BIOINFORMATICS

Vol. 20 no. 16 2004, pages 2597–2604
doi:10.1093/bioinformatics/bth291

Distribution of information in biomedical abstracts and full-text publications

M. J. Schuemie, M. Weeber, B. J. A. Schijvenaars, E. M. van Mulligen, C. C. van der Eijk, R. Jelier, B. Mons and J. A. Kors*

Department of Medical Informatics, Erasmus University Medical Center Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands

Received on November 19, 2003; revised on February 27, 2004; accepted on April 23, 2004
Advance Access publication May 6, 2004

ABSTRACT

Motivation: Full-text documents potentially hold more information than their abstracts, but require more resources for processing. We investigated the added value of full text over abstracts in terms of information content and occurrences of gene symbol—gene name combinations that can resolve gene-symbol ambiguity.

Results: We analyzed a set of 3902 biomedical full-text articles. Different keyword measures indicate that information density is highest in abstracts, but that the information coverage in full texts is much greater than in abstracts. Analysis of five different standard sections of articles shows that the highest information coverage is located in the results section. Still, 30–40% of the information mentioned in each section is unique to that section. Only 30% of the gene symbols in the abstract are accompanied by their corresponding names, and a further 8% of the gene names are found in the full text. In the full text, only 18% of the gene symbols are accompanied by their gene names.

Contact: m.schuemie@erasmusmc.nl

INTRODUCTION

The surge of scientific literature in the biomedical domain has made it hardly possible for researchers and medical professionals to keep track of developments in their own field of interest, let alone any information from related fields. Recently, many efforts to develop better information retrieval and extraction techniques to assist users in coping with this information overload have been suggested.

Traditionally, these techniques use the abstracts of papers, mostly due to wide availability of abstracts in databases such as MEDLINE (http://www.pubmed.gov). More recently however, full-text documents are more available, e.g. due to initiatives, such as BioMed Central (http://www.biomedcentral.com), Public Library of Science (http://www.plos.org) and PubMed Central (http://www.pubmedcentral.nih.gov).

*To whom correspondence should be addressed.

Because full-text documents are currently more difficult to obtain due to copyright protection, and are by definition longer, requiring more computing and storage, mining full-text documents is less practical than mining abstracts. When looking at the identification of gene names the problem is compounded. Identification of gene names in full text is more prone to error as papers mention chemical and biological entities other than genes with names similar to genes, and information can be contained in tables and figures that are difficult to process (Tanabe and Wilbur, 2002). Additionally, abstracts are assumed to contain the information most relevant to the paper, therefore having a higher information density than full text. Contrastingly, the full text generally contains more information, but this could be more dispersed. Therefore, the question arises how the information content of full-text documents compares to that of abstracts.

Little research has been done to evaluate the beneficial value of full-text documents compared with that of abstracts. Friedman *et al*. (2001) tested a system for the extraction of molecular pathways on one article and found that of the 19 unique molecular interactions mentioned in the text, only 7 were found in the abstract. Yu *et al*. (2002) used both abstracts and full-text documents to retrieve synonyms of genes and proteins, and found more synonyms, with a higher precision in the full text than the abstract.

Shah *et al*. (2003) performed a more systematic comparison of abstracts and full-text in *Nature Genetics*. In a set of 104 full-text articles that contained all the five standard sections Abstract, Introduction, Methods, Results and Discussion, they searched on keywords characterizing the text, assessing the keyword frequency in each section. They showed that the highest frequency of keywords occurred in the abstract. Furthermore, the content of the different sections was highly heterogeneous. In addition, Shah *et al*. (2003) investigated the appearance of a limited list of gene names and found that the abstract and introduction have the highest frequency of gene names. Shah *et al*. (2003) selected keywords by choosing single-word nouns that have a high K-value. This K-value was calculated using μ_{Iw}, a measure of the degree of inclusion

of noun w_j into noun $w_i \cdot \mu_{Iw}(w_i, w_j) = |W_i \cap W_j|/|W_i|$, where $|W_i \cap W_j|$ is the number of times that w_i and w_j appear together in a sentence and $|W_i|$ is the number of times w_i appears in the text. The K-value for a word w_i was then defined as

$$K_i = \sum_{j \neq i} \mu_{Iw}(w_i, w_j) = \left(\sum_{i \neq j} |W_i \cap W_j| \right) \Big/ |W_i|. \quad (1)$$

The K-value for a word was normalized to the maximum value found for K in that section. Keeping in mind that the nominator in the right-hand side of the above equation is equal to the number of times word i appears together with other words in a sentence, and the denominator is the number of sentences in which the word i appears, the highest K-values were assigned to words that appear in the, on an average, longest sentences (measured in nouns) in this section. However, it is unclear why words with a high K-value (i.e. words in relatively long sentences) should be preferentially considered keywords. Therefore, a different choice of keywords is used in the current study.

Given the paucity of results hitherto, the goal of this study is to assess the informational content of full-text documents as compared with abstracts, with a focus on medical information, and in particular information relating to genes. Additionally, we aim to determine how the information content of the document is distributed over different sections to identify parts of documents that are worth more attention when extracting information. We seek to improve on the research by Shah *et al*. (2003) by using more methodologically sound measures, by including both single and multiple word terms and a more extensive list of gene names, and by using a larger test corpus. Finally, because ambiguity of gene symbols is a recognized problem in information extraction (Pustejovsky *et al*., 2001), we investigate the presence of full-length gene names matching the gene-symbols, which would render the disambiguation problem trivial.

METHODS

Document set

We used a set of 1275 full-text publications from *Nature Genetics* (NG), from June 1998 (volume 19, issue 2) to November 2001 (volume 23, issue 3)[1], and all 2754 full-text publications from BioMed Central (BMC) (retrieved on 8 September 2003) containing 89 different journals. These included research articles as well as letters, news and view articles. Of these articles, 127 (3.2%) were not indexed in MEDLINE and were discarded because they mostly included letters and corrections with little relevance to the field, resulting in a test set containing 3902 documents.

[1]The dataset used by Shah *et al*. (2003) was a subset of this dataset.

Keyword identification

For the purpose of this study we assume that information in the text is represented by keywords, i.e. those words in the text that describe 'what the text is about'. To ensure that the keywords were relevant to the medical domain, and to simplify matters, a first selection was performed by identifying the text words that match the terms in a biomedical thesaurus. This was executed using the Collexis® engine (van Mulligen *et al*., 2000) (http://www.collexis.com), which normalizes text (i.e. reduces plural to singular form and upper case to lower case) and matches terms, possibly consisting of multiple words, to entries in the thesaurus. If the thesaurus terms appear literally in the text, or with small morphological variations, they are effectively always recognized by the system.

The thesaurus was MeSH 2002 (Medical Subject Headings, http://www.nlm.nih.gov/mesh/), the same thesaurus that is used by indexers at the National Library of Medicine to attach a list of terms to publications in the MEDLINE database. These manually attached MeSH terms will subsequently be called the 'MeSH headings'.

We used five different techniques to identify keywords. The first three techniques were based on MeSH terms:

(1) *MeSH headings*: These are the MeSH terms manually attached to a publication in MEDLINE by the human indexer. However, headings falling under the category Miscellaneous in the MeSH thesaurus were removed from the headings list because they were unlikely to appear literally in the paper (e.g. terms such as 'Support, U.S. gov.', 'Historical Biography (PT)' or 'Drug Evaluation, FDA Phase II').

(2) *Exploded MeSH headings*: These are MeSH headings extended with their children as defined in the thesaurus. For instance, if 'Parasitic Disease' was defined as a MeSH heading, then 'Malaria' would also be identified as a keyword.

(3) *TF*IDF* (*Term Frequency * Inverse Document Frequency*): This is a term relevance score commonly used in information retrieval. MeSH terms with a higher TF*IDF score are considered to be more relevant keywords than MeSH terms with a lower TF*IDF score. TF is the number of times the MeSH term appears in the document, and IDF is a measure for the uniqueness of the term in the whole document collection. We used $\mathrm{IDF} = {}^2\log(N/n)$, with N the total number of documents in the collection and n the number of documents containing the MeSH term.

Techniques 1 and 2 utilized the terms assigned by human indexers and were expected to yield the most informative keywords since humans are still the only entities capable of really understanding a text. However, because humans make errors and are subjective, technique 3 was included as a more objective approach.

Additionally, we used a self-constructed thesaurus of human gene names and symbols extracted from five genetic databases: GDB (http://www.gdb.org), Genew (http://www.gene.ucl.ac.uk/nomenclature), Locuslink (http://www.ncbi.nlm.nih.gov/locuslink), OMIM (http://www.ncbi.nlm.nih.gov/Omim) and SwissProt (http://us.expasy.org/sprot/). The combined thesaurus contained information regarding 25 004 genes, which were identified by 84 448 terms (gene names and symbols, including aliases). We included a fourth technique, using the aforementioned gene thesaurus, to find gene names and symbols in the text.

(4) *Gene terms*: Many gene symbols also have non-gene meanings (e.g. ESR, which can mean Estrogen Receptor 1 or Electronic Spin Resonance), or map to more than one gene. Of the 84 448 terms in our thesaurus, 3375 mapped to more than one gene. In an attempt to reduce the number of ambiguous terms, we required gene terms to contain at least one letter and one digit (where the first character must be a letter), or at least one space (i.e. gene names consisting of multiple words). A total of 66 806 gene terms conformed to these requirements[2]. A total of 2014 terms still map to more than one gene, however, we assume that the number of terms with non-gene meanings is reduced even further because the remaining terms are often typical gene names.

To gain more qualitative insight into the diversity of the content within sections, we also use a fifth technique:

(5) *MeSH terms per semantic type*: The MeSH hierarchy classifies terms into different semantic classes, their so-called semantic types. We established for every MeSH term that appears in the text its corresponding semantic type. To reduce complexity, we focused on three important categories within biomedical research: Organisms, Diseases and Chemicals and Drugs. Additionally, we included the genes from our thesaurus as a fourth semantic type.

Information measures

Two important concepts for describing the information content of a piece of text are the information density and the information coverage of that text. Information density refers to the average amount of information per unit of text (e.g. per word or sentence). Information coverage refers to the total amount of information described in a piece of text. We defined several specific information density and coverage measures, listed below. The information coverage measures were calculated in terms of the fraction of the total information in a paper that was described in a part of that paper.

(1) *Heading Density (HD)*: The number of instances of MeSH headings found in the text divided by the number of words.

(2) *Exploded Heading Density (XHD)*: Similar to HD, but included the children of the original MeSH headings.

(3) *Weighted MeSH Term (WMT) density*: The sum of the TF*IDF scores of the MeSH terms in the text, divided by the number of words. This measure can be viewed as a weighed density measure, since it took into account the weight of each term.

(4) *Gene Density (GD)*: The number of instances of gene names found divided by the number of words.

(5) *WMT fraction*: The sum of the TF*IDF scores of the MeSH terms mentioned at least once in a section, divided by the sum of the TF*IDF scores of the MeSH terms mentioned at least once in the entire document.

(6) *Heading Fraction (HF)*: The fraction of the MeSH headings encountered at least once in the text.

(7) *Exploded Heading Fraction (XHF)*: The fraction of the MeSH headings mentioned at least once, or of which one of its children was mentioned.

(8) *Gene Fraction (GF)*: The number of unique genes mentioned, either by symbol or name, in a section divided by the number of unique genes mentioned in the entire article.

(9) *Exploded Heading Uniqueness (XHU)*: The fraction of the MeSH headings, including children, mentioned in a section that was not mentioned in any other section.

(10) *Gene Uniqueness (GU)*: The fraction of genes mentioned in a section that were not mentioned in any other section.

(11) *Semantic Type Density (STD)*: The number of terms belonging to a specific semantic type in a section, divided by the total number of words in that section.

(12) *Semantic Type Fraction (STF)*: The number of terms of a specific semantic type mentioned at least once in a section, divided by the total number of terms of that type mentioned at least once in the entire document.

Section detection

Standard sections were detected by identifying section headings that contained ‘abstract’, ‘background’, ‘introduction’, ‘method’, ‘result’, or ‘discussion’. The abstract section was often identified by a specific mark-up tag in the source documents.

Disambiguation information

If gene symbols are directly followed by corresponding gene names, disambiguation of the symbols is straightforward. To assess how often this was the case in our document set, we

[2]Shah used a set of 539 genes whose names are comprised of three letters followed by one digit.

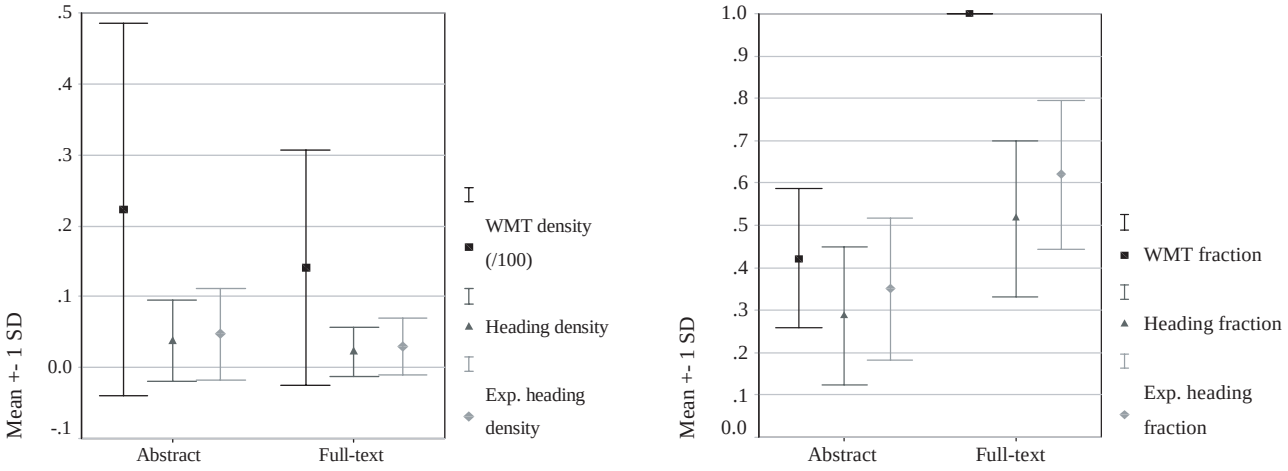

Fig. 1. (**a**) and (**b**) Information distribution between abstract and full text, for all 1834 papers (807 from NG, 1027 from BMC) with a MEDLINE abstract and MeSH headings. (WMT density is divided by 100 for visual purposes.)

determined for each gene symbol whether it was followed by a gene name, using the abbreviation expansion algorithm of Schwartz and Hearst (2003). In this analysis, we used all symbols from our gene thesaurus for gene–symbol identification, including those symbols that we deemed to be too ambiguous to be used for keyword identification, since we wanted to investigate the possibility of resolving this ambiguity.

Data analysis

Differences between groups and between dependent measurements were assessed with ANOVA tests. All statistical analyses were performed using SPSS 11.

RESULTS

Document structures

Of the 3902 analyzed NG and BMC publications, 2458 (63.0%) articles contained all five standard sections: Abstract (A), Introduction (I), Methods (M), Results (R) and Discussion (D). These articles were significantly longer than the other articles ($P < 0.001$), with a mean (SD) number of words of 3624 (1688) versus 1960 (1535) for the rest of the collection.

MEDLINE contained an abstract for 3500 (89.7%) publications, and a first comparison showed that these abstracts are always identical to the abstracts in the full text. A total of 1735 (44.5%) MEDLINE abstracts had no MeSH headings attached, and most of these were recent papers from the BioMed Central dataset. An average of 16.0 MeSH headings was assigned to the MEDLINE-indexed papers.

Abstract versus full text

Figure 1 shows the average scores of the information metrics for the abstract and for the full text of those documents with an abstract in MEDLINE. As was expected, the keyword density in the abstract is higher than in the full text, but the coverage of assigned MeSH headings in the full-text was significantly larger. (Note that the WMT fraction for the full text is 1 by definition). An ANOVA for repeated measurements showed that all effects are highly significant ($P < 0.001$).

Interestingly, not all attached MeSH headings are actually found in the text, even when children are included. This is not a problem for the information measures if the missed headings are missing in equal degrees from the abstract and the full text. To test this assumption, we determined whether the proportion between the exploded heading density in the abstract and the exploded heading density in the full text is different for documents with varying percentages of retrieved MeSH headings. The difference was not statistically significant[3], suggesting that the effect of missed MeSH headings is equally large in abstracts and full texts.

The average number of unique gene names found in the set of abstracts was 0.61 versus 2.35 in the full texts. Again, the difference was highly significant ($P < 0.001$).

Standard document sections

Figure 2 shows the distribution of information over the five standard sections for the 1050 documents that contained all these sections and had MeSH headings assigned. The keyword density was highest in the Abstract and lowest in the Methods and Discussion sections, whilst the keyword fraction was highest in the Results section. An ANOVA for repeated measurements indicated that all effects were highly significant ($P < 0.001$), except for HF ($P = 0.065$) and XHF ($P = 0.165$). A *post hoc* least significant distance (LSD)

[3] Pearson Correlation between ($XHD_{abstract}/XHD_{full\ text}$) and $XHF_{full\ text}$ is -0.037, $P = 0.115$.

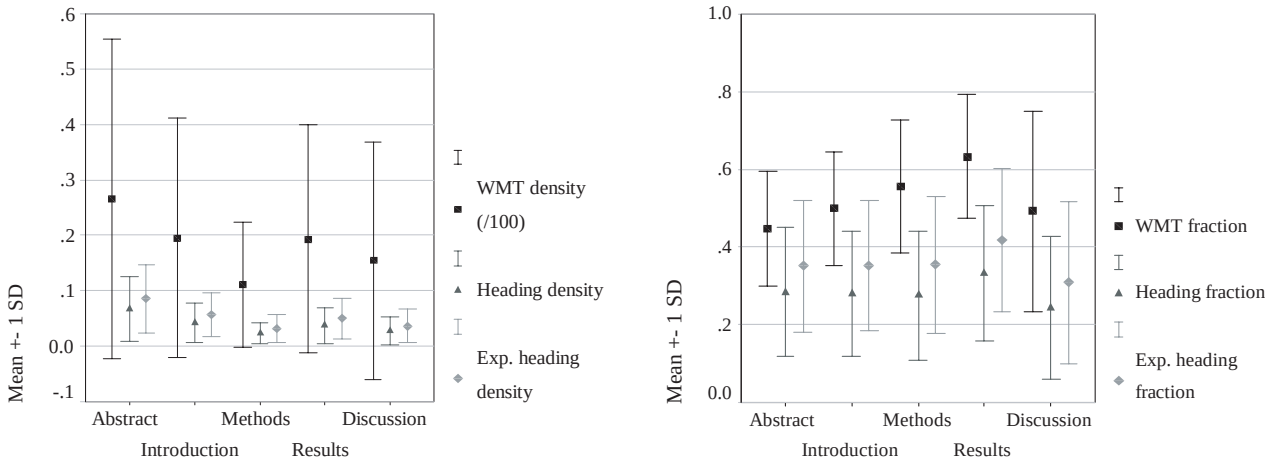

Fig. 2. (**a**) and (**b**) Information distribution over the different standard sections. (WMT Density is divided by 100 for visual purposes.) Based on all 1050 papers (114 from NG, 936 from BMC) to which MeSH headings have been assigned and containing all five standard sections.

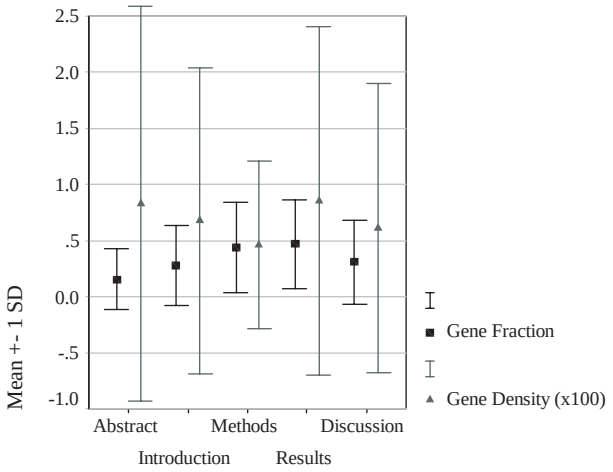

Fig. 3. Distribution of gene names over the different sections (Gene Density is multiplied by 100 for visual purposes). Based on all 1165 papers (106 from NG, 1059 from BMC) containing all five standard sections in which at least one gene name was found.

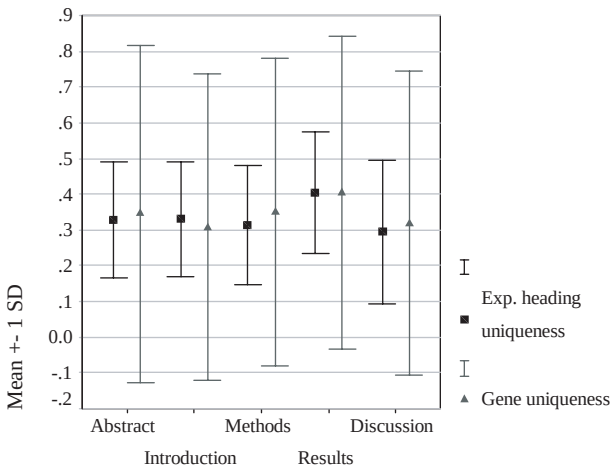

Fig. 4. Fraction of the MeSH headings (including children) and genes mentioned in a section that are not mentioned in any other section. Based on the 599 papers (106 from NG, 493 from BMC) to which MeSH headings have been assigned, with one or more occurrence of a gene name, and containing all five standard sections.

pairwise comparison between sections showed that most[4] of the differences between these sections were significant at the 0.05 level.

Figure 3 shows the distribution of gene names for the 1165 documents with at least one gene name and containing all five sections, indicating that the highest gene fraction was found in the Methods and Results sections. An ANOVA for repeated measurements indicated that all differences were highly significant ($P < 0.001$). A *post hoc* LSD pairwise comparison showed that most[5] of the differences between sections were also significant at the 0.05 level. Note that the actual gene name density could be higher than shown, as we ignored gene symbols that could be ambiguous, and because our thesaurus may have been incomplete.

Figure 4 shows the number of Exploded Headings and gene names that were uniquely found in a single section. Neither Exploded Headings Uniqueness nor Gene Uniqueness differed significantly between sections. A *post hoc* LSD pairwise comparison showed however that most[6] differences were significant at the 0.05 level.

[4] Not significant for WMT density, between I and R; for WMT fraction, between I and D; for HF, between A and I, between A and M and between I and M; for XHF, between A and I, between A and M and between I and M.

[5] Not significant for GF, between M and R; for GD, between A and R.

[6] Not significant for XHU, between A and I; for GU, between A and M, between A and D, between I and D, between M and D.

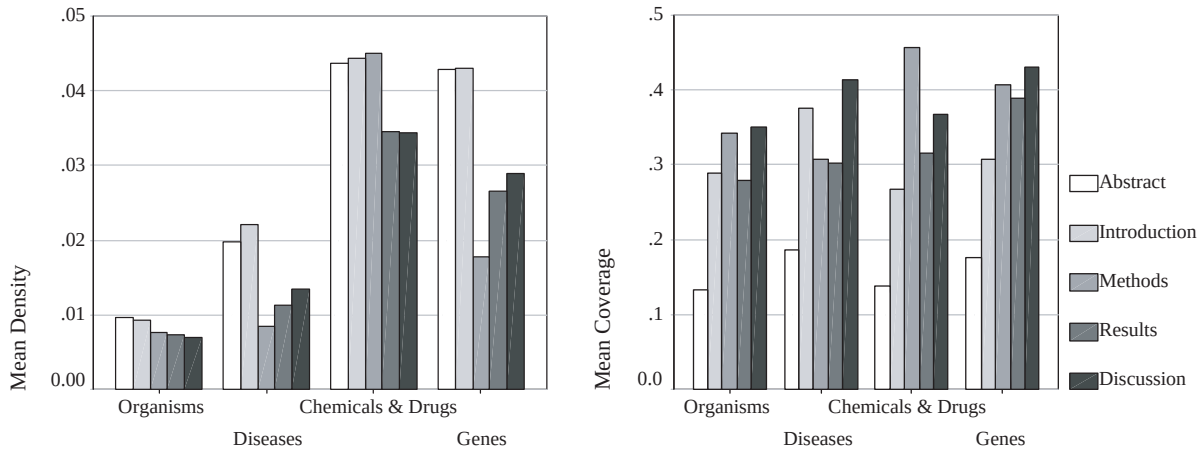

Fig. 5. (**a**) and (**b**) Information distribution of the four selected semantic types over the different standard sections. Based on the 2458 papers (114 from NG, 2344 from BMC) containing all five standard sections.

Table 1. Average number of gene symbols found in the abstract and full text, and the percentage for which a matching Long-form (LF) gene name was found

	Gene symbol found	Long-form found (%)	LF found in full-text only (%)	LF found only in abbreviation section (%)
All symbols				
Abstract	1.35	30	8	
Full text	9.69	18		2
Ambiguous symbols				
Abstract	0.12	53	12	
Full text	1.08	27		3

Based on all 3500 papers (809 from NG, 2591 from BMC) for which a MEDLINE abstract was found. The rightmost column shows, for the 610 papers (0 from NG, 610 from BMC) in which an abbreviations section was found, the percentage of short-forms in the full text for which a matching long-form was found only in the abbreviations section.

Figure 5a and b show the Semantic Type Density and Semantic Type Coverage for the four selected semantic types. It can be seen that the semantic types Diseases and Genes were found in relatively low density in the Methods section. Pairwise comparison with the other sections showed these differences to be highly significant ($P < 0.001$). In contrast, the widest variety of Chemicals and Drugs was discussed in the Methods section. Pairwise comparison between coverage in sections showed this difference also to be highly significant ($P < 0.001$).

Gene symbol—gene name alignment

Table 1 shows the average number of gene symbols that were found in abstract or full text (including the Abstract), and the percentage of these gene symbols that had a corresponding gene name. A matching gene name was found for only a small percentage of gene symbols in the abstract, even when searching the full text. For gene symbols found in the full text, the problem was even worse. The addition of an abbreviations section, found in 17% of the publications, increased the number of disambiguated gene symbols by only a small percentage.

We repeated our analysis for the subset of gene symbols that were assigned to more than one gene in the thesaurus ($n = 609$). We hypothesized that authors would be more prone to include the gene names of abbreviations that were ambiguous in their own research field. Although the results confirm this hypothesis, still only half of the gene symbols in the abstract were accompanied by a gene name.

DISCUSSION

Our results show that the information density is highest in abstracts. In contrast, the coverage of information in terms of biomedical and gene keywords is substantially higher in full text when compared with abstracts. Almost twice as many biomedical concepts were mentioned in the full text, and on average nearly four times as many genes. When looking at individual sections, the highest information coverage was detected in the Results section, whereas the density was lowest in the Methods section. This, therefore, would argue for placing particular emphasis on mining information-rich sections such as the Results, although one should keep in mind that a substantial part of the information mentioned in any section appears

to be unique to that section. The uniqueness of gene names in a section, however, could also be due to the fact that these genes were not related to the main topic of the article, but for instance to the method used as described in a Methods section.

The investigation into occurrences of relevant semantic types showed that the Methods section was richest in information on Chemicals and Drugs, whilst Diseases and Genes were mentioned less frequently in the Methods section than in other sections. Since named-entity extraction algorithms are reported to have difficulties in distinguishing between gene names and chemical entities (Tanabe and Wilbur, 2002), not applying these algorithms to the Methods section might improve their performance.

Remarkably, only ∼62% of the MeSH headings manually assigned to a paper could actually be retrieved from the full text, even when the children of those headings were included. In contrast Shah *et al*. (2003) reported that they could find ∼72% (an average of 4.91 of 6.80 headings) of all MeSH headings assigned to the documents in their collection, even without children. The difference between these results can be explained by taking into consideration the fact that Shah only used single word terms, thus ignoring the 60.6% of all MeSH headings that consist of more than one word which are often more specific and less likely to be found.

The reason that not all terms were found could be that the granularity of MeSH is insufficient for this task. For instance, there were papers with the heading 'Nuclear Proteins' that did not have the explicit term 'nuclear protein' in the text, but specific nuclear proteins were mentioned in several instances, none of which however were part of MeSH. Because the exploded heading density in the abstract relative to the full text is not different for documents with a low or high percentage of retrieved MeSH headings, we assume that this effect is equal for the different sections of a paper and that the MeSH headings can therefore be used in our measures for identification of relevant information.

Furthermore, the results of our different keyword measures were consistent and can therefore be assumed to be reliable. Even though we used different keyword measures, extended our scope to include multiple word terms, and used a substantially larger number of gene names and documents than Shah *et al*. (2003), our results concur on several points, most importantly on the fact that the highest information density in terms of generic keywords and genes is found in the abstract. However, the most significant difference was between the information coverage, which according to Shah's results was highest in the Introduction and Methods and lowest in the Results section, whilst our results showed it to be highest in the Results section. This difference is most likely due to differences between the keyword measure used by Shah and our measures.

Our inventory of gene symbol—gene name combinations showed that many gene symbols do not have a corresponding gene name in the abstract nor the full text. Even gene symbols that were known to be ambiguous in the biomedical domain were often not accompanied by their long-form. However, one should keep in mind that not all gene symbol—gene name combinations were found by the algorithm. But even when taking into account the 82% recall level reported by Schwartz and Hearst (2003), many ambiguous symbols remain. Therefore, homonymy in biomedical publications still remains a very serious problem to be considered in text-mining efforts and additional methods will have to be developed to disambiguate homonymous gene symbols in the literature and other information resources.

CONCLUSIONS

The overall density of information appears to be lower in full text when compared with abstracts. However, when using an information extraction tool that is capable of dismissing irrelevant data, this should not be a problem. The fact that the information content of full text was much greater strongly argues for using full text instead of abstracts when extracting information from literature. Within a single article, there were sections that contained more information than others, but a substantial part of the information in any section appeared to be unique to that section. Extraction from more text leads to more information. Conclusion: 'More is better.'

The restriction of using only abstracts in information retrieval and extraction, frequently triggered by performance issues, introduces the danger of serious information-loss. Processing the corresponding full text of millions of MEDLINE records, and other text repositories is not a trivial task. However, it would be desirable or even a future requirement for optimal literature-based knowledge extraction and discovery tools.

The ambiguity in MEDLINE abstracts relating to gene and protein symbols cannot be resolved automatically by searching for corresponding gene names in the abstract or full text. In fact, since more potential gene names appeared in the full text, the ambiguity problem here was even more aggravated. For further disambiguation, other knowledge sources, such as the context in which ambiguous symbols appear, will have to be employed.

ACKNOWLEDGEMENT

This research was supported in part by the European Commission under the ORIEL project, contract no. IST-2001-32688.

REFERENCES

Friedman,C., Kra,P., Yu,H., Krauthammer,M. and Rzhetsky,A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**(suppl. 1), 74–82.

Schwartz,A.S. and Hearst,M.A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. *Proceedings of the Pacific Symposium on Biocomputing* (PSB 2003), Kauai, pp. 451–263. http://helix-web.stanford.edu/psb03/

Shah,P.K., Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2003) Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, **4**, 20.

Tanabe,L.K. and Wilbur,W.J. (2002) Tagging gene and protein names in full text articles. *Bioinformatics*, **18**, 1124–1132.

van Mulligen,E.M., Diwersy,M., Schmidt,M., Buurman,H. and Mons,B. (2000) Facilitating networks of information. Proceedings of the American Medical Informatics Association Symposium. Philadelphia, PA. Harley and Belfus. Inc, pp. 868–872.

Yu,H., Hatzivassiloglou,V., Friedman,C., Rzhetsky,A. and Wilbur,W.J. (2002) Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. *Proceedings of the AMIA Symposium 2002*, pp. 919–923.