Ambiguity of Human Gene Symbols in LocusLink and MEDLINE: Creating an Inventory and a Disambiguation Test Collection

**Marc Weeber, PhD, Bob J. A. Schijvenaars, PhD, Erik M. van Mulligen, PhD,
Barend Mons, PhD, Rob Jelier, MSc, Christiaan van der Eijk, MSc, Jan A. Kors, PhD.
Department of Medical Informatics, Erasmus MC, PO Box 1738
3000 DR Rotterdam, The Netherlands**

ABSTRACT

Genes are discovered almost on a daily basis and new names have to be found. Although there are guidelines for gene nomenclature, the naming process is highly creative. Human genes are often named with a gene symbol and a longer, more descriptive term; the short form is very often an abbreviation of the long form. Abbreviations in biomedical language are highly ambiguous, i.e., one gene symbol often refers to more than one gene. Using an existing abbreviation expansion algorithm, we explore MEDLINE for the use of human gene symbols derived from LocusLink. It turns out that just over 40% of these symbols occur in MEDLINE, however, many of these occurrences are not related to genes. Along the process of making an inventory, a disambiguation test collection is constructed automatically.

INTRODUCTION

In the current era of genomics and proteomics, the massive analysis of biological data has become the default *modus operandi* in research. The data and possibly their interpretation with respect to, for instance, function and diseases, are put into huge online databases. In each of the databases, a gene has been assigned a unique identifier (UID). However, users will most likely not query a database with a UID; instead, they will search with a gene name or with other natural language terms indicating, for instance, a gene's function.

As genes and their products are discovered on a daily basis, new names have to be created constantly. Researchers are creative in this process, and few conventions exist. Gene names for Drosophila (fruit fly), for instance, are highly imaginative with examples such as *à la voile et à la vapeur, ken and barbie, lost in space, hu li tai sho, cheap date,* and *broken heart.* With respect to human genes, naming defaults to using a symbol, i.e., a combination of some, mostly capitalized, letters and digits and dashes. Also, there is often a longer or extended name. In many cases, the symbol is an abbreviation or acronym of the longer name. The symbol A2MP, for instance, expands to *alpha-2-macroglobulin pseudogene.*

Shows *et al* published the first guidelines for human gene nomenclature back in 1979 [1]. Since then, the Human Gene Nomenclature Committee (HGNC), part of the Human Genome Organisation (HUGO), has released guidelines and has developed a nomenclature database [2]. The practice of naming genes, however, shows that researchers do not strictly adhere to these guidelines, and it is often the case that a gene has more than one name or symbol, i.e. it has several *synonyms.* For instance, the symbols GNPDA and GNPI refer to one human gene, viz. *glucosamine-6-phosphate deaminase.* Also, one symbol may refer to more than one gene, which is a case of *homonymy.* The term NAP1, for instance, relates to at least five genes. One of the current tasks of the HGNC is therefore to assign one preferred name to a gene and compile a list of synonyms and also indicate homonyms.

Synonymy and homonymy pose serious challenges to genetic database indexing and retrieval systems. A user's query on GNPDA, for instance, should retrieve database records both on GNDPA and GNPI. In contrast, a query on NAP1 should first ask the user which of the different genes he or she is interested in, and then it should only retrieve records on the specified NAP1 gene. Synonymy can be handled adequately using a thesaurus, homonymy, on the other hand, asks for a more experimental approach of employing word sense disambiguation (WSD) algorithms. WSD algorithms decide on the basis of textual context which meaning is correct for a particular instance of a homonym. To train and test these algorithms, (manually) disambiguated data is necessary.

The extent of gene symbol usage in natural language text is not known. Additionally, the breadth of gene symbol homonymy, both for in-thesaurus and not-in-thesaurus meanings is unknown. The goal of this paper is therefore two-fold. First, we will provide an inventory of the actual use of gene symbols in

MEDLINE. Second, we will describe how to automatically compile a collection of homonyms with their disambiguated gene senses that can be used for testing disambiguation algorithms.

RELATED RESEARCH

Recently, WSD has seen an increased research interest, both in Natural Language Processing (NLP) and in biomedical informatics. Ambiguity is pervasive in natural language, and automatic systems that process language should be able to correctly disambiguate ambiguous expressions in order to achieve highly accurate results. The Senseval workshops[1] have become the major NLP arena for testing WSD algorithms.

Sometimes, it is claimed that language in restricted (research) environments is more specific and that there is less ambiguity. However, recent studies of both the UMLS and MEDLINE show that medical ambiguity is substantial [3-5]. Compared to normal language, medical language has one phenomenon for which ambiguity is paramount: Abbreviations. Medical terms often consist of multiple words, and the important terms are often abbreviated in the interest of economy. PSA, for instance, is used in MEDLINE as an abbreviation for *prostate specific antigen*, but also for *psoriasis arthritis* and *poultry science administration*, among others. Liu *et al* [5] showed that 81.2% of frequent MEDLINE abbreviations have more than one expansion, and thus are homonyms. Gene naming is a similar process: There is an abbreviation, or gene symbol, and there is the official gene name, the expansion of the abbreviation. In this paper, we use the term long form (LF) to refer to the full expansion, and short form (SF) to refer to the abbreviation or gene symbol.

Current research efforts in gene nomenclature can be grouped into two categories. First, there is the thesaurus-based approach in which groups of researchers try to compile a list of approved gene symbols. Most of these efforts are related to the genetic databases (HUGO, Swissprot). The other type of research is based on text mining in that different natural language text analysis techniques are employed to extract gene symbols from MEDLINE abstracts and digital full paper collections. This means, principally, that for each word in the text it has to be decided whether it is a gene symbol or another kind of word. A recent study on this form of gene tagging is [6]. As gene symbols are generally abbreviations of their longer form gene names, this kind of research boils down to employing abbreviation expansion algorithms [4, 5, 7-10]. Schwartz and Hearst [11] provide an extensive overview.

The research effort described in this paper tries to combine both approaches. It uses the Schwartz & Hearst abbreviation expansion algorithm to analyze MEDLINE. However, we are not interested in all abbreviations or all potential gene symbols, but restrict ourselves to those that are included in a thesaurus we derived from LocusLink[2], one of the more comprehensive genetic databases. Although this has several restrictions, we assume that the symbols in the thesaurus are referring (among others), to genes.

Disambiguation algorithms need data to be trained and tested for accuracy. The manual compilation of such test collections is a tedious exercise [3], and only few of those collections exist in the biomedical domain [3, 12]. Liu *et al* [7, 8] used an automatic approach to collect data for testing their abbreviation expansion algorithms by looking for explicit short form/long form (SF/LF) alignments in text. As an example, they use the SF CSF with four different LFs. Each MEDLINE abstract that contains CSF and one of the four long forms was retrieved and was assigned the meaning of the relevant long form. They compiled a test collection for 35 ambiguous three-character SFs. Liu *et al* [5] study the nature of three-character MEDLINE abbreviations and their coverage in the UMLS. Our approach is similar to Liu's, but our focus is on gene abbreviations, and we also will quantify the not-in-thesaurus meanings, i.e. other long forms of the homonymous short form.

MATERIALS AND METHODS

Using the LL3_030203 (February 2003) LocusLink data file, we compiled a thesaurus of gene symbols from the human gene records. We opted for LocusLink because it is one of the more comprehensive genetic databases with a rich set of gene name fields per record. We use the symbol record fields[3] to find symbols or short forms and the name record fields[4] to find the long forms. Note that we included the protein names as well. Although these are not strictly referring to genes, it is common practice in biomedicine do discuss genes by discussing their products (mRNA, proteins). This can also be observed in the naming of genes in

[1] http://www.senseval.org

[2] http://www.ncbi.nlm.nih.gov/LocusLink/

[3] OFFICIAL_SYMBOL, PREFERRED_SYMBOL, ALIAS_SYMBOL

[4] OFFICAL_GENE_NAME, PREFERRED_GENE_NAME, PREFERRED_PRODUCT, ALIAS_PROT

LocusLink. Many preferred gene names are actually protein names. LocusLink ID (LLID) 1, for instance, has the official symbol A1BG and the official gene name is alpha-1-B glyco*protein*. We think that the decision whether a symbol is a gene, protein or mRNA, if this distinction is needed at all, is better postponed to after identifying the symbol [13].

To align the SF to the potential LFs, we use the Schwartz & Hearst [11] algorithm of abbreviation expansion. The main reason for this choice is that it is a simple, efficient, and fast algorithm that does not need training. The accuracy of the algorithm is comparable to the more sophisticated (and computationally more intensive) machine learning algorithms that exist today. The principal idea of the algorithm is that it searches for a combination of a word within parentheses and it tries to match this word (on a character basis) to the preceding words.

After applying the algorithm, we obtain a list of aligned SF/LF pairs. As there are spelling variants in the LF (different uses of case, spaces instead of parenthesis or dashes, plural/singular issues), we normalize (and sort the words) of each LF using the UMLS[5] Specialist Lexicon [14]. When we use the term LF in the subsequent part of this paper, we actually use the normalized form of the LF. It is also possible that one SF aligns to more than one LF within one LocusLink record. If this is the case, the different LFs are synonyms and we still consider it as one SF/LF pair.

We assume that the expansion of the SF to the LF resolves the ambiguity of the SF. However, it turns out that this does not always hold true for the LocusLink data. There are SF/LF pairs that represent more than one gene meaning (i.e., LLID). In the current approach, it is not possible to distinguish these SF/LF pairs and they are therefore excluded from the analysis.

In gene symbol naming conventions, it is generally accepted that case does matter. Sometimes, case variants are used to distinguish the gene from its protein. However, there is no strict adherence to this convention. We therefore have a thesaurus variant in which all gene symbols are transformed to lowercase in order to match SFs case insensitive.

To study the use of gene symbols in MEDLINE, we compiled two databases. The SF database lists all occurrences of all LocusLink SFs in MEDLINE titles and abstracts from 1990 to 2002. For storage and computational efficiency, we only stored the SF in the exact LocusLink case variant. More specifically, we did not extract the lowercase variants from MEDLINE because it turned out that a number of lowercase SFs are highly frequent generic words.[6]

Using the Schwartz & Hearst algorithm, we also compiled the MEDLINE SF/LF database of all available SF/LF pairs in MEDLINE 1990-2002.

To compile the disambiguation test collection, we executed the following procedure. For every SF in LocusLink, we extract all corresponding LFs from the SF/LF MEDLINE. If there was an exact match with a LocusLink LF, we added the PubMed ID (PMID) to the collection and recorded the meaning, i.e. the LLID. For example, the SF A2M has the LocusLink LF of *alpha-2-macroglobulin.* If a (normalized) MEDLINE LF exactly matched this (normalized) LF, the PMID in which this expansion occurred, is assigned the LLID of 2.

During initial experimentation we observed that there is not always an exact match between a MEDLINE LF and a LocusLink LF while the meaning is identical. In order not to contaminate the NIT data with in-thesaurus data, we decided to apply a strict “not matching” rule, i.e. only if a MEDLINE LF did not have any word in common with all possible LocusLink LFs for that particular SF, then the PMID was assigned the LLID of 0 representing the NIT meaning. If there was some partial overlap between MEDLINE and LocusLink LFs, they were not included in the test collection. We executed this test collection compilation procedure for case sensitive and case insensitive SFs separately.

RESULTS

There is a total of 49,867 different gene symbols or short forms in LocusLink. 20,720 (41.5%) of those occur in MEDLINE. From the 24,786 SF that have an aligned LocusLink LF, 12,638 (51.0%) occur in MEDLINE. Figure 1 provides the frequency distributions for the LocusLink SFs in MEDLINE. Both distributions have a Zipfian shape, which means that few SFs occur very frequently in MEDLINE, and many SFs occur rarely.

There are 24,393 human genes in Locuslink, of which 12,533 (51.4%) have at least two symbols (or synonyms). 16,828 genes have a matching LF. Of these genes, 5,812 (34.5%) have more than one SF. When we ignore case variation, the number of genes with more than one SF decreases slightly (51.1% respectively 34.4%).

Before turning to the overall homonym data, we will present the data obtained for one example: PSA. In Locuslink, the SF PSA has been assigned to five different genes (Table 1). One of those does not have a LF. The fourth column provides the number of

[5] http://umlsks.nlm.nih.gov

[6] Among others: as, on, in, not, at, an, in, of, was, can, for, up.

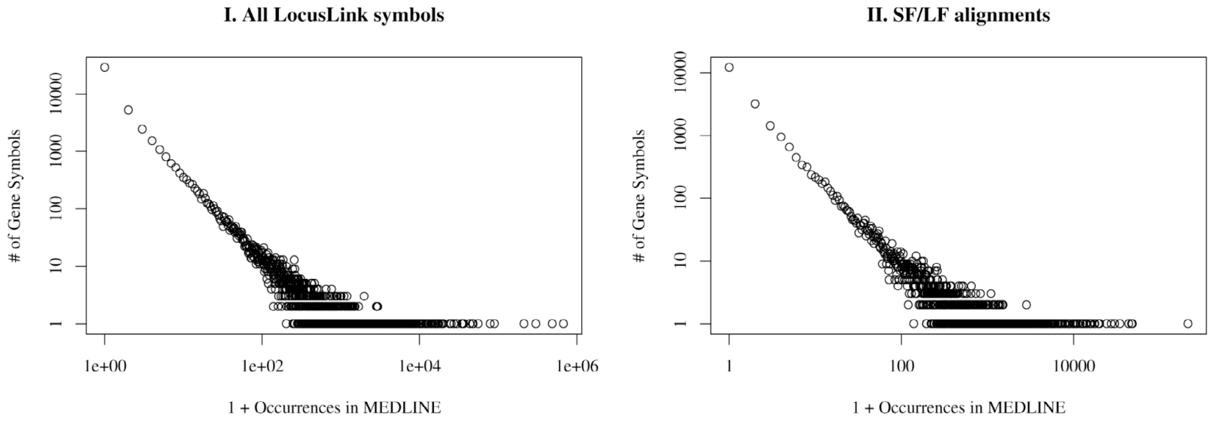

Figure 1. Frequency of LocusLink gene symbols in MEDLINE 1990-2002. Panel I provides the data for all 49,867 LocusLink gene symbols, Panel II for the 24,786 gene symbols that have an aligned long form.

PMIDs of each SF/LF combo. The SF/LF pair of PSA/protein s, alpha does not appear in MEDLINE, whereas the prostate specific antigen meaning of PSA is frequent. If we ignore case, we observe an increase in PMIDs (column 5): two additional PMIDs for LLID 9520, and 210 for the not-in-thesaurus (NIT) meaning. The final column shows the number of case variants in MEDLINE. For LLID 9520, the variants are PSA and Psa; for the NIT meaning, the variants are: PSA, PsA, Psa, PSa, psa, and pSA. In the MEDLINE SF/LF database there was a total of 182 different LFs for PSA.

Of the 49,867 different gene symbols in LocusLink, 1877 (3.8%), refer to more than one gene, and are therefore homonyms. Of the 24,786 SFs that have one or more LF alignments, 563 are homonyms (2.3%), which amounts to a total of 25,480 different SF/LF pairs. When case is ignored, the percentages remain about equal (4.0% and 2.3%). There were 61 cases with homonymous SF/LF pairs. An example is ZNF408/ zinc finger protein 408. This SF/LF pair refers to two different LLID. All 61 cases have been excluded from the analysis.

For each of the 24,786 different SFs, an additional LLID of 0 is included representing a potential not-in-thesaurus (NIT) meaning. There are 25,480 + 24,786 = 50,266 SF/LF entries possible. 6,340 of these SF/LF pairs (12.6%) were found at least once in MEDLINE. 3,012 had an in-thesaurus LF, 3328 had a NIT LF (52.5%). When the SF was matched case insensitive, there were 7,268 SF/LF pairs (14.6% increase over case sensitive) that appear in MEDLINE, of which 3,856 had a NIT LF (15.9% increase). Table 2 provides more detailed information. There are for instance 1,194 short forms that have 2 different meanings or LLIDs in MEDLINE. One of these two meanings is not-in-thesaurus for 1,187 of those SFs.

The disambiguation test collection consists of all SFs for which at least two different LLIDs were found, i.e. the final three rows in Table 2. This means there are 1,247 different SFs included.

The number of PMIDs per meaning varies considerably. The case sensitive test collection consists of 425,577 PMIDs of which 157,167 represent a LocusLink meaning. The case insensitive test collection consists of 493,555 PMIDs, a 16.0% increase, of which 176,164 have a LocusLink meaning (12.1% increase). Detailed counts can be obtained from the test collection itself.

Table 1. Example MEDLINE data for gene symbol or SF *PSA*. LLID = Locus Link ID, LF = long form, CS = case sensitive, CI = case insensitive, Variants = number of SF variants used in the case insensitive counts.

LLID	LF	# PMID CS	#PMID CI	# Variants
354	prostate specific antigen	3830	3830	1
5627	protein s, alpha	0	0	0
7996	-	-	-	-
9520	puromycin-sensitive aminopeptidase	10	12	2
29968	phosphoserine aminotransferase	1	1	1
0	not in thesaurus (NIT)	418	628	6

Table 2. MEDLINE counts of SF/LF combinations. CS = case sensitive, CI = case insensitive, NIT = not in thesaurus.

LLID / SF	CS: # SF (# NIT)	CI: # SF (# NIT)
0	19,750	18,938
1	3,789 (2088)	4,216 (2373)
2	1,194 (1187)	1,433 (1423)
3	49 (49)	54 (54)
4	4 (4)	6 (6)

DISCUSSION

The analysis of LocusLink genes showed that synonymy is a general phenomenon: more than half the genes have at least two different known symbols. Homonymy of gene symbols, however, is low in LocusLink (only 3.8%). It is likely that database curators try to reduce homonymy for indexing. However, the actual use of symbols in MEDLINE shows that homonymy is widespread. The main reason is that there are many not-in-thesaurus (NIT) expansions of the gene symbols. We also expect that the combination of different thesauri derived from different genetic databases will increase the in-thesauri homonymy.

Indexing systems should be able to distinguish between the different expansions. The automatically created test collection may assist in testing disambiguation algorithms for such purposes. The test collection developed for this paper is available at http://www.biosemantics.nl

Use of case is not uniform in natural language. Gene symbols may occur in different case variants. If the SF is found in MEDLINE with case insensitivity, the number of SF/LF pairs increases by about 15% Also, the number of PMIDs in the test collection increases by 16%; however, it turns out that the increase is higher for NIT meanings.

The test collection uses the alignment of long form to short form in order to disambiguate. Although the Schwarz & Hearst algorithm has not an 100% accuracy [11], the test collection is not likely to reflect the (few) flaws of the algorithm. Only if there is an exact match between the MEDLINE and the LocusLink LF the data is included in the test collection. This probably means that there are more correct SF/LF alignments in MEDLINE than we have included, but we expect a near 100% precision of our test collection.

The Zipfian distribution of short forms in MEDLINE (Figure 1) shows that abbreviations are a typical natural language phenomenon in that many entities occur with a low frequency while only a few entities have a high frequency of use.

Our thesaurus-based approach has the drawback of ignoring recently discovered genes because there is a delay between discovery and thesaurus incorporation. Recent text mining research for finding new gene names and synonyms [15, 16] may help overcome this drawback.

REFERENCES

1. Shows TB, Alper CA, Bootsma D, Dorf M, Douglas T, Huisman T, et al. International system for human gene nomenclature. Cytogenet Cell Genet 1979;25(1-4):96-116.
2. Wain HM, Lush M, Ducluzeau F, Povey S. Genew: the human gene nomenclature database. Nucleic Acids Res 2002;30(1):169-71.
3. Weeber M, Mork JG, Aronson AR. Developing a test collection for biomedical word sense disambiguation. Proc AMIA Symp 2001:746-50.
4. Liu H, Lussier YA, Friedman C. A study of abbreviations in the UMLS. Proc AMIA Symp 2001:393-7.
5. Liu H, Aronson AR, Friedman C. A study of abbreviations in MEDLINE abstracts. Proc AMIA Symp 2002:464-9.
6. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics 2002;18(8):1124-32.
7. Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. J Am Med Inform Assoc 2002;9(6):621-36.
8. Liu H, Lussier YA, Friedman C. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. J Biomed Inform 2001;34(4):249-61.
9. Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in biomedical articles. J Am Med Inform Assoc 2002;9(3):262-72.
10. Chang JT, Schütze H, Altman RB. Creating an online dictionary of abbreviations from MEDLINE. J Am Med Inform Assoc 2002;9(6):612-20.
11. Schwartz AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. Pac Symp Biocomput 2003:451-62.
12. Raileanu D, Buitelaar P, Vintar S, Bay J. Evaluation corpora for sense disambiguation in the medical domain. Proc Third International Conference on Language Resources and Evaluation (LREC) 2002:609-612.
13. Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics 2001;17 Suppl 1:S97-106.
14. McCray AT. The nature of lexical knowledge. Methods Inf Med 1998;37(4-5):353-60.
15. Liu H, Friedman C. Mining terminological knowledge in large biomedical corpora. Pac Symp Biocomput 2003:415-26.
16. Yu H, Hatzivassiloglou V, Friedman C, Rzhetsky A, Wilbur WJ. Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. Proc AMIA Symp 2002:919-23.