Automatic Indexing of Specialized Documents: Using Generic *vs.* Domain-Specific Document Representations

Aurélie Névéol **James G. Mork** **Alan R. Aronson**

{neveola,mork,alan}@nlm.nih.gov

National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894
USA

Abstract

The shift from paper to electronic documents has caused the curation of information sources in large electronic databases to become more generalized. In the biomedical domain, continuing efforts aim at refining indexing tools to assist with the update and maintenance of databases such as MEDLINE®. In this paper, we evaluate two statistical methods of producing MeSH® indexing recommendations for the genetics literature, including recommendations involving subheadings, which is a novel application for the methods. We show that a generic representation of the documents yields both better precision and recall. We also find that a domain-specific representation of the documents can contribute to enhancing recall.

1 Introduction

There are two major approaches for the automatic indexing of text documents: statistical approaches that rely on various word counting techniques [such as vector space models (Salton, 1989), Latent Semantic Indexing (Deerwester et al., 1990) or probabilistic models (Sparck-Jones et al., 2000)] and linguistic approaches that involve syntactical and lexical analysis [see for example term extraction and term variation recognition in systems such as MetaMap (Aronson, 2001), FASTR (Jacquemin and Tzoukermann, 1999) or IndDoc (Nazarenko and Ait El Mekki, 2005)]. In many cases, the combination of these approaches has been shown to improve the performance of a single approach both for controlled indexing (Aronson et al., 2004) and free text indexing (Byrne and Klein, 2003).

Recently, Névéol et al. (2007) presented linguistic approaches for the indexing of documents in the field of genetics. In this paper, we explore a statistical approach of indexing for text documents also in the field of genetics. This approach was previously used successfully to produce Medical Subject Headings (MeSH) main heading recommendations. Our goal in this experiment is twofold: first, extending an existing method to the production of recommendations involving subheadings and second, assessing the possible benefit of using a domain-specific variant of the method.

2 A k-Nearest-Neighbors approach for indexing

2.1 Principle

The k-Nearest-Neighbors (k-NN) approach views indexing as a multi-class classification problem where a document may be assigned several "classes" in the form of indexing terms. It requires a large set of labeled data composed of previously indexed documents. k-NN relies on the assumption that similar documents should be classified in a similar way. The algorithm consists of two steps: 1/documents that are most "similar" to the query document must be retrieved from the set of labeled documents. They are considered as "neighbors" for the query document; 2/an indexing set must be produced from these and assigned to the query document.

Finding similar documents

All documents are represented using a vector of distinctive features within the representation space. Based on this representation, labeled documents

BioNLP 2007: Biological, translational, and clinical language processing, pages 183–190,
Prague, June 2007.

may be ranked according to their similarity to the query document using usual similarity measures such as cosine or Dice. The challenge in this step is to define an appropriate representation space for the documents and to select optimal features for each document. Another issue is the number (k) of neighbors that should be selected to use in the next step.

Producing an indexing set

When applied to a single-class classification problem, the class that is the most frequent among the k neighbors is usually assigned to the query document. Indexing is a multi-class problem for which the number of classes a document should be assigned is not known, as it may vary from one document to another. Therefore, indexing terms from the neighbor documents are all taken into account and ranked according to the number of neighbors that were labeled with them. The more neighbors labeled with a given indexing term, the higher the confidence that it will be a relevant indexing term for the query document. This resulting indexing set may then be filtered to select only the terms that were obtained from a defined minimum number of neighbors.

2.2 Document representation

Generic representation

A generic representation of documents is obtained from the text formed by the title and abstract. This text is processed so that punctuation is removed, stop-words from a pre-defined list (of 310 words) are removed, remaining words are switched to lower case and a minimal amount of stemming is applied. As described by Salton (1989) words should be weighted according to the number of times they occur in the query document and the number of times they occur in the whole collection (here, MEDLINE). Moreover, words from the title are given an additional weight compared to words from the abstract. Further adjustments relative to document length and local weighting according to the Poisson distribution are detailed in (Aronson et al, 2000; Kim et al., 2001) where the PubMed Related Citations (PRC) algorithm is discussed. Further experiments showed that the best results were obtained by using the ten nearest neighbors.

Domain-specific representation

In specialized domains, documents from the literature may be represented with concepts or objects commonly used or studied in the field. For example, (Rhodes et al., 2007) meet specific chemistry oriented search needs by representing US patents and patent applications with molecular information in the form of chemical terms and structures. A similar representation is used for PubChem (http://pubchem.ncbi.nlm.nih.gov/) records. In the genetics domain, genes are among the most commonly discussed or manipulated concepts. Therefore, genes should provide a relevant domain-specific description of documents from the genetics literature.

The second indexing algorithm that we describe in this paper, know as the Gene Reference Into Function (GeneRIF) Related Citations (GRC) algorithm, uses “GeneRIF” links (defined in the paragraph below) to retrieve neighbors for a query document.

To form a specific representation of the document, gene names are retrieved by ABGene[1] (Tanabe and Wilbur, 2002) and mapped to Entrez Gene[2] unique identifiers. The mapping was performed with a version of SemRep (Rindflesch and Fiszman, 2003) restricted to human genes. It consists in normalizing the gene name (switch to lower case, remove spaces and hyphens) and matching the resulting string to one of the gene names or aliases listed in Entrez Gene.

For each gene, the GeneRIF links supply a subset of MEDLINE citations manually selected by NLM indexers for describing the functions associated with the gene. These sets were used in two ways:

To complete the document representation. If a citation was included in the GeneRIF of a given gene, the gene was given an additional weight in the document representation.

To limit the set of possible neighbors. In the generic representation, all MEDLINE citations contain the representation features, words. Therefore, they all have to be considered as potential neighbors. However,

[1] Software downloaded January 17, 2007, from http://www.ncbi.nlm.nih.gov/staff/lsmith/MedPost.html

[2] Retrieved January 17, 2007, from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

only a subset of citations actually contains genes. Therefore, only those citations need to be considered as potential neighbors. This observation enables us to limit the specific processing to relevant citations. Possible neighbors for a query document consist of the union of the GeneRIF citations corresponding to each gene in the document representation.

Table 1: Gene description of a sample MEDLINE document and its two nearest neighbors

PubMed IDs	ABGene	Entrez Gene IDs
15645653	abcc6	368
	mrp6	368; 6283
	ldl-r	3949
	pxe	368; 5823
	fh	2271
10835643	mrp6	368; 6283
	pxe	368; 5823
16392638	abcc6	368
	mrp6	368; 6283
	pxe	368; 5823

For each query document, the set of possible neighbors was processed and ranked according to gene similarity using a cosine measure. Table 1 shows the description of a sample MEDLINE citation and its two nearest neighbors.

Based on experiments with the PubMed Related Citations algorithm, ten neighbors were retained to form a candidate set of indexing terms.

3 Experiment

3.1 Application to MeSH indexing

In the MEDLINE database, publications of the biomedical domain are indexed with Medical Subject Headings, or MeSH descriptors. MeSH contains about 24,000 main headings denoting medical concepts such as *foot*, *bone neoplasm* or *appendectomy*. MeSH also contains 83 subheadings such as *genetics*, *metabolism* or *surgery* that can be associated with the main headings in order to refer to a specific aspect of the concept. Moreover, each descriptor (a main heading alone or associated with one or more subheadings) is assigned a “minor” or “major” weight depending on how substantially the concept it denotes is discussed in the article. “Major” descriptors are marked with a star.

In order to form a candidate indexing set to be assigned to a query document, the descriptors assigned to each of the neighbors were broken down into a set of main headings and pairs (i.e. a main heading associated with a single subheading). For this experiment, indications of major terms were ignored.

For example, the MeSH descriptor *Myocardium/cytology/metabolism would generate the main heading Myocardium and the two pairs Myocardium/cytology and Myocardium/metabolism.

3.2 Test Corpus

Both methods were tested on a corpus composed of a selection of the 49,863 citations entered into MEDLINE in January 2005. The 2006 version of MeSH was used for the indexing in these citations. About one fifth of the citations (10,161) are considered to be genetics-related, as determined by Journal Descriptor Indexing (Humphrey, 1999). Our test corpus was composed of genetics-related citations from which Entrez Gene IDs could be extracted – about 40% of the cases. The final test corpus size was 3,962. Appendix A shows a sample citation from the corpus.

3.3 Protocol

Figure 1 shows the setting of our experiment. Documents from the test corpus described above were processed to obtain both a generic and specific representation as described in section 2.2. The corresponding ten nearest neighbors were retrieved using the PRC and GRC algorithms. All the neighbors’ MeSH descriptors were pooled to form candidate indexing sets of descriptors that were evaluated using precision and recall measures. Precision was the number of candidate descriptors that were selected as indexing terms by NLM indexers (according to reference MEDLINE indexing) over the total number of candidate descriptors. Recall was the number of candidate descriptors that were selected as indexing terms by NLM indexers over the total number of indexing terms expected (according to reference MEDLINE indexing). For better comparison between the methods, we also computed F-measure giving equal weight to precision and recall - F1=2*PR/(P+R) and giving a higher weight to recall - F3=10*PR/(9P+R).

Four different categories of descriptors were considered in the evaluation:

MH: MeSH main headings (regardless of whether subheadings were attached in the reference indexing)

SH: stand-alone subheadings (regardless of the main heading(s) they were attached to in the reference indexing)

MH/SH: main heading/subheading pairs

DESC: MeSH descriptors, i.e. main headings and main heading/subheading pairs

Similarly, four different candidate indexing sets were considered: the indexing set resulting from PRC, the indexing set resulting from GRC, the indexing set resulting from the pooling of PRC and GRC sets and finally the indexing set resulting from the intersection of PRC and GRC indexing sets (common index terms).

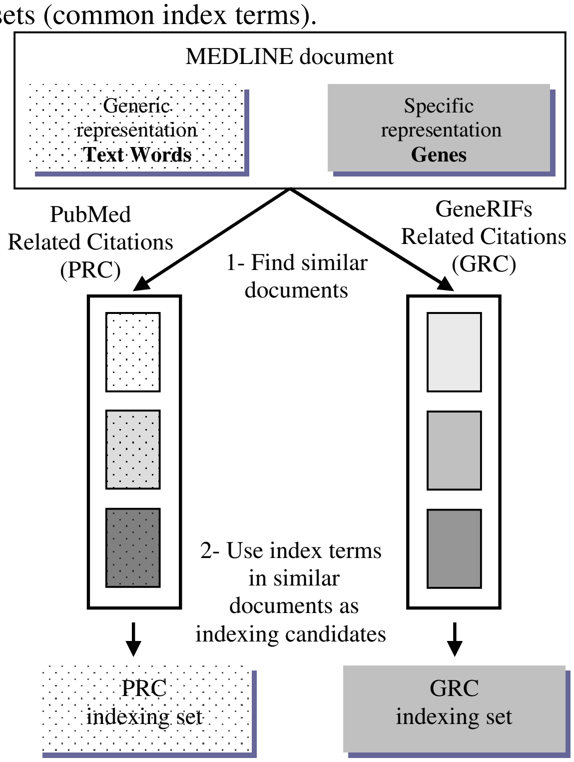

Figure 1: Producing candidate indexing sets with generic and domain-specific representations.

4 Results

Appendix B shows the indexing sets obtained from the GRC and PRC algorithms for a sample citation from the test corpus. Table 2 presents the results of our experiments. For each category of descriptors, the best performance was bolded. It can be observed that in general, the best precision and F1 scores are obtained with the common indexing set, the best recall is obtained with the pooling of indexing sets and the best F3 score is obtained with PRC algorithm, the pooling of indexing sets being a close second.

5 Discussion

5.1 Performance of the methods

As can be seen from the bolded figures in table 2, the best performance is obtained either from the PRC algorithm, or from a combination of PRC and GRC. When indexing methods are combined, it is usually expected that statistical methods will provide the best recall whereas linguistic methods will provide the best precision. Combining complementary methods is then expected to provide the best overall performance. In this context, it seems that the option of pooling the indexing sets should be retained for further experiments. The most significant result of this study is that the pooling of methods achieves a recall of 92% for stand-alone subheading retrieval. While the precision is only 19%, the selection of stand-alone subheadings offered by our methods is nearly exhaustive and it reduces by 70% the size of the list of allowable subheadings that could potentially be used. NLM indexers have declared this could prove very useful to enhance their indexing practice.

In order to qualify the added value of the specific description, we looked at the descriptors that were correctly recommended by GRC and not recommended by PRC. Check Tags (descriptors used to denote the species, age and gender of the subjects discussed in an article) seemed prominent, but only *Human* was significantly recommended correctly more often than it was recommended incorrectly (~2.2 times more correct than incorrect recommendations – 2,712 correct *vs.* 1,250 incorrect). No other descriptor could be identified as being consistently recommended either correctly or incorrectly.

For both methods, filtering the indexing sets according to the number of neighbors that lead to include the indexing terms results in an increase of precision and a loss of recall. The best trade-off (measured by F1) is obtained when indexing terms come from at least three neighbors (data not shown).

5.2 A scale of indexing performance

The problem with evaluating indexing is that, although inter-indexer variability is reduced when a controlled vocabulary is used, indexing is an open cognitive task for which there is no unique "right" solution.

Table 2: performance of the indexing methods on the four categories of descriptors

	SH				MH				SH/MH				DESC			
	P	*R*	*F1*	*F3*	*P*	*R*	*F1*	*F3*	*P*	*R*	*F1*	*F3*	*P*	*R*	*F1*	*F3*
GRC	21	72	32	58	8	49	14	32	3	23	6	14	6	38	10	25
PRC	27	88	41	**72**	13	61	22	**45**	8	56	15	**36**	11	59	18	**41**
Pool	19	**92**	32	67	9	**82**	16	44	5	**62**	9	29	7	**74**	13	38
Common	**36**	68	**47**	62	**22**	27	**24**	27	**18**	17	**17**	17	**21**	23	**22**	23

In practice, this means that there is no ideal unique set of descriptors to use for the indexing of a particular document. Therefore, when comparing an indexing set obtained automatically (e.g. here with the PRC or GRC methods) to a "gold standard" indexing set produced by a trained indexer (e.g. here, NLM indexers) the difference observed can be due to erroneous descriptors produced by the automatic methods. But it is also likely that the automatic methods will produce terms that are semantically close to what the human indexer selected or even relevant terms that the human indexer considered or forgot to select. While evaluation methods to assess the semantic similarity between indexing sets are investigated (Névéol et al. 2006), a consistency study by Funk et al. (1983) can shade some light on inter-indexer consistency in MEDLINE and what range of performance may be expected from automatic systems. In this study, Hooper's consistency (the average proportion of terms in agreement between two indexers) for stand-alone subheadings (SH) was 48.7%. It was 33.8% for pairs (MH/SH) and 48.2% for main headings (MH). In light of these figures, although no direct comparison with the results of our experiment is possible, the precision obtained from the common recommendations (especially for stand-alone subheadings, 36%) seems reasonably useful. Further more, when informally presenting the indexers sample recommendations obtained with these methods, they expressed their interest in the high recall as reviewing a larger selection of potentially useful terms might help them track important descriptors they may not have thought of using otherwise.

In comparison with other research, the results are also encouraging: the recall resulting from either PRC or pooling the indexing sets is significantly better than that obtained by Névéol et al. (2007) on a larger set of MEDLINE 2005 citations – 20% at best for main heading/subheading pairs with a dictionary-based method which consisted in extracting main heading and subheading separately from the citations (using MTI and string matching dictionary entries) before forming all the allowable pairs as recommendations.

5.3 Limitations of the experiment

In the specific description, the mapping between gene names and Entrez Gene IDs only takes human genes into account, which potentially limits the scope of the method, since many more organisms and their genes may be discussed in the literature. In some cases, this limitation can lead to confusion with other organisms. For example, the gene EPO "erythropoietin" is listed in Entrez Gene for 11 organisms including *Homo Sapiens*. With our current algorithm, this gene will be assumed to be a human gene. In the case of PMID 15213094 in our test corpus, the organism discussed in the paper was in fact *Mus Musculus* (common mouse). In this particular case, the check tag *Humans*, which was erroneous, could be found in the candidate indexing set. However, correct indexing terms could still be retrieved due to the fact that both the human and mouse gene share common functions.

Another limitation is the size of the test corpus, which was limited to less than 4,000 documents.

5.4 Mining the biomedical literature for gene-concept links

Other approaches to gene-keyword mapping exploit the links between genes and diseases or proteins as they are described either in the records of databases such as OMIM or more formally expressed as in the GeneRIF. Substantial work has addressed linking DNA microarray data to keywords in controlled vocabulary such as MeSH (Masys et al. 2001) or characterizing gene clusters with text words from the literature (Liu et al. 2004). However, no normalized “semantic fingerprinting” has been yet produced between controlled sets such as Entrez Gene and MeSH terms.

6 Conclusion and future work

In this paper, we applied a statistical method for indexing documents from the genetics literature. We presented two different document representations, one generic and one specific to the genetics domain. The results bear out our expectations that such statistical methods can also be used successfully to produce recommendations involving subheadings. Furthermore, they yield higher recall than other more linguistic-based methods. In terms of recall, the best results are obtained when the indexing sets from both the specific and generic representations are pooled.

In future work, we plan to refine the algorithm based on the specific method by expending its scope to other organisms than *Homo Sapiens* and to take the gene frequency in the title and abstract of documents into account for the representation. Then, we shall conduct further evaluations in order to observe the impact of these changes, and to verify that similar results can be obtained on a larger corpus.

Acknowledgments

This research was supported in part by an appointment of A. Névéol to the Lister Hill Center Fellows Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education. The authors would like to thank Halil Kilicoglu for his help with obtaining Entrez Gene IDs from the ABgene output. We also thank Susanne Humphrey and Sonya Shooshan for their insightful comments on the preparation and editing of this manuscript.

References

Alan R. Aronson, Olivier Bodenreider, H. Florence Chang, Susanne M. Humphrey, James G. Mork, Stuart J. Nelson, Thomas C. Rindflesch and W. John Wilbur. 2000. The NLM Indexing Initiative. *Proceedings of the Annual American Medical Informatics Association Symposium.* (AMIA 2000): 17-21.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the Annual AMIA Symposium.* (AMIA 2001):17-21.

Alan R. Aronson, James G. Mork, Cliff W. Gay, Susanne M. Humphrey and William J. Rogers. 2004. The NLM Indexing Initiative's Medical Text Indexer. *Proceedings of Medinfo 2004*: 268-72.

Kate Byrne and Ewan Klein. 2003. Image Retrieval using Natural Language and Content-Based techniques. In Arjen P. de Vries, ed. *Proceedings of the 4th Dutch-Belgian Information Retrieval Workshop (DIR 2003)*:57-62.

Scott Deerwester, Susan Dumais, Georges Furnas, Thomas Landauer and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of American Society for Information Science*, 6 (41):391-407.

Mark E. Funk, Carolyn A. Reid and Leon S. McGoogan. 1983. Indexing consistency in MEDLINE. *Bull. Med. Libr. Assoc.* 71(2):176-183.

Susanne M. Humphrey. 1999. Automatic indexing of documents from journal descriptors: a preliminary investigation. *J Am Soc Inf Sci Technol*. 50(8):661-674

Christian Jacquemin and Evelyne Tzoukermann. 1999. NLP for term variant extraction: Synergy of morphology, lexicon, and syntax. In T. Strzalkowski (Ed.), *Natural language information retrieval* (p. 25-74). Boston, MA: Kluwer.

Won Kim, Alan R. Aronson and W. John Wilbur. 2001. Automatic MeSH term assignment and quality assessment. *Proceedings of the Annual AMIA Symposium*: 319-23.

Ying Liu, Martin Brandon, Shamkant Navathe, Ray Dingledine and Brian J. Ciliax. 2004. Text mining functional keywords associated with genes. *Proceedings of MEDINFO 2004*: 292-296

Daniel R. Masys, John B. Welsh, J. Lynn Fink, Michael Gribskov, Igor Klacansky and Jacques Corbeil. 2001. Use of keyword hierarchies to interpret gene expression patterns. In: *Bioinformatics* 17(4):319-326

Adeline Nazarenko and Touria Ait El Mekki 2005. Building back-of-the-book indexes. In: *Terminology* 11(1):199–224

Aurélie Névéol, Kelly Zeng, Olivier Bodenreider. 2006. Besides precision & recall: Exploring alternative approaches to evaluating an automatic indexing tool for MEDLINE. *Proceedings of the Annual AMIA Symposium*: 589-93.

Aurélie Névéol, Sonya E. Shooshan, Susanne M. Humphrey, Thomas C. Rindflesch and Alan R Aronson. 2007. Multiple approaches to fine-grained indexing of the biomedical literature. *Proceedings of the 12th Pacific Symposium on Biocomputing*. 12:292-303

James Rhodes, Stephen Boyer, Jeffrey Kreulen, Ying Chen, Patricia Ordonez. 2007. Mining Patents Using Molecular Similarity Search. *Proceedings of the 12th Pacific Symposium on Biocomputing*. 12:304-315

Thomas C. Rindflesch and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform.* **36(6)**, 462-77

Gerald Salton. 1989. *Automatic text processing : The transformation, analysis, and retrieval of information by computer*. Reading, MA : Addison-Wesley.

Karen Sparck-Jones, Steve Walker and Stephen E. Robertson. 2000. A probalistic model of information retrieval: development and comparative experiments (part 1). *Information Processing and Management*, 36(3):779-808.

Lorraine Tanabe and W. John Wilbur. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics*. 2002 Aug;18(8):1124-32.

Appendix A: Title, abstract and reference indexing set for a sample citation

PubMed ID	15645653
Title	Identification of two novel missense mutations (p.R1221C and p.R1357W) in the ABCC6 (MRP6) gene in a Japanese patient with pseudoxanthoma elasticum (PXE).
Abstract	Pseudoxanthoma elasticum (PXE) is a rare, inherited, systemic disease of elastic tissue that in particular affects the skin, eyes, and cardiovascular system. Recently, the ABCC6 (MRP6) gene was found to cause PXE. A defective type of ABCC6 gene (16pl3.1) was determined in two Japanese patients with PXE. In order to determine whether these patients have a defect in ABCC6 gene, we examined each of 31 exons and flanking intron sequences by PCR methods (SSCP screening and direct sequencing). We found two novel missense variants in exon 26 and 29 in a compound heterozygous state in the first patient. One is a missense mutation (c.3661C>T; p.R1221C) in exon 26 and the other is a missense mutation (c.4069C>T; p.R1357W) in exon 29. These mutations have not been detected in our control panel of 200 alleles. To our knowledge, this is the first report of mutation identification in the ABCC6 gene in Japanese PXE patients. The second patient was homozygous for 2542_2543delG in ABCC6 gene and heterozygous for 6 kb deletion of LDL-R gene. This case is the first report of a genetically confirmed case of double mutations both in PXE and FH loci.
MeSH reference indexing set	Adult Aged Female Humans Japan Multidrug Resistance-Associated Proteins/*genetics *Mutation, Missense Pedigree Pseudoxanthoma Elasticum/*genetics

Appendix B: Sample indexing sets obtained from the GRC and PRC algorithms for a sample citation

PubMed ID	15645653
GRC indexing set* (top 15 terms)	<u>Humans</u> (10) <u>Multidrug Resistance-Associated Proteins</u> (9) Mutation (8) Male (7) <u>Female</u> (7) <u>Multidrug Resistance-Associated Proteins/genetics</u> (7) <u>Pseudoxanthoma Elasticum</u> (6) <u>Pseudoxanthoma Elasticum/genetics</u> (6) <u>Pedigree</u> (5) Exons (4) DNA Mutational Analysis (4) Mutation/genetics (4) <u>Adult</u> (4) Introns (3) *Aged* (3)
PRC indexing set* (top 15 terms)	<u>Multidrug Resistance-Associated Proteins</u> (10) <u>Multidrug Resistance-Associated Proteins /genetics</u> (10) <u>Pseudoxanthoma Elasticum</u> (10) <u>Pseudoxanthoma Elasticum/genetics</u> (10) Mutation (7) DNA Mutational Analysis (6) <u>Pedigree</u> (5) Genotype (4) Polymorphism, Genetic (4) Alleles (4) Mutation/genetics (3) Haplotypes (3) Models, Genetic (3) Gene Deletion (3) Exons (3)

* Terms appearing in the reference set are underlined; the number of neighbors – out of the 10 nearest neighbors – labeled with each term is shown between brackets after the term.