Minireview

Automated extraction of information in molecular biology

Miguel A. Andrade[a,b], Peer Bork[a,b,*]

[a]European Molecular Biology Laboratory, Meyerhofstr. 1, D-69012 Heidelberg, Germany
[b]Max Delbrück Center for Molecular Medicine, Department of Bioinformatics, P.O. Box 740238, D-13092 Berlin-Buch, Germany

Received 16 May 2000

Edited by Gunnar von Heijne

Abstract We review data mining techniques in molecular biology, specifically those that extract information from the scientific literature itself. As more of the biological literature is published electronically, there is an opportunity, and even a need, to automatically summarize the literature in a customized way, for example by associating keywords to a topic. These keywords can be extracted from relevant publications. The process of keyword extraction can be automated and optimized to keep literature pointers automatically up-to-date or to filter relevant information from the literature. To illustrate these points, OMIM (Online Mendelian Inheritance in Man), a database of human inherited diseases, was linked to the literature and keywords were derived that covered distinct aspects such as genetic information on the one hand and disease-specific protein and phenotypic information on the other. They were used to extract information that is helpful for keeping entries about disease up-to-date. © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

1. Introduction

New scientific discoveries are based on the existing knowledge which has to be accessible and thus usable by the scientific community. In the 19th century, the spread of scientific information was still done by writing letters with new discoveries to a small number of colleagues. Printed journals took over this job professionally. We are now on another transition into electronic media. Electronic storage allows the customized extraction of information from the literature and its combination with other data resources such as heterogeneous databases. In fact, it is not only an opportunity, but also a pressing need as the volume of scientific literature is increasing immensely (see Fig. 1). Furthermore, the scientific community is growing so that even for a rather specialized field it becomes impossible to stay up-to-date just through personal contacts in that particular community. The growing amount of knowledge also increases the chance for new ideas based on combining solutions from different fields, i.e. one has to be able to retrieve information from areas one is not so familiar with. For example, a molecular biologist researching a human gene may need to link research carried out by geneticists (finding the DNA sequence of the gene), biochemists (characterizing the protein coded by that gene), and physiologists (characterizing a disease related to a defect in that gene). There is a necessity of accessing and integrating all scientific information to be able to judge the own progress and to get inspired by new questions and answers.

The classical way to overview the literature is via reviews (and the fraction of reviews per original research paper is increasing); however, already in the 1960s, an alternative method developed in the form of Current Contents®, published monthly, which provide lists of papers indexed by keywords and by author name (http://www.isinet.com/products/cc/). With the advent of Current Contents® on diskette in the late 1980s, computers made the querying easier and allowed keyword searches. An obvious problem was that of the storage of the data. This was solved by the development of the Internet, which made bibliographic databases more accessible. The best example is MEDLINE, a collection of over 10 million citations compiled by the National Library of Medicine (http://www.nlm.nih.gov/) covering publications since 1966. Since 1997, this database is freely accessible through PubMed (http://www3.ncbi.nlm.nih.gov/Entrez/index.html), the National Library of Medicine's search service. Today, it is possible to select a number of articles using queries that can be very complex (including combinations of different attributes of the publication and logical operators) and to read a short summary associated (the abstract) usually provided by the authors. With the increasing distribution of journals in electronic format, the full text of any paper will soon be only one click away. There are already some initiatives heading to provide such repositories in Molecular Biology (e.g. PubMed central, http://www.pubmedcentral.nih.gov/)[1].

However, access to literature does not solve the problem of the selection of information. For example, specific topics such as the molecule 'protein kinase' or the disease 'neurofibromatosis' are mentioned in 13 212 and 3043 papers, respectively, only since 1990. Reading or even browsing all of those publications is something that most researchers will not contemplate.

Databases with specialized content might be one solution and the creation of those is currently very popular. A typical problem with these databases is that they often mix heterogeneous kind of information. For example, in sequence databases it is common to find that the function of a protein is

*Corresponding author.
E-mail: peer.bork@embl-heidelberg.de

[1] The World Wide Web deserves to be mentioned as a public place for deposition of molecular biology (or of any) information. However two problems hamper its use as a source of scientific information in comparison to papers: it is not peer reviewed, and it is too heterogeneous. For example, you may try a search with the words 'anti-cancer' and 'drug' in any of the web search engines (e.g. Altavista, http://www.altavista.com/).

PII: S0014-5793(00)01661-6

labelled as 'it binds protein A' without source information on the respective experiment. If one is lucky to identify the respective article it may turn out that this statement comes from the discussion part where it is noted that "perhaps, it *might* bind protein A" based on very weak evidence. Maintenance and annotation of those databases become increasingly difficult due to the information flood. Although sequence databases are increasingly subjected to automatic annotation (e.g. [1,2]), they still face a huge information shortage and only some functional features are currently completely covered.

Clearly, there is a necessity of developing methods for automatic extraction of relevant information from any source of scientific data, especially sources such as literature written in human language (also known as natural language). Linguists have been working since the 1960s in the computational analysis of natural language, a difficult task given its ambiguities and complexities. However, applications that deal with scientific text have better chances of success given that scientific language is per se simpler than common language: the vocabulary is smaller and the definition of terms is more accurate. In Section 2, we will review the recent advances in the field and some of the terminology. Section 3 will deal with applications to molecular biology. Section 4 will give a simple example: the deduction and use of keywords related to human disease.

2. Computational approaches to the extraction of information from text

Data mining defines the compendium of techniques to identify pieces of information contained in textural sources. For text sources (corpus) written in natural language (NL) (usually English[2]) there are computational techniques for text analysis [4,5].

Information retrieval (IR) techniques [6] are used to select documents that are relevant according to a user's needs. Information extraction (IE) techniques [7,8] are used to extract relevant information from text according to pre-specified templates (e.g. for a terrorist action, extract place, date, victim and outcome). They do not need an understanding of the text under analysis [9], which is approached by natural language processing (NLP), but they can benefit from it.

NLP can be applied at the level of words out of context (for lexical matching and morphological analysis or stemming [10]) or at the level of sentences (for syntactic parsing, namely, analyzing a sentence to determine its structure, usually in order to identify noun sentences and their components).

Understanding a text can ultimately be possible only if the system can refer to an ontology (or controlled vocabulary), i.e. the association of words to meanings, maybe including hierarchical relations between them [11]. They can be general (e.g. WordNet from the Cognitive Science Laboratory, Princeton University, http://www.cogsci.princeton.edu/~wn/) or specific to a domain of knowledge, e.g. to medicine as the unified medical language system (UMLS, http://www.nlm.nih.gov/research/umls/umlsmain.html) or to eukaryotic genes as in gene ontology (http://www.geneontology.org; [12]).

The latest IE algorithms are periodically tested against texts from the news on various subjects in the Message Understanding Conferences (MUCs, http://www.muc.saic.com/). Similarly, IR systems are tested in Text Retrieval Conferences (TReC, http://trec.nist.gov/). General applications of IE/IR systems are numerous varying from indexing systems that work with a controlled vocabulary (e.g. RUBRIC [13]), others that derive it from the analyzed text itself (e.g. CLARIT [14]), or others that use a structured vocabulary and do more complex NLP parsing (e.g. FERRET [15]; or Condorcet [16]).

3. Applications of text analysis to molecular biology

Due to easy access and availability, the most widely used sources of information are abstracts of scientific publications. They already contain a concise description of the most important parts of the information carried by a paper and can be more informative than selected parts of the corresponding papers [17].

Most of the systems for abstract mining use neither NLP nor any kind of ontology. Exceptions include the work of Otha et al. [18], which expands IR/IE queries using words close to those used in the query according to a pre-compiled dictionary. Proux et al. [19] use a dictionary of the fly *Drosophila melanogaster* genes to extract protein–protein interactions in this organism. Some systems that do not rely on an ontology use the matching to pre-specified templates as a way of detecting protein names [20] or protein–protein interactions [21–23]. Statistics of word co-occurrence have been used for the extraction of keywords related to protein functionality [24] and for the inference of protein–protein interactions [25]. More elaborated systems take advantage of NLP (at the simple level of detecting and analyzing noun sentences) for the extraction of anti-cancer drugs [26] or gene or protein–protein interactions [27,30].

Given the fact that full text papers are becoming available in electronic format soon and that there is a need to integrate literature with other information resources, many groups are

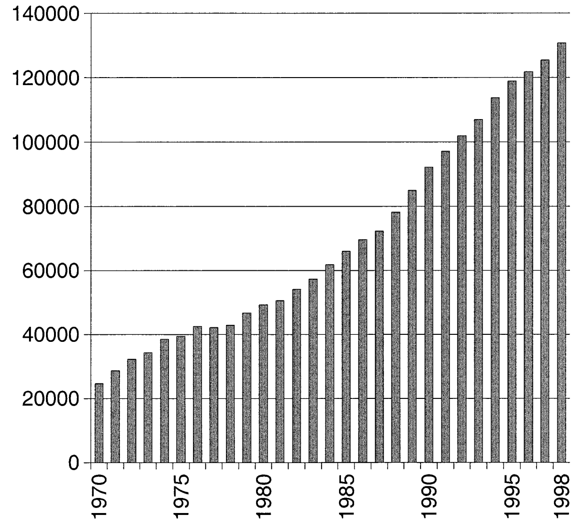


Fig. 1. Number of abstracts found in MEDLINE by querying PubMed with the word 'protein' vs. year.

[2] As pointed out by Bryson [3] (p. 2) "for better or worse, English has become the most global of languages, the lingua franca of business, science, education, politics, and pop music". However, "there is no reliable way of measuring the quality or efficiency of any language" ([3], p. 8), and therefore English cannot be said to be the best language for science. Latin and German have been used before. The use of English is just a historical accident.

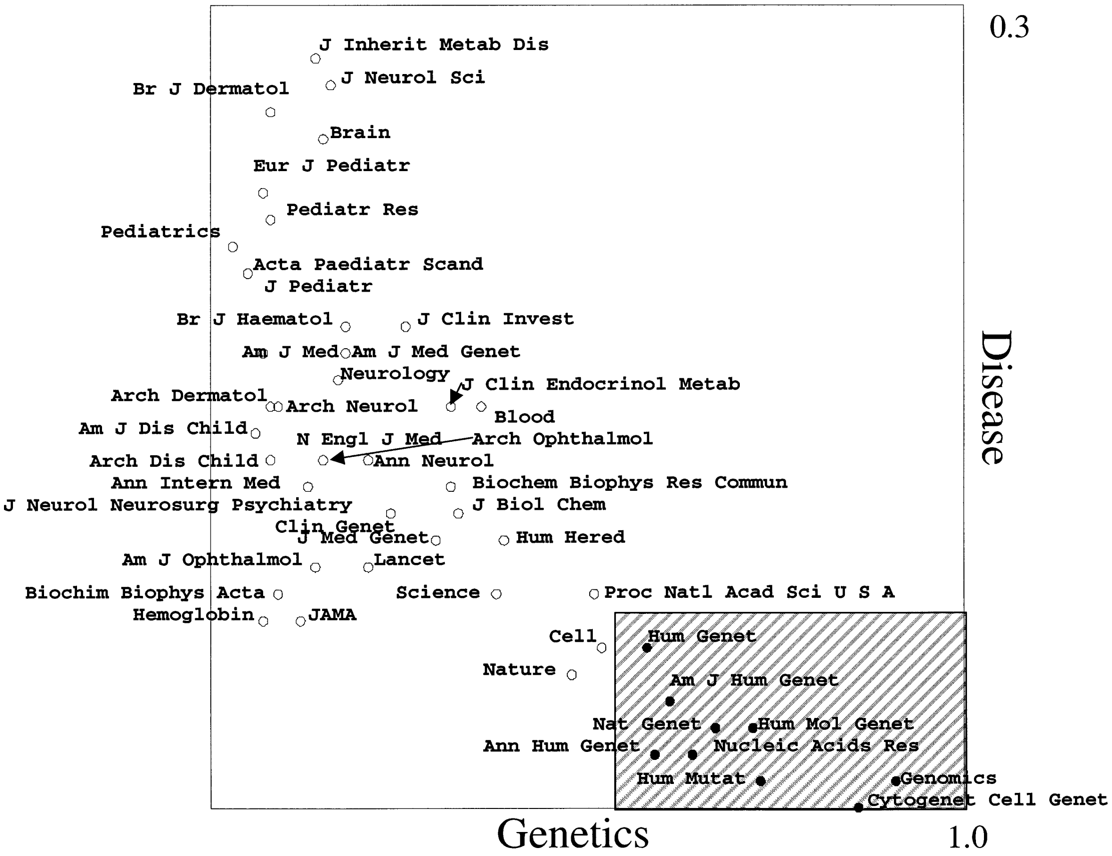
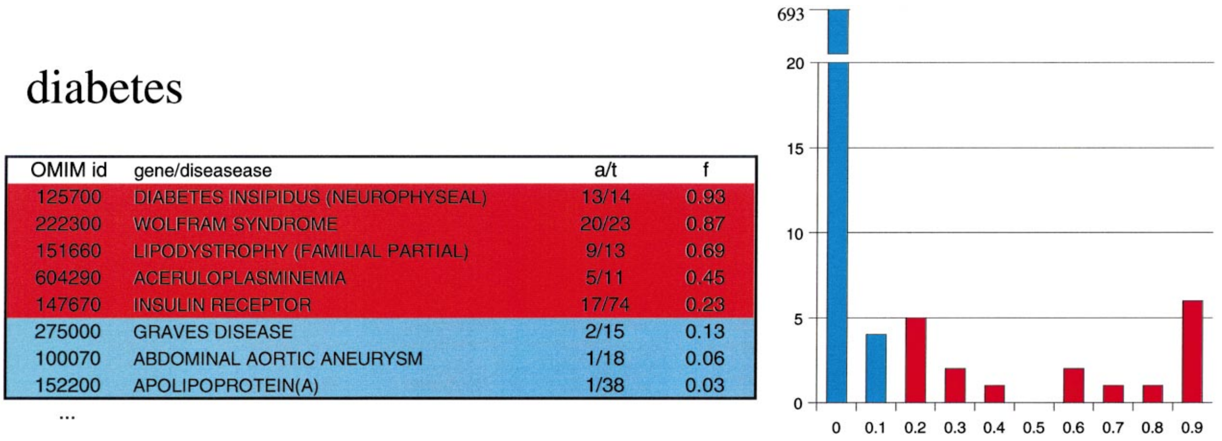


Fig. 2. Distribution of journals by word content in the title of the papers linked in the set of entries under analysis. *X*-axis: fraction of papers from the journal containing at least one of the following words related to genetic mapping (genetics, locus, marker, chromosome, localization, assign, link, clone-, candidate, map, mutation, mutant, screen, polymorphi-, deletion, allele). *Y*-axis: fraction of papers from the journal containing at least one of the following words related to phenotype descriptions (activit-, impair, clinical, treated, treatment, review, case, biochemical, defect, abnormal, deficient). The region marked in the graph contains journals that were considered to belong to the domain of genetics. Note that some journals containing the word 'gene' or 'genetics' did not fall in this region. The analysis was restricted to journals highly cited in OMIM (more than 100 references).



OMIM id	gene/diseasease	a/t	f
125700	DIABETES INSIPIDUS (NEUROPHYSEAL)	13/14	0.93
222300	WOLFRAM SYNDROME	20/23	0.87
151660	LIPODYSTROPHY (FAMILIAL PARTIAL)	9/13	0.69
604290	ACERULOPLASMINEMIA	5/11	0.45
147670	INSULIN RECEPTOR	17/74	0.23
275000	GRAVES DISEASE	2/15	0.13
100070	ABDOMINAL AORTIC ANEURYSM	1/18	0.06
152200	APOLIPOPROTEIN(A)	1/38	0.03

...

Fig. 3. Computation of the usage of the word 'diabetes' in the 721 OMIM entries. Table: OMIM id; description of the disease/protein; *a*/*t*, number of abstracts linked to that OMIM entry containing at least once the word 'diabetes'/total number of linked abstracts; *f*, fraction of abstracts containing the word 'diabetes'. Graph: counts for number of OMIM entries with *f* in a given range. For example, 693 entries had $f < 0.1$. Observe that there is a first minimum for $f = 0.1$–0.2. $f > 0.2$ was taken as threshold for selecting the word 'diabetes' as keyword. The top five diseases given as examples in the table are above the threshold. They are all clearly related to 'diabetes' as it can be seen from the name of the disease (e.g. diabetes insipidus) or from sentences extracted from the abstracts: e.g. "wolfram syndrome is characterized by optic atrophy insulin dependent diabetes mellitus vasopressin sensitive diabetes insipidus and neurosensory hearing loss" or "a mutation in the insulin receptor gene that impairs transport of the receptor to the plasma membrane and causes insulin resistant diabetes". 'Diabetes' is mentioned incidentally for the three bottom diseases, but they are not likely to be related to it, e.g. "the genetic susceptibility to graves disease and type 1 insulin dependent diabetes mellitus is conferred by genes in the human leukocyte antigen region on the short arm of chromosome 6 but several other genes are presumed to determine disease susceptibility" or "there were differences in risk factors for hernia and diabetes mellitus among the probands with peripheral arterial aneurysm (aaa) or arteriomegaly but none for relatives". According to this distribution, the word 'diabetes' is a potential keyword, which will be selected for sets of abstracts with fraction values above the first minimum (blue bars in graph, or blue rows in table).

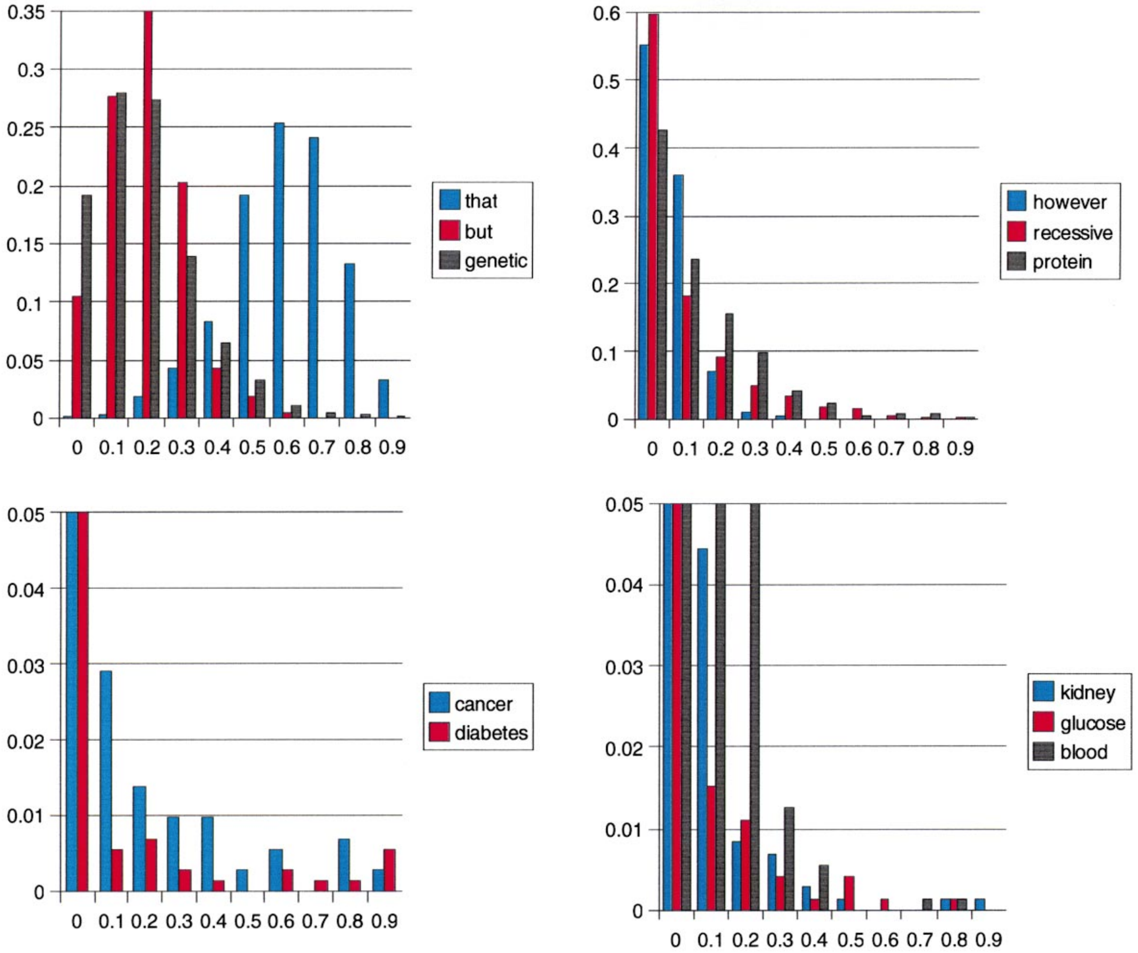


Fig. 4. Typical distributions of word usage over the 721 OMIM entries analyzed. Top left: presenting one maximum; top right: continuously decreasing; bottom: at least one minimum (left, relatively frequently used 'potential' keywords, right, rarely used 'potential' keywords).

currently establishing data mining capabilities, although little of that effort is so far reflected in the biological literature. In order to illustrate a simple IE application, and to show the power even of simple statistical approaches, we have performed a case study that makes use of curated knowledge (a database that integrates medical, phenotypic, biochemical and genetic information on inherited diseases, OMIM (Online Mendelian Inheritance in Man); http://www.ncbi.nlm.nih.gov/omim/, [28]) and the abstracts available in MEDLINE (http://www.nlm.nih.gov/). The goal is two-fold. First, to establish customized keyword lists for any disease enabling to distinguish between genomic and proteomic information. Second, to develop a keyword extraction system for any disease allowing the automatic update of the information on diseases as well as mining relevant literature not been yet linked to the database.

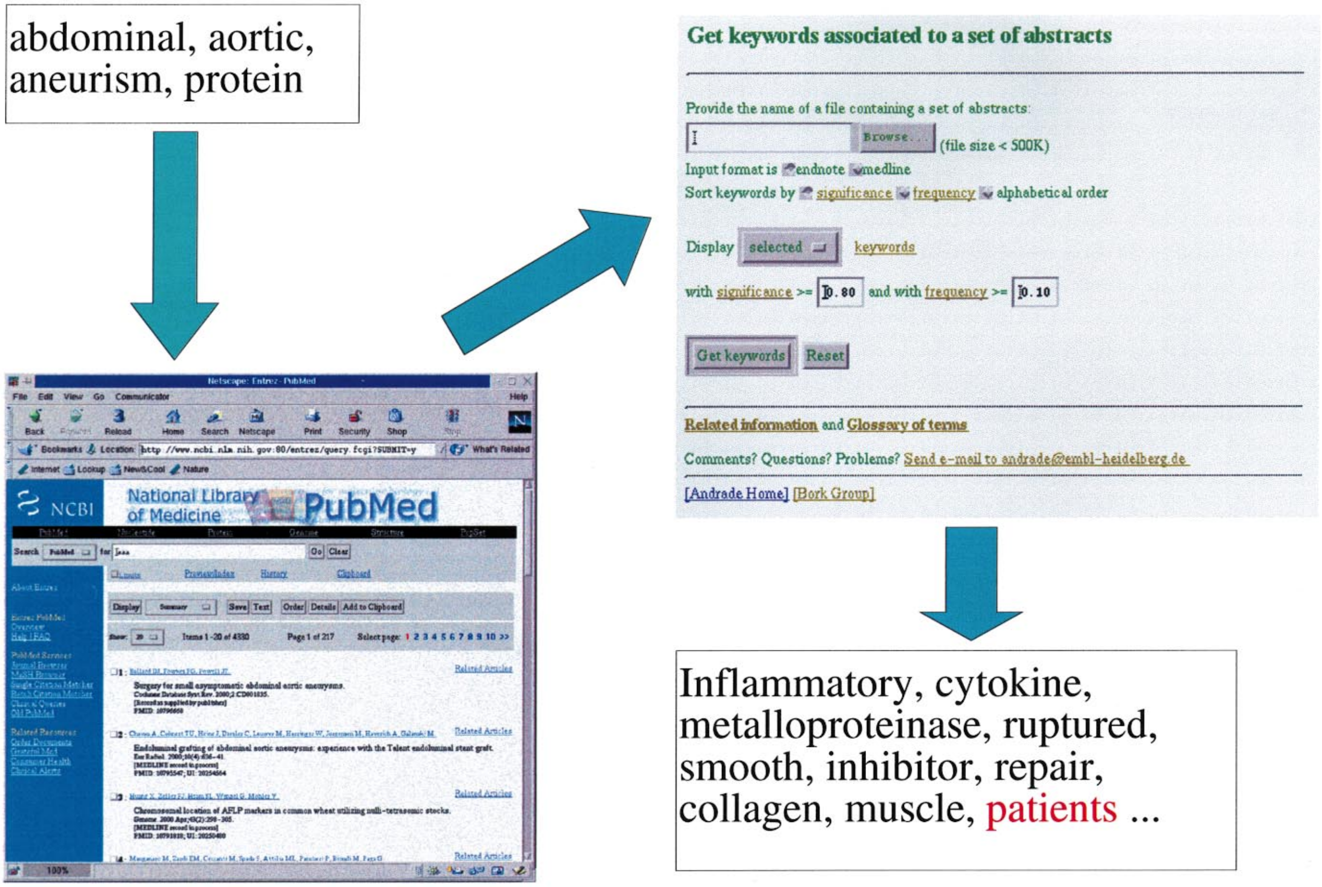
4. A simple case study

OMIM [28] is a catalogue of human diseases assumed to be genetically inherited. It is very well curated and thus extremely useful in many respects. Each entry in OMIM contains data about one disease, including information on related genes and mutations. This information is derived from the scientific literature. Given the limited human resources and the increasing volume of diverse literature, it is becoming more likely that recent information relevant to the database entries is not integrated appropriately. Another problem inherent to curated databases is the subjectivity involved in the phrasing, selection of facts, and the literature digest. Although a human curator is clearly superior to any automated system, an automatic IE system could be helpful for consistency, updates, and on-the-fly selection of information according to particular users' interests. With this goal in mind, we have developed a method that extracts keyword lists that describe certain aspects of a disease.

4.1. Genetic vs. phenotypical information

Our source of keywords is the text of the abstracts linked to each different disease entry in OMIM. Potential keywords for the description of a disease should discriminate significantly between one or more diseases and the rest. However, OMIM entries have links to papers detailing the genetics of the disease (e.g. describing linkage analysis or chromosomal location of related genes), which are not likely to contain keywords describing biochemistry or phenotypes. Therefore, a first step is to distinguish papers belonging to these two information profiles (genetic vs. biochemical, phenotypic and medical).

As a simplification, we reduced the problem to the characterization of journals instead of papers. First, we compiled an arbitrary list of words (or wordstems) likely to be related to mapping and mutation data (locus, marker, chromosome, lo-



abdominal, aortic, aneurism, protein

Netscape: Entrez - PubMed
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?SUBMIT=y
What's Related
NCBI
National Library of Medicine
PubMed
Search PubMed for
Go Clear
Display Summary Save Text Order Details Add to Clipboard
Items 1-20 of 4330
Page 1 of 217
Select page: 1 2 3 4 5 6 7 8 9 10 >>
Related Articles
100%

Surgery for small asymptomatic abdominal aortic aneurysms.

Endoluminal grafting of abdominal aortic aneurysms: experience with the Talent endoluminal stent graft.

Chromosomal location of AFLP markers in common wheat utilizing nulli-tetrasomic stocks.

Get keywords associated to a set of abstracts

Provide the name of a file containing a set of abstracts:

I Browse... (file size < 500K)

Input format is endnote medline

Sort keywords by significance frequency alphabetical order

Display selected keywords

with significance >= 0.80 and with frequency >= 0.10

Get keywords Reset

Related information and Glossary of terms

Comments? Questions? Problems? Send e-mail to andrade@embl-heidelberg.de

[Andrade Home] [Bork Group]

Inflammatory, cytokine, metalloproteinase, ruptured, smooth, inhibitor, repair, collagen, muscle, patients ...

Fig. 5. Use of the server for analyzing a set of abstracts. In this example, the words composing the name of a disease plus 'protein' were used for a search in PubMed. The resulting set of abstracts was saved in MEDLINE format and the file was used as input in the server which gave back words highly used in the submitted abstracts and possibly relevant to the research on the protein functionality related to the disease. The server is accessible from the web address http://www.embl-heidelberg.de/~andrade/papers/disease_kw/.

calization, assign, link, clon-, candidate, map, mutation, mutant, screen, polymorphi-, deletion, allele). Then, we computed the fraction of papers linked in OMIM entries containing any of those words in their title. It was already possible to classify those journals into two groups. From that classification we derived the words present in the titles that better discriminated between the two groups of journals. The resulting list of words or wordstems (genetics, locus, marker, chromosome, localization, assign, link, clon-, candidate, map, mutation, mutant, screen, polymorphi-, deletion, allele) was very similar to the list chosen initially. This automatically derived list was used for re-classifying the journals and a set of journals was classified as belonging to the domain of genetics (Fig. 2). Papers from these journals were not further considered. In order to be statistically significant, we restricted the next step of our analysis to OMIM entries with 10 or more links to the remaining journals (721 entries linked to 28 177 references).

4.2. *Distribution of words related to OMIM entries*

In order to recognize potential keywords for the description of a disease, we compared the word usage in the abstracts linked to the 721 OMIM entries selected above. For each word, we recorded the fraction of abstracts linked to an entry containing at least once that word. Then we counted how many entries felt in a given range of fraction values and the list of values (for fraction values from zero to one) constituted the distribution of fraction values for the word (see one example for the word 'diabetes' in Fig. 3). See some examples of distributions in Fig. 4. We defined as potential keywords those words for which the distribution presented at least one minimum. Infrequent words that appear randomly distributed across diseases have continuously decreasing distributions, i.e. no threshold could be applied that would distinguish between a set of diseases and the rest[3]. The analysis was restricted to words present in at least 10 abstracts (10 624 different words). Of them, 2247 distributions were found to have at least one minimum, and the corresponding words were annotated as potential keywords.

4.3. *Selection of keywords for OMIM entries*

Given the set of abstracts linked to an OMIM entry, we 'selected' keywords by choosing those 'potential' keywords that were present in a fraction of abstracts higher than the first minimum in their usage distribution over the 721 OMIM entries analyzed (see Fig. 3). An additional measure (more informative than that of fraction of usage) is the fraction of the OMIM entries analyzed for which the frequency of the

[3] We may not dare to say that they follow an extreme value distribution, but this rationale has been already followed by others in order to derive *P* values associated to word frequencies (e.g. Sawted [29]).

word usage was identical or lower (i.e. with a certain significance).

Potential and selected keywords derived for each OMIM entry can be retrieved through a web server accessible from the web address http://www.embl-heidelberg.de/∼andrade/papers/disease_kw/. The server offers links to sentences containing the keywords, and to the corresponding MEDLINE entries (through PubMed).

Taken as an example, the automatic analysis of OMIM entry 100 070 (abdominal aortic aneurysm) mostly confirmed the keyword selection by human annotators. Main 'selected' keywords refer to the abbreviation of the disease 'aaa' or to the name of the disease itself. However, 'ruptured' (only mentioned once in the original entry) appears as something not obvious. Examination of the context shows that it is associated to the name of the disease: 'ruptured aaa'.

4.4. Updating the information: extraction of keywords from a new set of abstracts

The previous analysis allows us to select keywords for an extended set of abstracts related to a disease. Following the example used above, we will update our knowledge on the abdominal aortic aneurysm. A search in MEDLINE using those three terms indicates 8340 papers (as of 27th April 2000). Restricting the search by including the term 'protein' (since we are interested in the phenotype of the disease) and to papers published only in 1998 and later (since we want an update) gives a more handy number of 94 papers (see the example in the web pages). Some interesting potential keywords appear with a high significance. For example, the keyword 'metalloproteinases' has a usage fraction of only 0.13, but a significance of 0.9972. If we examine the distribution of this word, we can see that only three of the 721 original OMIM entries analyzed contained the word in more than 10% of the abstracts. Indeed, increments in the concentration of some metalloproteinases in the aorta have been recently reported to produce a reduction of the elastin concentration and the subsequent degradation of the extracellular matrix, which characterize the disease. Other potential keywords that appear slightly above the 10% cut-off are 'cytokine' and 'inhibition'. The levels of cytokine seem to be altered in patients with the disease but the mechanism by which this happens is unknown; 'inhibition' is used in reference to protease inhibitors and its high frequency reflects the strategy followed by the researchers in this field for curing the disease.

Note that neither 'cytokine' not 'metalloproteinase' were referred to in the original OMIM entry. With this simple approach, we have selected abstracts containing information that can be used for updating OMIM entries. Furthermore, the set of keywords derived constitutes an objective summary of key literature for a certain disease, which might guide database curation.

We have made publicly available the use of this algorithm through a web server. A file with abstracts can be submitted to the server for the automated extraction of keywords related to disease (see Fig. 5).

5. Conclusion

Tools for data mining are becoming increasingly important in biology and medicine because of the growth of the related scientific knowledge. As a consequence, databases and scientific literature have to be integrated and information therein has to be filtered and categorized. Information extraction will become an important part of bioinformatics, despite the fact that current applications to molecular biology are still very preliminary.

Acknowledgements: Thanks to David Torrents for many fruitful discussions during the preparation of the test case and of this review. Thanks also to the maintainers of the OMIM and MEDLINE databases. This work has been supported by grants from the BMBF.

References

[1] Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1998) Bioinformatics 14, 656–664.

[2] Fleischmann, W., Möller, S., Gateau, A. and Apweiler, R. (1999) Bioinformatics 15, 228–233.

[3] Bryson, B. (1990) Mother Tongue, Penguin Books Ltd, London.

[4] Salton, G. (1989) Automatic Text Processing. Addison-Wesley series in Computer Science, Addison-Wesley, Reading, MA.

[5] Barnbrook, G. (1996) Language and Computers, Edinburgh University Press, Edinburgh.

[6] Lewis, D. and Jones, K. (1996) Commun. ACM 39, 92–101.

[7] Cowie, J. and Lehnert, W. (1996) Commun. ACM 39, 80–91.

[8] Gaizauskas, R. (1998) J. Doc. 54, 70–105.

[9] Smeaton, A. (1997) in: Information Extraction. A Multidisciplinary Approach to an Emerging Information Technology (Pazienza, M., Ed.), Lecture Notes in Computer Science Vol. 1299, pp. 115–138, Springer, Berlin.

[10] Ulmschneider, J.E. and Doszkocs, T. (1983) Online Rev. 7, 301–315.

[11] Schulze-Kremer, S. (1998) Pac. Symp. Biocomput. 3, 693–704.

[12] The Gene Ontology Consortium (2000) Nat. Genet. 25, 25–29.

[13] McCune, B.P., Tong, R.M., Dean, J.S. and Shapiro, D.G. (1985) IEEE Trans. Softw. Eng. SE-11, 939–945.

[14] Paijmans, H. (1993) J. Am. Soc. Inf. Sci. 44, 383–392.

[15] Mauldin, M.L. (1991) Conceptual Information Retrieval. A Case Study in Adaptive Partial Parsing, Kluwer Academic Publishers, Boston, MA.

[16] Oltmans, J.A.E. (2000) PhD thesis, Centre for Telematics and Information Technology Enschede, Enschede, Ph.D.-thesis Series No. 99-27.

[17] Hersh, W.R., Evans, D.A., Monarch, I.A., Lefferts, R.G., Handerson, S.K. and Gorman, P.N. (1992) Indexing Effectiveness of Linguistic and Non-Linguistic Approaches to Automatic Indexing, Elsevier Science Publishers, Amsterdam.

[18] Ohta, Y., Yamamoto, Y., Okazaki, T., Uchiyama, I. and Takagi, T. (1997) ISMB 5, 218–225.

[19] Proux, D., Rechenmann, F., Julliard, L., Pillet, V. and Jacq, B. (1998) Genome Inform. Workshop 9, 72–80.

[20] Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T. (1998) Pac. Symp. Biocomput. 3, 705–716.

[21] Blaschke, C., Andrade, M.A., Ouzounis, C. and Valencia, A. (1999) ISMB 7, 60–67.

[22] Ng, S. and Wong, M. (1999) Genome Inform. Workshop 10,

[23] Humphreys, K., Demetriou, G. and Gaizauskas, R. (2000) Pac. Symp. Biocomput. 5, 502–513.

[24] Andrade, M.A. and Valencia, A. (1998) Bioinformatics 14, 600–607.

[25] Stapley, B. and Benoit, G. (2000) Pac. Symp. Biocomput. 5, 526–537.

[26] Rindflesch, T.C., Tanabe, L., Weinstein, J.N. and Hunter, L. (2000) Pac. Symp. Biocomput. 5, 514–525.

[27] Hishiki, T., Collier, N., Nobata, C., Okazaki-Ohta, T., Ogata, N., Sekimizu, T., Steiner, R., Park, H.S. and Tsujii, J. (1998) Genome Inform. Workshop 9, 81–90.

[28] McKusick, V. (1994) Mendelian Inheritance in Man, Catalog of Human Genes and Genetic Disorders, Johns Hopkins University Press, Baltimore, MD.

[29] MacCallum, R.M., Kelley, L.A. and Sternberg, M.J.E. (2000) Bioinformatics, in press.

[30] Thomas, J., Milward, D., Ouzounis, C., Pulman, S. and Carroll, M. (2000) Pac. Symp. Biocomput. 5, 538–549.