Combination of Genetic Databases for Improving Identification of Genes and Proteins in Text

Jan A. Kors, Martijn J. Schuemie, Bob J.A. Schijvenaars, Marc Weeber, Barend Mons
Department of Medical Informatics
Erasmus University Medical Center
Rotterdam, The Netherlands
{j.kors,m.schuemie,r.schijvenaars,m.weeber,b.mons}@erasmusmc.nl

Abstract

Many genes and proteins have alternative symbols and names in addition to their official ones. This complicates their identification and hampers the retrieval and extraction of information about genes from the literature. In order to merge complementary nomenclature information contained in genetic databases, we developed an algorithm that integrates this information into one thesaurus. Application of the algorithm to five human genetic databases substantially increased the number of synonyms per gene over that contained in each of the databases separately, as well as the total number of genes. The combined thesaurus can be a helpful resource in information retrieval and text mining applications.

1 Introduction

Absence of standards of nomenclature for genes and proteins or lack of adherence to available standards have created a plethora of symbols and names (Pearson, 2001; Tuason et al., 2004). More often than not, genes or proteins have several synonyms, so that different symbols or names refer to the same gene. This has considerable implications for the retrieval and extraction of information from text. For example, literature queries that do not use all available synonyms for a gene may miss important documents. In applications that automatically try to distill relationships between genes and gene products from large text corpora (Jenssen et al., 2001; Shatkay and Feldman, 2003; Hoffmann and Valencia, 2004; Wren et al., 2004), important relationships may be missed because the same gene can hide under various aliases.

To deal with these problems one may utilize information about synonymous gene and protein symbols and names (together these will be referred to as "terms"), which is present in many genetic databases. However, this information is likely to differ between databases, both in the number of genes contained in a database and in the synonyms listed per gene. A combination of various databases should be advantageous, in that the combined thesaurus will contain more genes and more gene symbols than each of the constituent databases. Here we present an algorithm that merges the gene and protein terms from different databases into one combined thesaurus, and show the increase in genes and synonyms when the algorithm is applied to five human genetic databases.

2 Material and Methods

We downloaded (February 2005) information about human genes and proteins from five curated databases: Genew, GDB, Entrez Gene, OMIM, and Swiss-Prot. We chose not to distinguish between genes and proteins. In practice, gene and protein terms are often used interchangeably and the distinction is difficult to make (Hatzivassiloglou et al., 2001). For each database entry, gene and protein symbols (including aliases), names, and identification codes were extracted. Only the names were normalized using the lvg tool (http://umlslex.nlm.nih.gov/lvg/current/). The number of identification codes per gene varied from database to database. Each database maintains its own set of identification codes, but may

also include cross-references to one or more of the other databases. Also gene identification codes from Unigene and RefSeq were extracted if available. The original databases, including Unigene, were searched for information about obsolete identification codes and their possible replacements, and the extracted codes were corrected or excluded as appropriate. Database entries with the status "withdrawn" were not included.

The combining algorithm consists of two stages (Fig. 1). First, genes from the different databases with any matching identification code are grouped. If there are conflicting identification codes, a procedure is entered to determine whether there are subgroups of genes that represent the same gene. Identification codes, symbols, and names of similar genes are merged, and a second stage is entered. In this stage, genes that have no common identification codes but share any term are grouped. Depending on the extent of the overlap in terms between the genes, a decision is made whether they are to be considered as identical.

```
input_genes = Ø                                   # start with empty gene list
for all genetic databases D_i do                  # add information from databases
    add gene identifiers, names, and symbols from D_i to input_genes
end
temp_genes = combine_genes(input_genes, ID)                  # stage 1: match genes on identification numbers
output_genes = combine_genes(temp_genes, SYMBOL_OR_NAME)     # stage 2: match on symbols or names, output_genes
                                                                 contains the genes in the combined thesaurus
combine_genes(input_list, match_type)
    while at least one gene g in input_list do
        R = get_related_genes(g, input_list, match_type)         # get all genes with any match of match_type
        G = Ø
        for all genes g_i in R do
            if g_i similar to a gene g_j in G then       # see decision table below
                merge g_i with g_j                       # combine the information from both genes
            else
                add g_i to G                             # new gene not previously seen
            end
        end
        add G to output_list
        remove R from input_list
    end
    return output_list
end

get_related_genes(gene, gene_list, match_type)
    R = {g_i | g_i ∈ gene_list ∧ g_i and gene have at least one item of match_type in common}
    for all g_i in R and g_i != gene do
        R = R ∪ get_related_genes(g_i, gene_list, match_type)
    end
    return R
end
```

No. of discordant ID codes	0	1	2	3	4
3	D	D	D	D	D
2	D	D	D/S (O = 1)	D/S (O ≥ .5)	S
1	D	D/S (O ≥ .5)	S	S	S
0	D/S (O ≥ .5)	S	S	S	S

No. of concordant ID codes

Decision table to determine whether two genes are similar. The decision is based on the number of identification codes that both genes have in common (horizontal axis) and that conflict (vertical). The elements in the table indicate the decision outcome: D=different; S=same; D/S=either different or same, dependent on the symbol overlap, O. Overlap is defined as the number of symbols both genes have in common divided by the number of symbols of either gene, whichever is smaller; if the specified condition is fulfilled, the genes are considered the same, else different.

Figure 1. Pseudo-code of the algorithm that combines the information from different genetic databases into one thesaurus.

Database	Entrez Gene	GDB	Genew	OMIM	Swiss-Prot
Entrez Gene	25671 (96.7)	15644	20399	9018	11267
	117380 (71.6)	*45493*	*60755*	*16617*	*23908*
GDB		16119 (60.7)	15394	8229	9762
		51865 (31.6)	*37680*	*12233*	*15793*
Genew			20671 (77.9)	8433	10870
			62971 (38.4)	*10986*	*16969*
OMIM				9073 (34.2)	7048
				24443 (14.9)	*11110*
Swiss-Prot					11644 (43.9)
					54982 (33.5)

Table 1. Number of genes and terms in five genetic databases and the overlap between databases. Black numbers indicate genes, grey numbers in italics indicate terms. Numbers in parentheses on the diagonal denote percentages of genes and terms in the combined thesaurus covered by the separate databases.

3 Results

The combined thesaurus contains information on 26,552 human genes and proteins, with a total of 163,896 symbols and names, and is available as Supplementary Data online. The overlap with the original databases is shown in Table 1. EntrezGene is the most comprehensive of the databases considered, covering 96.7% of the genes and proteins and 71.6% of the terms in the new thesaurus; Genew, maintained by the HUGO Gene Nomenclature Committee, covers 77.9% of the genes and 38.4% of the terms. The average number of terms per gene in the original databases varies from 2.69 (in OMIM) to 4.72 (in Swiss-Prot); the combined thesaurus has an average of 6.17 terms per gene.

Part of the different terms per gene may be attributed to spelling variations. To assess the effect of this variability on the number of terms, we applied four simple rewrite rules to reduce each term to a canonical form: (1) terms ending in one or more digits had a space or hyphen preceding the digits removed; (2) Roman numerals I to IX at the end of a term were replaced by the corresponding digit; (3) Greek symbols (alpha, beta) were replaced by the first character of the symbol; (4) all terms were converted to lower case. For the original databases, the decrease in number of terms after rewriting varied between 9 (OMIM) and 2,334 (Swiss-Prot). The number of terms in the combined thesaurus decreased by 14,223 (to 149,673 terms), showing that there is spelling variation of terms across databases beyond that present in each database. Still, the total number of terms is considerably larger than in the largest individual database (Entrez Gene, 115,522 terms after rewriting).

To assess how well the combined thesaurus covers gene and protein terms used in the literature, we used it to find terms in a set of 67,991 Medline abstracts. Each abstract is referenced in Entrez Gene for a specific gene, and we assume this gene or its product to be mentioned at least once in the abstract. If the abstract was referenced for more than five genes, it was not included because of the chance that part of the genes would only be mentioned in the full text rather than in the abstract. On this test set, we obtained a recall (percentage of abstracts in which the correct term was found) of 74.7%. For comparison, when using the original databases recall varied between 37.9% (OMIM) and 67.5% (Entrez Gene). When we used a version of the combined thesaurus in which the names were normalized with the lvg tool, recall increased to 79.5%.

4 Discussion

Our results indicate that the combination of information from standard genetic databases expands the number of genes beyond that found in each database separately, at the same time increasing the average number of synonyms per gene, and thus can help to alleviate the synonym problem. This holds even true for the most comprehensive database, Entrez Gene. While Entrez Gene covers most genes in the combined thesaurus, it contains only 71.6% of the terms. When the combined thesaurus was used to find terms in a large set of abstracts, recall improved by 10% in comparison to Entrez Gene. Further recall improvement may be feasible by applying term variation rules (Tsuruoka and Tsujii, 2004).

Synonym identification may further improve by culling information from a greater number of genetic databases, although it is prudent to require rigorous database curation in order not to include spurious terms. Additionally, synonyms may automatically be extracted. Yu et al. (2002) employed pattern recognition rules to extract synonymous gene symbols, but required that synonyms be listed within the same abstract or article.

Recently, machine learning approaches have been proposed to automatically recognize biomedical entities in text without the help of a thesaurus (see Zhou et al. (2004) for a recent overview). These techniques may supplement our thesaurus-based approach to further increase the coverage of the gene thesaurus, but do not give clues whether newly detected symbols are synonyms for existing genes. In this respect, they are of limited value in solving the synonym problem.

Several other investigators have previously combined nomenclature information from different databases (Jenssen et al., 2001; Koike and Takagi, 2004). Combination in these approaches was mostly based on correspondence between terms rather than on identification codes. Also, the effect of a combination strategy on the number of genes and terms has not been assessed before.

Limitations: There are several potential limitations of our approach. First, there is a time lag between the publication of a new or alternative symbol or name and its becoming available in a thesaurus. This problem is becoming less bothersome now that many biological journals require that a new gene symbol be registered by the appropriate nomenclature committee before publication. Still, a combined thesaurus must be updated frequently. An automatic approach as described here would allow such an update, possibly even daily.

Second, in the second stage of our algorithm genes can be combined that have common terms but no common identification codes. Care is taken to only combine genes that have considerable term overlap. Nevertheless, it cannot be excluded that some genes are taken to be the same that are not because the overlapping terms in fact are ambiguous and refer to multiple genes. Since relatively few of the 27,273 genes that result after the first stage are combined in the second stage to yield the final 26,552 genes, the impact of this potential difficulty will probably not be large.

Conclusion: The combination of information from different genetic databases can alleviate the synonym problem in information retrieval and extraction and can help biologists to find pertinent biological information more straightforwardly. The combination algorithm was applied to human genetic databases, but may also be used to create comprehensive thesauri for other organisms.

References

Hatzivassiloglou, V., P.A. Duboue, and A. Rzhetsky. 2001. Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics,* 17:S97-106.

Hoffmann, R. and A. Valencia. 2004. A gene network for navigating the literature. *Nat Genet,* 36:664.

Jenssen, T.K., A. Laegreid, J. Komorowski, and E. Hovig. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet,* 28:21-28.

Koike, A. and T. Takagi. 2004. Gene/protein/family name recognition in biomedical literature. In *BioLINK Workshop*, pp. 9-16. Association for Computational Linguistics, Boston.

Pearson, H. 2001. Biology's name game. *Nature,* 411:631-632.

Shatkay, H. and R. Feldman. 2003. Mining the biomedical literature in the genomic era: an overview. *J Comput Biol,* 10:821-855.

Tsuruoka, Y. and J. Tsujii. 2004. Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform,* 37:461-470.

Tuason, O., L. Chen, H. Liu, J.A. Blake, and C. Friedman. 2004. Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pac Symp Biocomput,*238-249.

Wren, J.D., R. Bekeredjian, J.A. Stewart, R.V. Shohet, and H.R. Garner. 2004. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics,* 20:389-398.

Yu, H., V. Hatzivassiloglou, C. Friedman, A. Rzhetsky, and W.J. Wilbur. 2002. Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. *Proc AMIA Symp,*919-923.

Zhou, G., J. Zhang, J. Su, D. Shen, and C. Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics,* 20:1178-1190.