BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Information extraction from full text scientific articles: Where are the keywords?

Parantu K Shah[1,2], Carolina Perez-Iratxeta[1,2], Peer Bork*[1,2] and Miguel A Andrade[1,2,3]

Address: [1]Biocomputing, European Molecular Biology Laboratory, Heidelberg, Germany, [2]Department of Bioinformatics, Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany and [3]Present address: Bioinformatics group, Ottawa Health Research Institute, Ottawa, Canada

Email: Parantu K Shah - shah@embl.de; Carolina Perez-Iratxeta - cperez@embl.de; Peer Bork* - bork@embl.de; Miguel A Andrade - andrade@embl.de

* Corresponding author

Published: 29 May 2003

BMC Bioinformatics 2003, **4**:20

Received: 5 March 2003
Accepted: 29 May 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/20

Abstract

Background: To date, many of the methods for information extraction of biological information from scientific articles are restricted to the abstract of the article. However, full text articles in electronic version, which offer larger sources of data, are currently available. Several questions arise as to whether the effort of scanning full text articles is worthy, or whether the information that can be extracted from the different sections of an article can be relevant.

Results: In this work we addressed those questions showing that the keyword content of the different sections of a standard scientific article (abstract, introduction, methods, results, and discussion) is very heterogeneous.

Conclusions: Although the abstract contains the best ratio of keywords per total of words, other sections of the article may be a better source of biologically relevant data.

Background

Most applications of information extraction from the scientific medical bibliography use the Abstract of the publication (for review see for example [1–3]). In the context of information extraction in molecular biology it is usually understood that the information to be extracted from an article are words regarding biological concepts that could synthesize the main points of the article (keywords). Therefore the Abstract of a paper is a good target for information extraction because by definition an abstract synthesizes the content of the article. Moreover, abstracts are available in public databases. However, nowadays most journals are also available in electronic version, and thus full text articles can be used for information extraction.

It is obvious that the full text of an article contains more information than its Abstract. However, in approaching full text analysis several problems must be tackled. On the one hand, the storage of full text articles requires more disk space and the analysis needs more computational capacity. On the other hand, an Abstract, as a summary, contains a high frequency of relevant terms (keywords), but this may not be the case of the rest of the article.

Other questions regard the quality of the information carried by different sections of an article. First of all, is the information in full text organized enough so that keywords can be extracted? Secondly, different biological concepts (for example, gene and protein names, tissue names, organisms, experimental conditions, etc.) may be

located in different parts of the article. Or it could be that a word has a different meaning depending on the section where it is located (the word has a context dependent meaning). For example, regarding gene names, those found in the Methods section may refer mostly to analytical tools rather than being relevant to the biological phenomenology described in the whole article. In summary, it would be good to quantify and qualify the information in a full text article before embarking in large scale extraction of particular items of information.

With this goal in mind, we analyzed in this work the kind of information that is attached to different parts of an article and we tried to quantify how much information can be found in each section of an article. This should help to state some guidelines for researchers attempting to extract particular keywords (words synthesizing the content of the article) from full text articles.

Results

Text Corpus

As previously stated, the major objective of this work was to compare the information defined as keyword content carried by different sections of a paper, especially the differences between the Abstract and the rest. Therefore, as source for our analysis we used a set of full text articles with a regular section structure, in our study having a defined Abstract, Introduction, Methods, Results, and Discussion (A, I, M, R, D). Another requirement was certain homogeneity of style across the articles (for example, a similar length of the Methods section) and, since there is great interest in the field of data mining on the detection of gene names, the subject should be related to Genetics. Thus, we chose the 104 articles published in *Nature Genetics* from June 1998 (volume 19, issue 2) to June 2001 (volume 28, issue 2), which comply with the AIMRD structure. Note that other journals, or even the Letters of the very same *Nature Genetics*, might have a different structure (for example, lacking separated I, M, R, D sections).

Selection of Keywords

To simplify matters, and following our previous work [4], we focused on the extraction of relevant words (keywords) regarding objects, detected as nouns from natural text by a standard grammatical tagger (TreeTagger, Helmut Schmid, IMS, Stuttgart University, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/). In order to derive keywords from the section of an article, we first compute the associations between the words in the section. Here, we take the sentence as the unit of text to look for associations, that is, two words are associated in the context of a section if they co-occur repeatedly in sentences within that section (see METHODS).

Since words associated strongly to many other words are relevant to the matter that is dealt in the article [5] we use a score (K) that is higher for words with many and strong relations to other words (see METHODS). This measure is used to select words as keywords, in this case, related to objects such as proteins, genes, organisms, etc.

In order to evaluate the performance of the keyword detection, we observed how the selected keywords matched the MeSH (Medical Subject Headings, http://www.nlm.nih.gov/mesh/) terms attached by indexers at the National Library of Medicine to these 104 articles (18.6 on average). Since MeSH terms can be composed of several words (for example, "Learning Disorders"), we selected those composed of a single word (6.80 terms on average). We noted that the most unspecific (for example, *animal*) were often not present in the text and thus could not be matched by a keyword as opposed to species names (*mouse, mycobacterium, human*), or anatomical terms (*hippocampus, cerebellum, breast*). Of those single-word MeSH terms, 4.91 were found on average in the article (as nouns), and 2.22 were among the set of selected keywords (above $K >= 0.3$). Obviously, a more accurate comparison to MeSH terms would require the detection of bigrams, and trigrams (keywords composed of multiple words), but this is out the scope of our work. The recall when matching the original MeSH terms (6.80 on average) went down from 4.91 / 6.80 = 0.72 in the dictionary of 470.6 different nouns present in an article to 2.22 / 6.80 = 0.33 in the 66.6 keywords selected. However, since the size of the list of all nouns found in an article (470.6) is much larger than the number of keywords (66.6), the precision in matching the MeSH terms of an article increased from 4.91 / 470.6 = 0.010 to 2.22 / 66.6 = 0.033.

Keyword Selection by Section

The number of words selected upon a threshold in the K value varies for different sections (see Figure 1). The first observation is that there are a small number of words that have much better K scores than the rest. This means that the organization of words makes it possible to extract keywords for all the five considered sections.

The number of selected words is very similar for all sections for very high values of K (above 0.8). Above a threshold on K ($K >= 0.5$; see Table 1) the resulting number of keywords is quite similar for Introduction and Methods (around 15 for each) with the other three sections producing around nine keywords. However, if one accounts for the size of the sections it is obvious that the frequency of keywords (selected with $K >= 0.5$) per noun is the best in the Abstract (0.18), followed by the Introduction (0.08), with Methods, Results, and Discussion lagging behind. This justifies data mining strategies that focus in the analysis of Abstracts in order to minimize computational

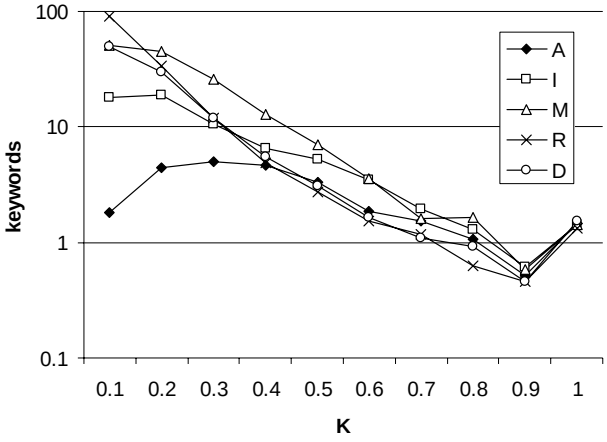

Figure 1
Average number of keywords versus *K* for A, I, M, R, and D sections. The average number of nouns per section is, A = 52, I = 171, M = 404, R = 600, D = 331.

resources. However, this result already indicates that not all keywords are in the Abstract, and that therefore mining the rest of the article may be worthy.

Sections Display Heterogeneous Information

As a way to show that the keyword content in different sections is heterogeneous, we examined which keywords (if any) were selected in all the sections of an article. Our results indicate that, as it could be expected, not many keywords are present in every section and those are not very relevant. Even for a low threshold of $K >= 0.3$, there is on average only one of such general keywords per article. Those are often non-informative words such as "gene", or "protein". This indicates that the information is unevenly distributed across the sections of the article, that is, different sections contain different kind of information.

We illustrate the heterogeneity of the information by section with the keywords selected (for $K >= 0.5$) for a particular article [6] (Figure 2). This work deals with a mutation of the *Nf1* gene of mouse (an exon loss) that produces learning deficits. The only keyword present in every section is the organism under study, the *mouse*. If the Methods section is excluded, only one single more keyword (*mutation*) is selected. Other three-section overlaps give more interesting keywords such as the name of the gene under study (*Nf1*, *neurofibromin*), a domain contained in the resulting protein (*GAP*), the method for testing learning performance of mice (*maze*), or the resulting phenotype (*impairment, lethality*). Keywords unique to different sections tend to correspond to the different information contained in each section. For example, the keywords unique to the Methods section deal with reagents and techniques (*antibody, amersham, tris, primer*).

In order to quantify the differences and similarities of content across the article we have used the number of keywords that are shared between different sections (Table 2). The values indicate that the Methods section is the most different of all. In Methods, the content is usually focused on the techniques and protocols used, and not so much on the biological phenomena that is the main subject of the article. This alone explains why those keywords present in every section (for example *protein*, *gene*) are scarce and uninteresting.

Regarding similarities between sections, A, I, and D are evenly similar among them, and R is the closest to M, as it is shown when plotting the distance matrix of Table 2 as a dendogram (see Figure 3). This is probably due to the fact that the Results section deals with the protocols used, although not as explicitly as the Methods section. The Discussion focuses again on the biological results (stressing their relation to the current knowledge) without detailing the techniques that have already been explained in Methods and justified in Results.

This result indicates that each section contains certain keywords that are unique to the section. In the following we try to characterize what are the differences in content between sections.

Qualitative Analysis of Subjects per Section

To make a deeper analysis of the kind of information present in each of the sections, we classified in seven categories a set of words present in our corpus of 104 articles (among the most frequent nouns). In order to do so as unambiguously as possible, we selected words that matched MeSH descriptors also consisting on that single word and belonging to only one major MeSH category (see METHODS). We added another category not present in MeSH, that of "Units, Dimensions, & Parts" in order to account for many terms that are currently not MeSH terms but are of interest to us.

Table 1: Keyword selection per section.

	all	K >= 0.3	K >= 0.4	K >= 0.5
A	52.17	19.44	14.42	9.77
I	171.32	31.03	20.47	14.00
M	404.19	54.24	28.50	15.80
R	599.98	24.74	12.74	7.85
D	331.04	26.16	14.25	8.75

Average number of nouns per section (all), or number of those selected as keywords for three different thresholds on the *K* score.

Table 2: Average number of keywords (*K* >= 0.5) shared by two sections for the corpus of 104 articles.

	A	I	M	R	D
A		2.01	0.92	1.77	2.20
I	2.01		0.81	1.34	2.02
M	0.92	0.81		1.55	1.02
R	1.77	1.34	1.55		1.99
D	2.20	2.02	1.02	1.99	

Table 3: Detection of gene names appearing only in the Methods section.

		Ref
Incorrect:	Restriction endonucleases	
	Msp1	v27.n3.277
	Pst1	v19.n4.340
	Sac1	v27.n4.375
	Vector name	
	Psg5	v23.n3.287
	Cell strain	
	Tig3	v26.n3.291
Correct (technical context):	Definition of a Yeast strain	
	Can1, *Leu2*, *Lys2*, *Trp1*	v26.n4.415
	In array	
	Faf1	v20.n3.266
	Growth detection	
	Mcm5, *Mcm6*	v25.n3.263
	Platelet mRNA analysis	
	Pbp2	v23.n2.166
	Primers used to determine embryo sex	
	Zfy1, *Zfy2*	v27.n1.31
Correct:	Analysis of mutant phenotypes	
	Pmd1	v24.n4.355
	cDNA probe	
	Rab2	v19.n2.134
	SNP found in cDNA	
	Add3, *Npr2*	v22.n3.239
	Identifier given	
	Pom1	v28.n3.223
	Detection of meiosis specific genes	
	Mei4, *Mek1*, *Sps4*, *Zip1*	v26.n4.415

Ref: reference of the article in *Nature Genetics* by volume, issue number and first page of the article.

The results (See Figure 4a) indicate that the large sections are a good source of keywords, obviously Methods gathering many terms related to techniques. Introduction, Results and Discussion contain a good deal of information regarding diseases. However, again, the Abstract section is shown as the best source for most subjects regarding frequency of keywords (Figure 4b) except for those typical of the Methods section (Techniques & Equipment; Chemicals & Drugs).

Distribution of Gene Names

Since the detection of gene and protein names is a very important subject, broadly used for the detection of macromolecular interactions (see for example [7]), and because, as stated in the introduction, we are concerned

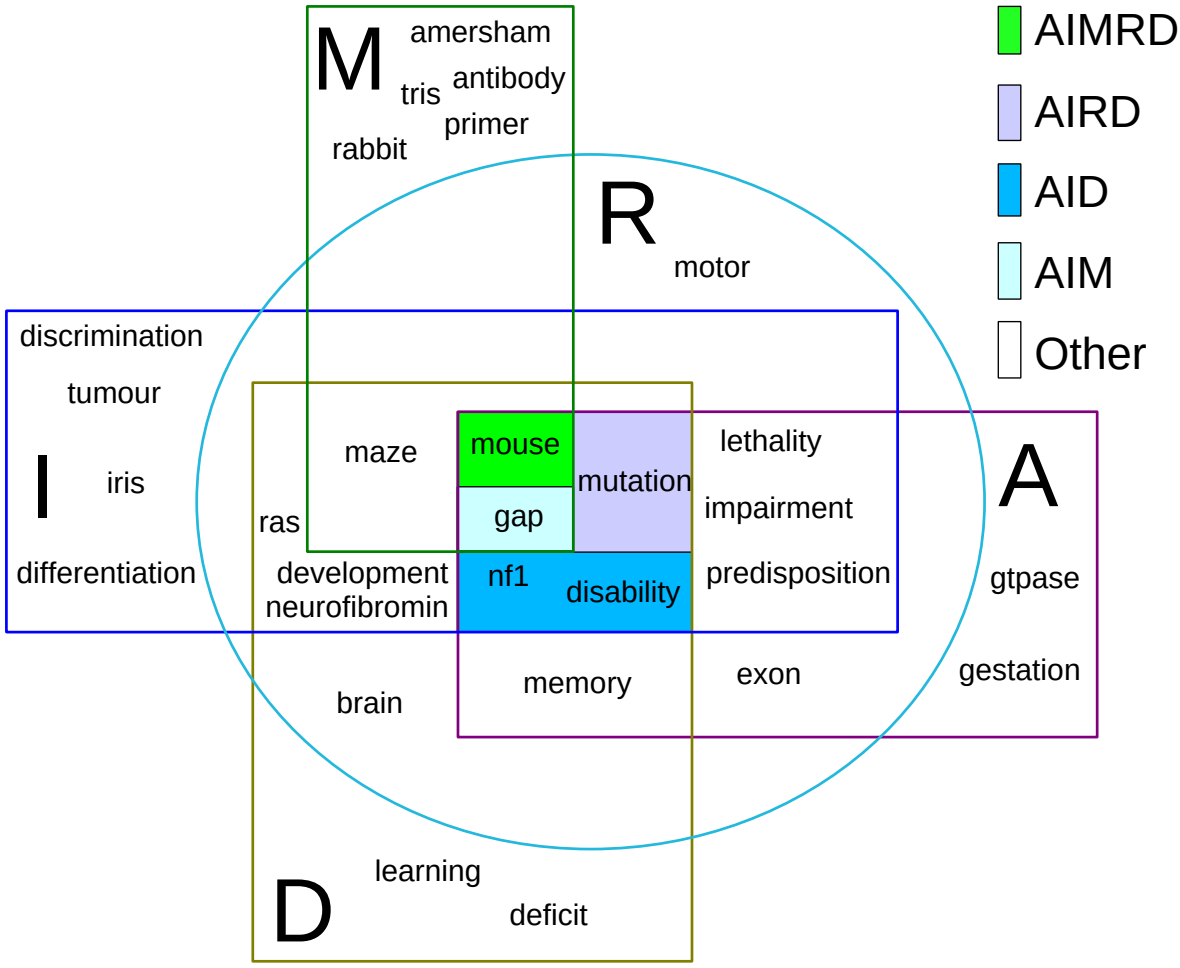

Figure 2
The keywords selected for an article [6] with a *K* >= 0.5 are represented as they appear in the different sections of the article.

about the relevance of matching gene names in different sections of an article, we examined the distribution of gene names across sections.

From a long list of genes names derived from the SWISS-PROT database [8], we selected a very restricted set of 539 genes whose names are composed of three letters followed by one single digit, thus very difficult to be mistaken to other words not being genes. For example, there are gene names called *Not* or *That*. Shorter names (e.g. *A6*) can also be a problem. A total of 224 gene names out of the 539 was matched in 76 of the 104 articles. The Results section was the one with the greatest number of unique gene names (Figure 5a). Again, the Abstract, and then the Introduction, are the sections with the highest frequency of these names (Figure 5b).

In order to illustrate the problems that affect gene-name identification if context is ignored (even using gene names apparently easy to recognize) (discussed for example in [9]) we checked manually the context of gene names that were exclusively mentioned in the Methods section. Of the 224 genes, just 24 were mentioned in the Methods section of the corresponding 14 articles and not elsewhere (see Table 3). In five of the 14 articles, the name was referring to a non-gene object (three restriction endonucleases, a vector name, and a fibroblast cell strain). In five articles, the gene was mentioned in a technical context (usually, the gene mRNA level was used for analysis of cell state) and no biological process involving the gene was described. In only five articles we found the mention of the gene name relevant (See Table 3). Additionally, we noted that of these 24 gene names, at least two (*Pbp2*,

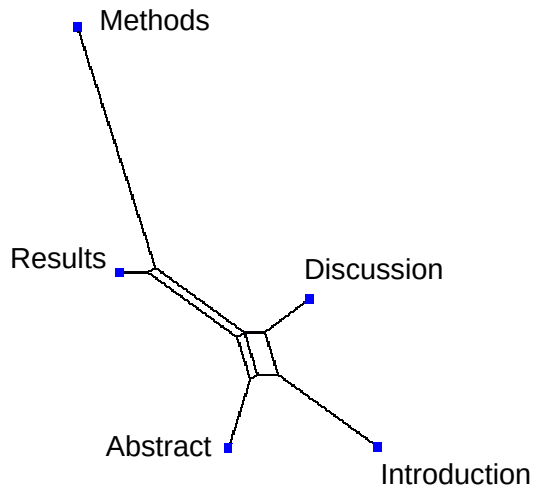

Figure 3
In order to display graphically the similarity between sections regarding keyword content, we took the inverse of the average number of shared keywords (Table 2) as a measure of dissimilarity between sections, and we plotted it as a dendogram (using [19]).

Pom1) could refer to two non-homologous (unrelated) genes, and another one (*Sac1*) to four; such polysemous gene names complicate gene identification from text. Biologists are aware of such problems (see for example [10]). In summary, extreme caution should be applied with gene names appearing uniquely in the Methods section because the context of gene names there is very different to that seen in the rest of the article. If automated methods to extract gene names from text are applied to the Methods section, those that explore the context of gene names using part-of-speech tagging (for example, [11]) or Hidden Markov Models (for example, [7]) should then perform better than those that just take co-occurrences of gene names [12,13].

Discussion

There is a clear need for doing information extraction of biological data from full text scientific articles and the means for doing it are there with computers better suited for faster computation every day and new methodologies for Natural Language Processing that can be used for biomedical literature (see for example [14]). Regarding the source of data, the full text electronic versions of journals are now more the rule than the exception, with initiatives in the way towards the construction of large public repositories of such information (although hotly debated; see about PubMed Central [15,16]).

In this work we have shown that the distribution of information (as keywords) in full text articles is heterogeneous and that there is certain correspondence of article sections with different kind and density of relevant data. The Abstracts are shown as the best repository from the point of view of having many keywords in a short space, justifying previous information extraction approaches. The lack of large repositories of full text articles in contrast to the current eleven million of references (many of them with their abstract) in the MEDLINE database, are another advantage of the Abstract approach.

However, we have shown that there is much more relevant information (at least in a ratio of 1:4 regarding gene names, anatomical terms, organism names, etc.) in the rest of the article. We have demonstrated that the information is structured enough to get important numbers of relevant keywords, but that for certain words (such as gene names) caution has to be taken regarding the context of the word.

We propose that the text mining of full text articles should be approached with different strategies for different sections. Beyond the Abstract, the Introduction looks like the best place to look for protein and gene names (and interactions) since it is probably describing current knowledge. The Discussion section, that interprets the results and put them in context with the current knowledge, looks like the third best place for mining such information, with Methods probably as the worst place. The Results section could be problematic given its mixed nature between Methods and the rest.

Regarding other subjects, such as keywords about biological concepts (species, tissues, diseases, etc.), again the Abstract and then the Introduction section look like the best sections to mine regarding frequency of such keywords, but Results and especially Discussion seem better from a quantitative point of view. The Methods section is clearly appropriated for looking for technical data, measurements, and chemicals. Respect to chemicals, again, their context can be completely different in this section compared to the rest.

Conclusions

Extraction of biological information from full text looks promising, but context must be regarded. Part of this context is given by the situation of the text under analysis within the article. Therefore, tuning the extraction of information to the section is probably a good strategy, and for particular tasks some sections should be avoided.

We have shown that the kind of simplistic annotation that constitutes tagging a fragment of an article as belonging to a characteristic section is already useful for text mining.

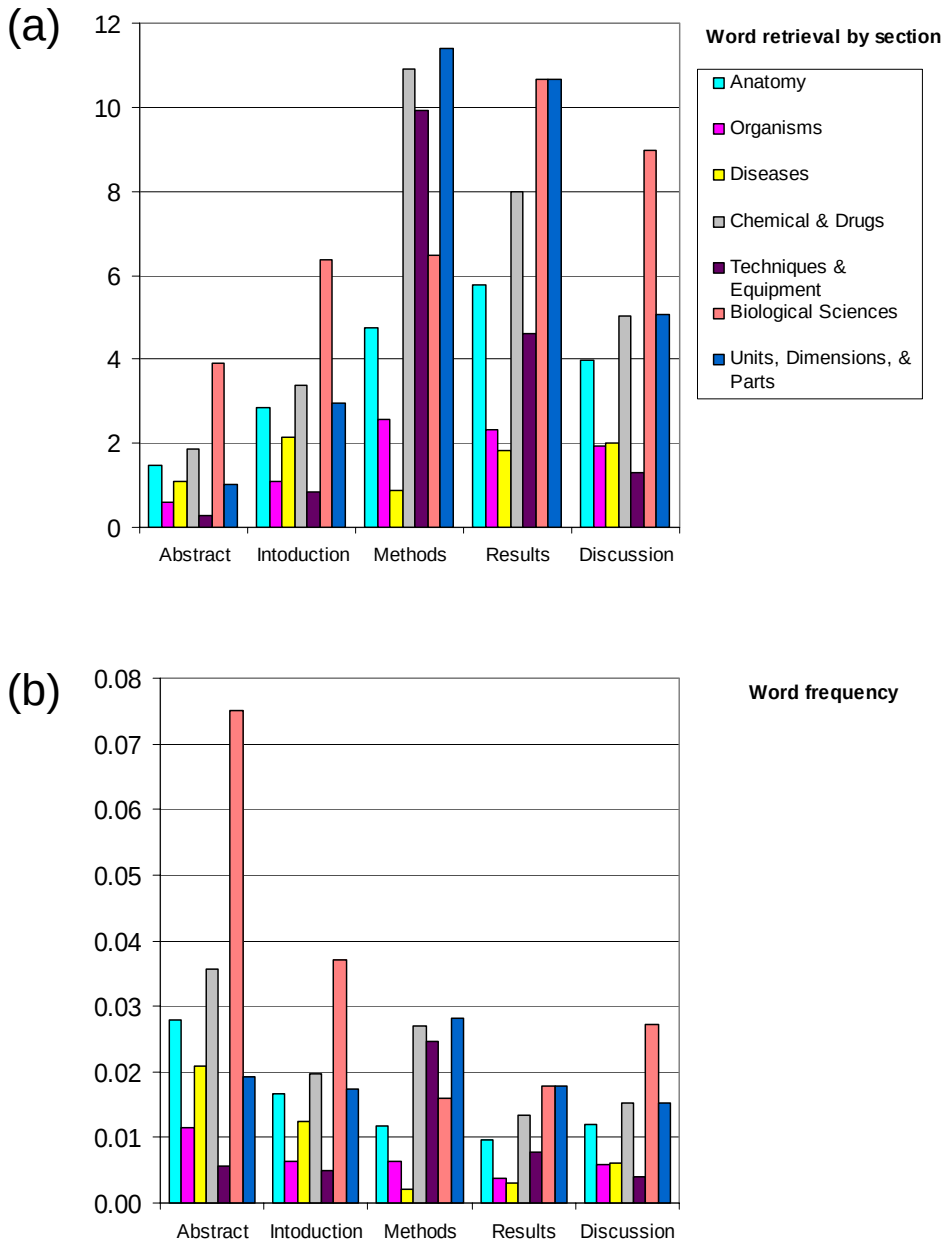

Figure 4
Word categories present in the five sections under analysis. Classes according to MeSH are A (Anatomy), B (Organisms), C (Diseases), D (Chemicals & Drugs), E (Techniques & Equipment), G (Biological Sciences). An additional class X was defined in this work (Units, Dimensions, & Parts). The number of words used for the analysis was 36 (class A), 14 (B), 11 (C), 47 (D), 33 (E), 41 (G), 49 (X). **(a)** Average number of occurrence of words of each subset per section. **(b)** Frequency of words of each subset per total number of words for each section.

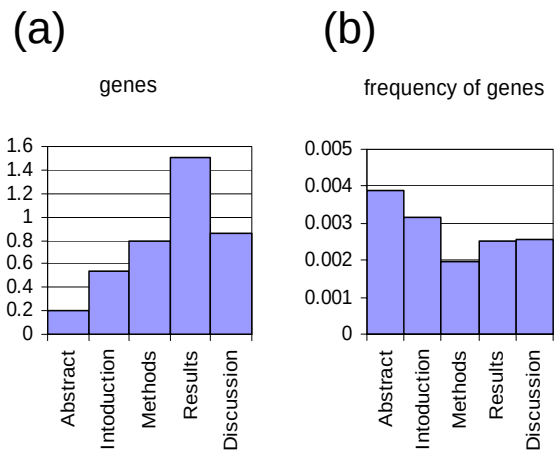

Figure 5
Distribution of matches to a set of 224 gene names across sections. **(a)** Average number of unique gene names per section. **(b)** Frequency of different gene names per total of nouns for each section.

But further tagging using markup codes in XML style [17] identifying biological objects and concepts (under development; see for example [18] or the GENIA project http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/) could ultimately make text mining a children's game. We hope for future interfaces for writers of Molecular Biology articles that should do the job upon validation by the authors (for example, marking every occurrence of a gene name with a unique and stable link to any of the existing gene sequence databases). For this to happen, the collaboration between both scientists and publishers will be very important.

Methods

Derivation of Associations between the words of a section

Given a section from an article, we split the text in sentences using a standard part of speech tagger (TreeTagger). We only computed associations between the words identified from the tagging as nouns. Following [4], the association between two words (w_i, w_j) (for example, "cell" and "cycle") can be modeled as the degree of inclusion of one word into the other $(\tilde{I}_W)$ which can defined as the fuzzy binary relation given by: $\mu_{\tilde{I}_W}(w_i, w_j) = \frac{|W_i \cap W_j|}{|W_i|}$, that is, the ratio of the number sentences where both words w_i and w_j co-occur to the number of sentences the word w_i occurs. This is an asymmetric relation very appropriate to model hierarchical relations between words as they happen in natural text. For example, in some Cell Biology context, the word "cycle" could appear always related to the word "cell" (as in "cell cycle"), but the word "cell" can be related to many other words such as in "cell growth", "cell membrane", or "cell nucleus". Accordingly, the inclusion value of the word "cycle" into "cell" will be close to one and the inclusion value of the word "cell" into the word "cycle" will be close to zero.

Selection of Keywords

We identify a word as relevant for the text analyzed if it establishes many and strong relations to other words (following [4]). Therefore, in a given section, we define a score for a word w_i that is equal to $K_i = \sum_{j \neq i} \mu_{\tilde{I}_W}(w_i, w_j)$, normalized to the maximum value found for K of any word in that section. Then, the keywords of the section are defined as those words that have a K score above a certain value.

Classification of Words in Subjects

In order to classify words into categories we used the following procedure. We chose the MeSH (Medical Subjects Headings) classification from the National Library of Medicine. All MeSH terms (including official synonyms) composed of one single word were selected and then the stem of the word was computed using TreeTagger. The words present in our corpus of 104 articles were ordered by frequency and all words occurring more than 200 times were selected. Those matching the selected single-word MeSH headers from six categories (A, B, C, D, E, and G; See the caption of Figure 4 for descriptions) were selected as belonging to those classes. In order to avoid possible miss-annotations, words matching more than one category were discarded. Manual analysis of the resulting table of associations was carried out in order to check the associations and to make new ones. A new class not present in MeSH (the X class of "Units, Dimensions, & Parts") was generated in order to include a large number of terms mainly present in the Methods section.

Authors' Contributions

PS carried out the analysis of the keyword distribution from a database of full text articles. CP developed and applied the method to compute keywords. MA prepared the figures (except fig 2 by PS) and conceptualised the structure of the paper. PB and MA co-directed the project and contributed to the final manuscript. All authors collaborated during the whole length of the project. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to the developers and maintainers of the different databases used in this work (SWISSPROT, MeSH, *Nature Genetics* full text repository), to Helmut Schmid (IMS, Stuttgart University) for distributing

TreeTagger, to Harindar S. Keer for help with the data management, and to the members of our group at EMBL-Heidelberg for fruitful discussions.

References

1. Andrade MA and Bork P: **Automated extraction of information in molecular biology** *FEBS Lett* 2000, **476:**12-17.
2. Blaschke C, Hirschman L and Valencia A: **Information extraction in molecular biology** *Brief Bioinform* 2002, **3:**154-165.
3. de Bruijn B and Martin J: **Literature mining in molecular biology.** *Proceedings of the EFMI Workshop on Natural Language: Processing in Biomedical Applications* Edited by: *Baud R and Ruch P. Nicosia, Cyprus*; 2002:1-5.
4. Perez-Iratxeta C, Keer HS, Bork P and Andrade MA: **Computing fuzzy associations for the analysis of biological literature** *Biotechniques* 2002, **32:**1380-1385.
5. Perez-Iratxeta C, Bork P and Andrade MA: **XplorMed: a tool for exploring MEDLINE abstracts** *Trends Biochem Sci* 2001, **26:**573-575.
6. Costa RM, Yang T, Huynh DP, Pulst SM, Viskochil DH, Silva AJ and Brannan CI: **Learning deficits, but normal development and tumor predisposition, in mice lacking exon 23a of Nf1** *Nat Genet* 2001, **27:**399-405.
7. Collier N, Nobata C and Tsujii J: **Extracting the names of genes and gene products with a Hidden Markov Model** *COLING'2000* 2000:201-207.
8. Bairoch A and Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000** *Nucleic Acids Res* 2000, **28:**45-48.
9. Stapley BJ and Benoit G: **Bibliometrics: information retrieval and visualization from co-occurrence of gene names in MedLine abstracts** *Proceedings of the Pacific Symposium on Biocomputing Oahu, Hawaii, World ScientificPress*; 2000:526-537.
10. Luan S, Kudla J, Harter K, Gruissem W and Chory J: **Renaming genes and duplication of gene names in the literature** *Plant Cell* 2001, **13:**2391-2392.
11. Tanabe L and Wilbur WJ: **Tagging gene and protein names in biomedical text** *Bioinformatics* 2002, **18:**1124-1132.
12. Stapley B.J. Benoit G.: **Bibliometrics: Information Retrieval and Visualization from co-occurrence of gene names in MedLine abstracts** *Proceedings of the Pacific Symposium onBiocomputing Oahu, Hawaii, World ScientificPress*; 2000:526-537.
13. Jenssen TK, Laegreid A, Komorowski J and Hovig E: **A literature network of human genes for high-throughput analysis of gene expression** *Nat Genet* 2001, **28:**21-28.
14. de Bruijn B and Martin J: **Getting to the (c)ore of knowledge: mining biomedical literature** *Int J Med Inf* 2002, **67:**7-18.
15. Roberts RJ: **PubMed Central: The GenBank of the published literature** *Proc Natl Acad Sci U S A* 2001, **98:**381-382.
16. Eisen MB, Brown PO and Varmus HE: **Public-access group supports PubMed Central** *Nature* 2002, **419:**111.
17. St. Laurent Simon: **XML Elements of Style** *New York, McGraw-Hill*; 2000.
18. Ettinger M: **The complexity of comparing reaction systems** *Bioinformatics* 2002, **18:**465-469.
19. Huson DH: **SplitsTree: analyzing and visualizing evolutionary data** *Bioinformatics* 1998, **14:**68-73.