ABSTRACT

The increase of information in biology makes it difficult for researchers in any field to keep current with the literature. The MEDLINE database of scientific abstracts can be quickly scanned using electronic mechanisms. Potentially interesting abstracts can be selected by matching words joined by Boolean operators. However, this means of selecting documents is not optimal. Nonspecific queries have to be effected, resulting in large numbers of irrelevant abstracts that have to be manually scanned. To facilitate this analysis, we have developed a system that compiles a summary of subjects and related documents on the results of a MEDLINE query. For this, we have applied a fuzzy binary relation formalism that deduces relations between words present in a set of abstracts preprocessed with a standard grammatical tagger. Those relations are used to derive ensembles of related words and their associated subsets of abstracts. The algorithm can be used publicly at http://www.bork.embl-heidelberg.de/xplormed/.

INTRODUCTION

In molecular biology, there is an accelerated evolution taking place of both the objects of study and the terminology used to describe them, accompanied by increasing specialization and interrelation of fields, resulting in a growth in the amount of research papers published (1,5). Researchers in the field are often compelled to analyze the scientific literature to synchronize one’s own research with the current state of knowledge. Selection of papers relevant to a particular subject is usually done by electronic query on MEDLINE, a database of scientific literature references including abstracts, via several Web servers. The field of information or document retrieval deals with this task (10,11).

Typically, the querying mechanism consists of retrieving the documents matching a series of words joined by Boolean expressions. This way of retrieving documents is simple, though insufficient. In practice, the users have to do nonspecific queries producing large amounts of papers that hide the relevant ones. Indeed, the results of a search are displayed as a list of papers without an overview of the results, and the users have to painfully examine the abstracts of the papers one by one. Even worse, they will just quickly browse through the list of titles.

To alleviate this problem, we propose a protocol that digests the results of a query in MEDLINE and builds a summary of the more relevant terms and the relations between them (i.e., a thesaurus). This gives an overview of the subjects dealt within the results of the query and allows the selection of subsets of papers related to a subject in one or several iterations.

Although there are other applications for the analysis of sets of MEDLINE abstracts [using word frequency (2) or machine learning (3)], they are not targeted to improving document retrieval but to concept discovery via large-scale analysis. Here we are restricted to the words contained in a collection of documents [also called local context analysis (12)].

There are methods for automatically building thesauri that use statistical measurements of word co-occurrence (4,7,9). However, associations between words can have a distinct semantic nature that cannot be grasped with a simple measure of co-occurrence. For that reason, we chose the model for building a fuzzy pseudo-thesaurus described by Miyamoto (6) that is better suited to handle information transmitted through natural language (13).

The system has been implemented as a Web server, XplorMed, which has been described elsewhere (8). Here we focus on the algorithm behind the system, and we show the performance of the system with a detailed example and a benchmark.

MATERIALS AND METHODS

Initial Data and Preprocess

The collection of documents that is used as input by our system is the set of abstracts result of a MEDLINE query (Figure 1, Step 1). Our analysis is restricted to nouns (extracted with a pub-
licly available grammatical tagger that performs a part-of-speech text annotation; TreeTagger, Helmut Schmid, IMS, Stuttgart University, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html). In addition, other noninformative words are also discarded (e.g., units of measure).

Once we have derived the set of words in each abstract, we try to describe the semantic relations between them using fuzzy binary relations (FBRs) (13), which allow description of the strength of the association between two elements.

Degree of Relatedness and Degree of Inclusion between Two Words

An FBR \( R_w \) in a set \( W \) is defined on the Cartesian product \( W \times W \) where the pairs \((x,y)\) may have varying degrees of membership \( \mu_{R_w}(x,y) \) within the relation; that is, it is a fuzzy set, \( R_w = \{(x,y), \mu_{R_w}(x,y)\}, (x,y) \in W \times W \). Let \( Q \) be the set of abstracts to be analyzed. We denote by \( W \) the set of all the words present in \( Q \). Adapting the model of Miyamoto (6) for building a fuzzy pseudo-thesaurus, we define two FBRs in \( W \times W \), \( S_w \) and \( I_w \).

\( S_w \) is the degree of relatedness between two words. We will consider that two words of \( W \), \( w_i \) and \( w_j \), are highly related in the particular context of \( Q \) if they tend to appear very often in the same abstract (e.g., “cell” and “cycle”, which have independent meanings but could be used together in one context to form the more specific concept “cell cycle”). The membership function of \( S_{w_i}(w_i,w_j) \) is estimated by the ratio of the number of abstracts where \( w_i \) and \( w_j \) co-occur and the total number of abstracts where either \( w_i \) or \( w_j \) occur,

\[
\mu_{S_w}(w_i,w_j) = \frac{|W_i \cap W_j|}{|W_i \cup W_j|}
\]

where \(|W_k|\) denotes the cardinality of the subset of \( Q \) where \( w_k \) occurs.

\( I_w \) is the degree of inclusion of one word into another. It expresses the fact that words related to general concepts might include other less general words (e.g., “kinase” can be modified by “aspartate”, “casein”, and “protein”, forming “aspartate kinase”, “casein kinase”,

![Diagram](https://example.com/xplorMedDiagram.png)

**Figure 1. XplorMed procedure.** The boxes depict the steps of the procedure and the ellipses the actions that the user can take. A MEDLINE search produces a set of abstracts (cylinder). Step 1: this set is used as input to the system. Step 2: the system selects words from the abstracts ordered by the strength of their association to other words. Step 3: the selected words are joined into classes of associated words. Step 4: one or more word classes can be used to select a subset of related abstracts (smaller cylinder). The new selection can be used for a new round of analysis closing one iteration cycle.
and “protein kinase”, respectively). The value of \( \mu_{w_i}(w_i, w_j) \) is estimated by the ratio of the number of abstracts where \( w_i \) and \( w_j \) co-occur and the total number of abstracts where \( w_i \) occurs,

\[
\mu_{w_i}(w_i, w_j) = \frac{|W_i \cap W_j|}{|W_i|}
\]

**Keyword Detection**

We can identify words relevant to a collection of abstracts (keywords) because they are likely to establish many and strong relations to other words. To measure this relevance, we define a score for each word \( w_i \), equal to

\[
K_i = \sum_{j \neq i} \mu_{w_i}(w_i, w_j)
\]

normalized to the maximum. The words \( w_i \) with higher scores are assumed to be the keywords.

We consider only the pairs of words whose \( \mu_{w_i}(w_i, w_j) \) is larger than a threshold \( \alpha \) that can be properly varied \((K_w \text{ being either } I_w \text{ or } S_w)\). Such subsets of pairs are called the \( \alpha \)-cuts of the FBR. The remaining network of relations (Figure 2) is a set of overlapping classes of words that are semantically related (e.g., cell\( \rightarrow \)kinase\( \rightarrow \)tyrosine or cell\( \rightarrow \)cancer\( \rightarrow \)breast). For each selected word, we compute one class of words as the chain of words of the path that can be constructed from the maximal acyclic graph spanned by the inclusion relation (Figure 1, Step 3).

**Selection of the Subset of Abstracts Related to a Class of Words**

Given a class of words, the subset of abstracts related could be extracted from \( Q \) with the simple but strict criteria of recovering those abstracts that contain the words belonging to the class. This would result in retrieval with a poor recall: related abstracts would be missed if they do not contain any exact term of the class but a synonym, abbreviation, or other related concepts.

To improve the recall of the retrieval, all words related to the narrowest word of the class (that with the lowest \( K \)-score) with an inclusion value above a given threshold are added to the word class. In the example of Figure 2, the word class formed by “cell\( \rightarrow \)cancer\( \rightarrow \)breast” could be extended with the terms “bcr” and “estradiol”. Then, an abstract could be selected by this word class even if “breast” is not mentioned but the abbreviation “bcr” is used (for “BCR”, which in this context stands for “breast cancer receptor”).

The abstracts are scored according to the presence of the words of the class (Figure 1, Step 4). The selection of abstracts with best scores can be used for a new analysis by the system, beginning an iteration cycle at the keyword computation step (Figure 1, Step 2).

**Table 1. Main Words Associated with the MIP-99 Query**

<table>
<thead>
<tr>
<th>( K )</th>
<th>word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>mip</td>
</tr>
<tr>
<td>0.602</td>
<td>cell</td>
</tr>
<tr>
<td>0.317</td>
<td>protein</td>
</tr>
<tr>
<td>0.221</td>
<td>chemokine</td>
</tr>
<tr>
<td>0.184</td>
<td>1alpha</td>
</tr>
<tr>
<td>0.105</td>
<td>receptor</td>
</tr>
<tr>
<td>0.101</td>
<td>expression</td>
</tr>
<tr>
<td>0.100</td>
<td>macrophage</td>
</tr>
<tr>
<td>0.074</td>
<td>response</td>
</tr>
<tr>
<td>0.074</td>
<td>il</td>
</tr>
<tr>
<td>0.068</td>
<td>hiv</td>
</tr>
<tr>
<td>0.061</td>
<td>patients</td>
</tr>
<tr>
<td>0.056</td>
<td>alpha</td>
</tr>
<tr>
<td>0.053</td>
<td>lung</td>
</tr>
<tr>
<td>0.053</td>
<td>rantes</td>
</tr>
</tbody>
</table>

\( K \) is the association score of the word.

**Table 2. List of Meanings of MIP Found in MEDLINE**

- Macrophage infectivity potenciator
- Macrophage inflammatory protein
- Major intrinsic proteins
- Maternally inherited fragile germanium
- Maximal inspiratory pressure
- Maximum intensity projection
- Mechanically induced potentials
- Medial intraparietal area
- Melation induced premeiotically
- Microwave induced plasma
- Mitochondrial DNA polymerase
- Mitochondrial intermediate peptidase
- Moleculan implanted polymers
- Monos-isopropylated
- Mouse 1,4,5-insolubilins phosphate
- Mytillus inhibitory peptide
- Preconditioning Metabolic inhibition
- mip@xxx.edu

1382 BioTechniques  
Vol. 32, No. 6 (2002)
RESULTS

In this section we illustrate the application of the procedure to the selection of papers about a protein family from a nonspecific query. Then, we report a test of the performance of the system.

Test Case: Analysis of “MIP”

Assume that a researcher comes across the term “major intrinsic proteins” during a study on water channels and wants to know more about those proteins. “MIP” is the commonly used abbreviation to address them. A simple MEDLINE search using the NCBI’s Entrez server (http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi) with “MIP [tw]” produced more than 3000 references. For practical reasons, the analysis was limited to the 325 abstracts of the papers published during the year 1999 and annotated by PubMed with the MeSH category “Chemical & Drugs”.

The most important terms were selected according to an association score $K > 0.05$ (Table 1). These are not related to the expected context: “water channels”. The reason is that the refer-

Table 3. Word Classes More Abundant in the Query MIP-99

<table>
<thead>
<tr>
<th>$n$</th>
<th>word class</th>
</tr>
</thead>
<tbody>
<tr>
<td>153</td>
<td>mip $\rightarrow$ cell</td>
</tr>
<tr>
<td>126</td>
<td>mip $\rightarrow$ protein</td>
</tr>
<tr>
<td>118</td>
<td>mip $\rightarrow$ alpha</td>
</tr>
<tr>
<td>111</td>
<td>mip $\rightarrow$ chemokine</td>
</tr>
<tr>
<td>94</td>
<td>mip $\rightarrow$ macrophage</td>
</tr>
<tr>
<td>77</td>
<td>mip $\rightarrow$ expression</td>
</tr>
<tr>
<td>64</td>
<td>mip $\rightarrow$ il</td>
</tr>
<tr>
<td>57</td>
<td>mip $\rightarrow$ level</td>
</tr>
<tr>
<td>56</td>
<td>mip $\rightarrow$ alpha</td>
</tr>
<tr>
<td>56</td>
<td>mip $\rightarrow$ rantes</td>
</tr>
<tr>
<td>52</td>
<td>mip $\rightarrow$ cell $\rightarrow$ t</td>
</tr>
<tr>
<td>50</td>
<td>mip $\rightarrow$ production</td>
</tr>
</tbody>
</table>

... 4 mip $\rightarrow$ protein $\rightarrow$ aquaporin

$n$ indicates the number of abstracts containing all the words of the word class.

Table 4. Evaluation of the Performance of XplorMed on 30 Papers

| PMID | title | N(med) | N(med) | N(med) | N(med) | N(med) | N(Xpl) | N(Xpl) | N(Xpl) | R(med) | R(med) | R(med) | R(Xpl) | R(Xpl) | R(Xpl) | R(Xpl) | R(Xpl) |
|------|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 10415 | 1. Control of endodermal endodermal development by Hoxc1 | 58.00 | 1.50 | 0.50 | 0.50 | 0.60 | 0.40 | 0.40 | 0.40 | 0.40 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 |
| 10416 | 1. Control of endodermal endodermal development by Hoxc1 | 58.00 | 1.50 | 0.50 | 0.50 | 0.60 | 0.40 | 0.40 | 0.40 | 0.40 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 |

PMID is the identifier of the test paper. The words in bold in the title were used for a query in MEDLINE in the years 1998 and 1999. The reference gives the volume and page of the paper (taken from Nature Genetics, years 2000 or 2001). A word class suggested by XplorMed was picked up that was related to the subject of the paper. N is the number of papers cited in the test paper (the reference set), N(med) is the number of papers in the search in MEDLINE (the medline set), and N(Xpl) is the number of papers remaining after selection by one XplorMed word class (the XplorMed set), with the numbers of papers found in the reference set between brackets. R(med) and R(Xpl) refer to the recall with respect to the reference set by the medline and the XplorMed sets, respectively. In general, it was intuitive to find an XplorMed word class producing an increment in the precision.
ences retrieved deal with several subjects that correspond to different meanings of MIP (Table 2). The word classes resulting from a 0.75 α-cut on the inclusion relation $T_w$ are displayed in Table 3. At this point, manual intervention was needed to select for the next step a word class that agreed with the subject “water channel”. In particular, “mip→protein→aquaporin” was chosen. Only four abstracts contained the three words. However, a total of 13 abstracts gave a significant score using the extended word class (i.e., adding the words that are included on the narrower member of the class, “aquaporin”; see Materials and Methods). Nine of those 13 abstracts were dealing with “Major Intrinsic Proteins” (the first false positive being ranked in the eighth place). Manual check of the unselected abstracts indicated that none of them was referring to these proteins (no false negative). The iteration of the procedure with these 13 abstracts produces a new set of words that are more precise as keywords for the protein family than “MIP”, such as “channel”.

**Benchmark**

We performed a benchmark of the system to evaluate its support to a manual process of literature retrieval. We contrasted the bibliography referenced in a series of papers with both the abstracts obtained by manual search in MEDLINE and the subsequent selection done using XplorMed. As test items we chose 30 papers from recent issues of *Nature Genetics* with eight or more references to papers published during the years 1998 and 1999. Those references are the reference set of each test item. From the title of each test item, we chose a set of words to perform a keyword search on the MEDLINE database limited to the years 1998 and 1999. The selected papers constitute the medline set (containing at least a 10% of the reference set). The medline set of each test item was used as input to XplorMed. A word class computed by XplorMed that was in agreement to the subject of the test item was used to select a smaller set of abstracts (the XplorMed set).

The comparisons were done by means of recall and precision measurements with respect to the reference set. Recall is defined as the fraction of the elements of a set that was present in the reference set. Precision is defined as the fraction of the elements of a set that was present in the reference set. The results are detailed in Table 4. The recall with respect to the reference set of a search in MEDLINE was on average 0.375 (standard deviation = 0.209). This low recall is not surprising given the heterogeneous nature of the bibliography that usually includes methods and very general papers that may not be strictly related to one particular subject. What is remarkable is that a further selection using XplorMed on this set did not reduce dramatically the recall (average 0.282, standard deviation = 0.161), and produced a significant improvement in the average precision from 0.063 (standard deviation = 0.063) to 0.136 (standard deviation = 0.156).

**CONCLUSION**

The analysis of relations between words including dependencies is very appropriate for the detection of words with a relevant meaning in a collection of documents. Nevertheless, this relevance depends on the interest of the person doing the analysis. Therefore, we have chosen an approach that guides a process of document retrieval.

The possibility of selecting sets of words from a list and different α-cuts of the FRBs derived for them makes the procedure very flexible for the user. Different levels of description may be desirable sometimes and can be controlled by varying only two parameters (α and $K$). The benchmark indicates that the selection via XplorMed helps to find sets of abstracts focused on a subject.

The system proposed here has some obvious limitations. The user does not have to be an expert on the subject of research but should generate a collection of abstracts with not too many unrelated subjects on it. In this respect, the example that we showed of the MIP query is an extreme case.

**ACKNOWLEDGMENTS**

The authors are indebted to Helmut Schmid (Institut für Maschinelle Sprachverarbeitung, Stuttgart University) for developing TreeTagger and making it freely available to the scientific community.

**REFERENCES**


Received 21 November 2001; accepted 15 March 2002.

Address correspondence to:
Dr. Carolina Perez-Iratxeta
European Molecular Biology Laboratory
Meyerhofstr. 1, Heidelberg 69002, Germany
e-mail: cperez@embl-heidelberg.de

BioTechniques 1385