Research Report

Computing Fuzzy Associations for the Analysis of Biological Literature

BioTechniques 32:1380-1385 (June 2002)

Carolina Perez-Iratxeta[1,2], Harindar S. Keer[3], Peer Bork[1,2], and Miguel A. Andrade[1,2]

[1]European Molecular Biology Laboratory, Heidelberg, Germany, [2]Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany, [3]Indian Institute of Technology Kanpur, India

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.unistuttgart.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 $\tilde{R}_w$ in a set W is defined on the Cartesian product $W \times W$ where the pairs (x,y) may have varying degrees of membership $\mu_{\tilde{R}_w}(x,y)$ within the relation; that is, it is a fuzzy set, $\tilde{R}_w = \{((x,y), \mu_{\tilde{R}_w}(x,y)), (x,y) \in W \times W\}$. 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 $W \times W$, $\tilde{S}_w$ and $\tilde{I}_w$.

$\tilde{S}_w$ 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 $\tilde{S}_w$, $\mu_{\tilde{S}_w}(w_i,w_j)$ 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,

$$\mu_{\tilde{S}_w}(w_i, w_j) = \frac{|W_i \cap W_j|}{|W_i \cup W_j|}$$

where $|W_k|$ denotes the cardinality of the subset of Q where w_k occurs.

$\tilde{I}_w$ 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",

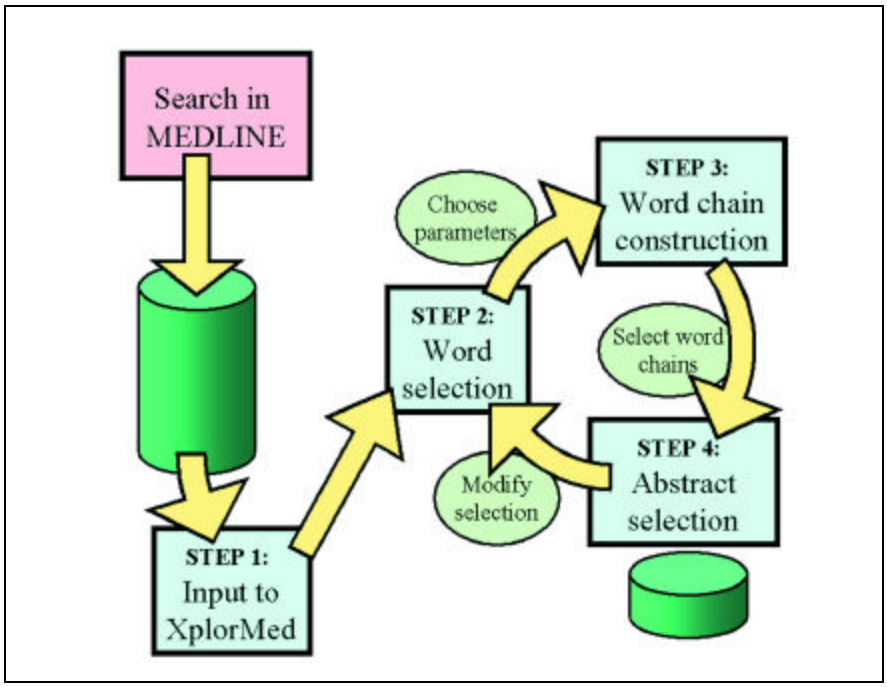

Figure 1. *XplorMed* procedure. The boxes depict the steps of the procedure and the ellipses the actions that the user can take. A MEDLINE search produces a set of abstracts (cylinder). Step 1: this set is used as input to the system. Step 2: the system selects words from the abstracts ordered by the 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 $\mu_{\tilde{I}_w}(w_i, w_j)$ 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,

$$\mu_{\tilde{I}_w}(w_i, w_j) = \frac{|W_i \cap W_j|}{|W_i|}$$

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

$$K_i = \sum_{j \neq i} \mu_{\tilde{I}_w}(w_i, w_j)$$

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 $\mu_{\tilde{R}_w}(w_i, w_j)$ is larger than a threshold α that can be properly varied ($\tilde{R}_w$ being either $\tilde{I}_w$ or $\tilde{S}_w$). Such subsets of pairs are called the α-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→kinase→tyrosine or cell→cancer→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 criterion 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→cancer→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

K	word
1.000	mip
0.602	cell
0.317	protein
0.221	chemokine
0.184	1alpha
0.105	receptor
0.101	expression
0.100	macrophage
0.074	response
0.074	il
0.068	hiv
0.061	patients
0.056	alpha
0.053	lung
0.053	rantes

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 permutation
Maximal inspiratory pressure
Maximum intensity projection
Mechanically induced potentials
Medial intraparietal area
Metilation induced premeiotically
Microwave induced plasma
Mitochondrial DNA polymerase
Mitochondrial intermediate peptidase
Molecularly imprinted polymers
Mono-isopropylated
Mouse 1,4,5-inositol phosphate
Mytillus inhibitory peptide
Preconditioning Metabolic inhibition
mip@xxx.edu

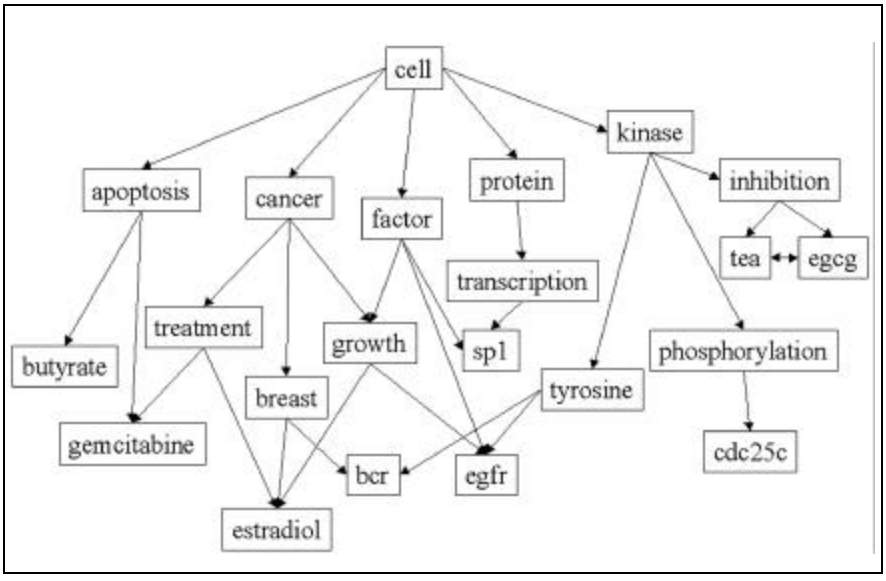

Figure 2. Part of the net of dependency relations derived after the analysis of a set of abstracts selected from MEDLINE that contained all the words "protein", "kinase", and "cancer". The word found at the top of the hierarchy was "cell". Chemical objects such as drugs, genes, and proteins are found at the bottom of the hierarchy. From left to right: sodium butyrate and gemcitabine induce cell apoptosis; estradiol is related to the inducement of breast cancer; BCR is breast cancer receptor; EGFR is epidermal growth factor receptor; both BCR and EGFR have tyrosine kinase activity; Sp1 is a transcription factor; Cdc25C is a tyrosine phosphatase; EGCG is epigallocatechin-3-gallate, a tea polyphenol that inhibits MAP kinase.

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

n	word class
153	mip → cell
126	mip → protein
118	mip → 1alpha
111	mip → chemokine
94	mip → macrophage
77	mip → expression
64	mip → il
61	mip → level
57	mip → alpha
56	mip → rantes
56	mip → mcp
52	mip → cell → t
50	mip → production
...	
4	mip → protein → aquaporin

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

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

PMID	title	ref	word class	N	N(med)	N(xpl)	R(med)	R(Xpl)	P(med)	P(Xpl)
10615123	MLH3: a **DNA mismatch repair gene** associated with mammalian microsatellite instability.	24:27	gene, tumour	17	334(8)	209(8)	0.471	0.471	0.024	0.027
10615124	Control of **endodermal** endonina **development** by Hes-1.	24:35	cell, embryo	11	83(2)	65(2)	0.182	0.182	0.024	0.031
10615127	Novel dominant mutations in Saccharomyces cerevisiae **MSH6**.	24:53	mismatch, repair, dna	14	63(10)	46(6)	0.714	0.571	0.159	0.167
10615134	Molecular mechanism for duplication 17p11.2- the homologous recombination red procal of the **Smith-Magenis** microdeletion.	24:84	smith, syndrome, sms, chromosome	11	24(2)	13(2)	0.182	0.182	0.083	0.154
10615135	DNA methyltransferase Dnmt1 associates with **histone deacetylase** activity.	24:88	histone, acetylation	17	311(13)	61(3)	0.765	0.176	0.042	0.049
10742094	**Mutations** of SCN1A, encoding a neuronal **sodium channel**, in two families with GEFS+2.	24:343	mutation, channal, gene	10	166(3)	161(3)	0.300	0.300	0.018	0.019
10742095	NPHS2, encoding the **glomerular protein** podocin, is mutated in autosomal recessive steroid-resistant nephrotic **syndrome**.	24:349	protein, slit, nephrin	16	95(4)	6(4)	0.250	0.250	0.042	0.667
10742099	Mutations in **ATRX**, encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation.	24:368	mutation, atrx	9	9(2)	4(1)	0.222	0.111	0.222	0.250
10742110	The SH2 tyrosine phosphatase **shp2** is required for mammalian limb development.	24:420	protein, kinase, factor, growth	10	35(3)	13(2)	0.300	0.200	0.086	0.154
11175783	The putative forkhead transcription factor FOXL2 is mutated in **blepharophimosis**/ptosis/epicanthus inversus syndrome.	27:159	maturation, pool, follicle	14	38(3)	7(1)	0.214	0.071	0.079	0.143
11175785	A candidate **prostate cancer susceptibility** gene at chromosome 17p.	27:172	prostate, cancer, gene	10	66(5)	61(5)	0.500	0.500	0.076	0.082
11175795	Pten and p27KIP1 cooperate in **prostate** cancer **tumor suppression** in the mouse.	27:222	cancer, prostate, cell, gene	9	7(6)	5(1)	0.111	0.111	0.013	0.018
11242108	**Recombination**al **DNA** double-strand breaks in mice precede **synapsis**.	27:271	recombination, site	13	22(3)	15(2)	0.231	0.154	0.136	0.133
11242109	X-linked anhidrotic **ectodermal dysplasia** with immunodeficiency is caused by impaired NF-kappaB signaling.	27:277	gene, x	15	146(3)	9(2)	0.200	0.133	0.021	0.222
11242110	**DiGeorge syndrome** phenotype in mice mutant for the T-box gene, Tbx1.	27:286	syndrome, gene	8	152(5)	34(4)	0.625	0.500	0.033	0.118
11242115	Constitutively activating mutation in **WASP** causes X-linked severe congenital neutropenia.	27:313	protein, domain	10	274(9)	34(2)	0.900	0.200	0.033	0.059
11242118	Deficiency of methyl-CpG binding protein-2 in CNS neurons results in a **Rett**-like phenotype in mice.	27:327	syndrome, patients	9	116(5)	85(4)	0.556	0.444	0.043	0.047
11279519	Sequence diversity in **CYP3A** promoters and characterization of the genetic basis of polymorphic CYP3A5 expression.	27:383	liver, microsomes	13	324(2)	179(2)	0.154	0.154	0.006	0.011
11279522	**Trinucleotide expansion** in haploid germ cells by gap repair.	27:407	repeat, expansion	13	454(8)	333(8)	0.615	0.615	0.018	0.024
11279529	TSLC1 is a tumor-suppressor gene in human non-small-cell **lung cancer**.	27:427	suppressor, tumour, cancer, protein	8	228(4)	180(3)	0.500	0.375	0.018	0.017
11326274	First genetic evidence of GABA(A) receptor dysfunction in **epilepsy**: a **mutation** in the gamma2-subunit gene.	28:46	channel, potassium	12	157(3)	27(3)	0.250	0.250	0.019	0.111
11326276	Beta-catenin-sensitive isoforms of **lymphoid enhancer factor**-1 are selectively expressed in colon cancer.	28:53	factor, catenin	9	36(1)	36(1)	0.111	0.111	0.028	0.028
11326279	The gene defective in **leukocyte adhesion deficiency** II encodes a putative GDP-fucose transporter.	28:69	gdp, fucose	11	65(6)	12(6)	0.545	0.545	0.092	0.500
11381259	Deletion of the **hypoxia**-response element in the **vascular endothelial growth factor** promoter causes motor neuron degeneration.	28:131	growth, factor, angiogenesis	18	204(5)	102(5)	0.278	0.278	0.025	0.036
11381260	Regulation of the Caenorhabditis **elegans longevity** protein DAF-16 by insulin/IGF-1 and germline signaling.	28:139	life, span	15	38(5)	27(5)	0.333	0.333	0.132	0.185
11381262	Mutant glycosyltransferase and altered glycosylation of alpha-dystroglycan in the **myodystrophy** mouse.	28:151	gene, region, mapping	10	8(2)	5(2)	0.200	0.200	0.250	0.400
11381264	Rapid gene mapping in Caenorhabditis elegans using a high density **polymorphism map**.	28:160	gene, region	12	261(3)	183(3)	0.250	0.250	0.011	0.016
11431692	Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage **T7** gene 4-like **protein** localized in mitochondria.	28:223	deletion, dna, abnormality	9	337(3)	7(2)	0.333	0.222	0.009	0.286
11431693	The SNRPN promoter is not required for genomic imprinting of the Prader-Willi/**Angelman** domain in mice.	28:232	will, prader	21	126(12)	86(10)	0.571	0.476	0.095	0.116
11431698	Genetic interactions between tumor suppressors **Brca1** and **p53** in apoptosis, cell cycle and tumorigenesis.	28:266	cancer, breast	10	95(4)	71(1)	0.400	0.100	0.042	0.014

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 respect to the reference set by the medline and the *XplorMed* sets, respectively. R(med) is always larger than R(Xpl) because the *XplorMed* set is a selection of the medline set. P(med) and P(Xpl) refer to the precision 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 $\tilde{I}_w$ 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 FBRs 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

1. **Andrade, M.A. and P. Bork.** 2000. Automated extraction of information in molecular biology. FEBS Lett. *476*:12-17.
2. **Andrade, M.A. and A. Valencia.** 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. Bioinformatics *14*:600-607.
3. **Iliopoulos, I., A.J. Enright, and C.A. Ouzounis.** 2001. Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. Pac. Symp. Biocomput. *5*:384-395.
4. **Jing, Y. and W.B. Croft.** An association thesaurus for information retrieval. Recherche d'Information Assistee par Ordinateur; 1994 Oct 11-13; New York. Rockefeller University, New York, p. 146-160.
5. **Lawrence, P.A.** 2001. Science or alchemy. Nat. Rev. Genet. *2*:139-141.
6. **Miyamoto, S.** 1990. Fuzzy sets in information retrieval and cluster analysis. Theory and decision library. Kluwer Academics Publishers, Dordrecht.
7. **Peat, H.J. and P. Willett.** 1991. The limitations of term co-occurrence data or query expansion in document retrieval systems. J. Am. Soc. Inform. Sci. *42*:378-383.
8. **Perez-Iratxeta, C., P. Bork, and M.A. Andrade.** 2001. XplorMed: a tool for exploring MEDLINE abstracts. Trends Biochem. Sci. *26*:573-575.
9. **Qiu, Y. and H.P. Frei.** Proceedings of the SIGiR-93, 16th ACM International Conference on Research and Development in Information Retrieval; 1993 June 27-July 1; Pittsburgh, PA.
10. **Salton, G.** 1963. Associative document retrieval techniques using bibliographic information. J. ACM *10*:440-457.
11. **Salton, G.** 1978. Generation and search of clustered files. ACM T. Database Syst. *3*:321-346.
12. **Xu, J. and W.B. Croft.** 2000. Improving the effectiveness of informational retrieval with local context analysis. ACM T. Inform. Syst. *18*:79-112.
13. **Zimmermann, H.J.** 1996. Fuzzy Set Theory and Its Applications, 3rd ed. Kluwer Academics Publishers, Boston.

Received 21 November 2001; accepted 15 March 2002.

Address correspondence to:

Dr. Carolina Perez-Iratxeta
European Molecular Biology Laboratory
Meyerhofstr. 1, Heidelberg 69012, Germany
e-mail: cperez@embl-heidelberg.de