrelevant documents were used as input.

Heuristics can be useful tools for increasing recall in entity identification, as well as for helping us ensure that we are performing true entity identification, as opposed to entity location. Tanabe and Wilbur (in press) point out the value of combining knowledge sources in the entity identification task; our heuristics seem especially promising in part because they are based on a combination of two sources: (1) the expertise of NLP application developers about the sorts of variability that need to be dealt with in NLP systems (e.g. in text normalization), and (2) on empirical data about variability in the names themselves. Future work should concentrate on three areas. The first is extending our study of variability to include other dimensions of contrast, such as the ones that we point out that our study ignored, so that we can increase the inventory of heuristics. The second is integrating our heuristics with a system that identifies weak matches with gene names, i.e. candidates for application of the heuristics. The third is elucidating the place of orthographic variability within all causes of pattern match failure.

References

Hatzivassiloglou, Vasileios; Pablo A. Duboue; and Andrey Rzhetsky (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics* 17, Suppl. 1: S97-S106.

Krauthammer, Michael; Andrey Rzhetsky; Pavel Morozov; and Carol Friedman (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene* 259:245-252.

Tanabe, Lorraine and W. John Wilbur (in press). Tagging gene and protein names in biomedical text. *Bioinformatics*.