APPLICATION NOTE

A protocol for the update of references to scientific literature in biological databases

Carolina Perez-Iratxeta,[1,2] Nagore Astola,[1] Francesca D Ciccarelli,[1,2] Parantu K Sha,[1,2] Peer Bork,[1,2] Miguel A Andrade[1,2]

[1]European Molecular Biology Laboratory, Heidelberg, Germany; [2]Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany

Abstract: Entries in biological databases are usually linked to scientific references. To generate those links and to keep them up-to-date, database maintainers have to continuously scan the scientific literature to select references that are relevant for each single database entry. The continuous growth of both the corpus of scientific literature and the size of biological databases makes this task very hard. We present a protocol intended to assist the updating of an existing set of literature (abstract) links from a single database entry with new references. It consists of taking the set of MEDLINE® neighbour references of the existing linked abstracts and evaluating their relevance according to the existing set of abstracts. To test the applicability of the algorithm, we did a simple benchmark of the system using the references associated with the entries of a protein domain database. Human experts found the references that the algorithm scored highly were more relevant to the database entry than those scored lowly, suggesting that the algorithm was useful.

Keywords: information retrieval, database maintenance, MEDLINE

Availability: All the scripts are available by request to the authors and TreeTagger is available at http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/

Contact: Carolina Perez-Iratxeta (cperez@embl.de)

Introduction

The number and size of biological databases has increased enormously in recent years, especially in the field of molecular biology. The rapid generation of data from complete genome sequencing, structural genomics, proteomics, gene expression analysis etc makes online databases indispensable work tools for the researcher. But at the same time, the accelerating pace of data production makes databases difficult to maintain. This difficulty concerns not only the incorporation of new objects, but also the continuous inclusion of new references to the scientific literature regarding the material already present in the database. The latter is an information retrieval problem (Salton 1963, 1978).

Normally, an established database entry is already linked to a small set of references to the literature, often manually selected by dedicated curators. Here, we propose a method that extracts word associations from the abstracts of the linked articles and uses these to score new references. Such a method can be used in biological databases to support the manual selection of new links to literature.

The first step consists of retrieving from the MEDLINE® database of scientific literature, via the PubMed® server (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed), the neighbour abstracts of the original hand-picked set of abstracts. The MEDLINE neighbours of an abstract are a pre-computed set of 'near' papers supposed to deal with similar subjects. To define abstracts 'near' to a given one, each abstract is coded as a frequency vector of its words. As a measure of distance between two abstracts, the cosine of the angle formed by the two corresponding word vectors is used (Wilbur and Yang 1996).

Unfortunately, it usually happens that the set of papers obtained by this mechanism is large and unspecific. The most interesting references are buried among irrelevant ones. The protocol we propose is intended to be applied at this stage to sort the results. It can be termed a local context analysis protocol (Xu and Croft 2000).

Description of the protocol

Given an entry of a database to be updated, its set of already linked abstracts is considered (the primary set). First, the neighbours of the primary set are collected (for example,

Correspondence: Carolina Perez-Iratxeta, Ottawa Health Research Institute, 501 Smyth Road, K1H 8L6, Ottawa, ON, Canada; tel +1 613 737 8899 ext 73255; fax +1 613 737 8803; email cperez-iratxeta@ohri.ca

via the PubMed server). This set is filtered by selecting those references that are neighbours of at least two of the abstracts of the primary set. We denote the resulting set as the secondary set.

The secondary set is then scored by relevance to the primary set, in a two-step procedure. The first step consists of computing the keywords and main associations between words in the primary set. The computation is done only over the nouns, since this works well for the classification of scientific text (Perez-Iratxeta et al 2002). Nouns and sentences are detected using a standard part-of-speech tagger, TreeTagger, that can be publicly accessed at http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/.

Keywords are calculated using a simple model of fuzzy binary relations (Miyamoto 1990). Given a collection of abstracts, one word w_i is considered to be highly included in a broader word w_j, if most of the times that w_i occurs, the broader word w_j is also present in the same sentence.

For example, both 'lipid' and 'transfer' can be included in the broader concept of 'protein' in a set of abstracts dealing with proteins and mentioning lipid transfer proteins.

To formalise this, we define a fuzzy set, $\tilde{I}_W$, containing all the possible pairs of words (w_i, w_j) whose membership function $\mu_{\tilde{I}_W}(w_i, w_j)$ is an estimation of the 'degree of the inclusion' of w_i in w_j:

$$\mu_{\tilde{I}_W}(w_i, w_j) = \frac{|W_i \cap W_j|}{|W_i|} \quad (1)$$

In this expression, the numerator stands for the number of sentences in which both w_i and w_j occur, and the denominator for the number of sentences where w_i occurs.

Following Perez-Iratxeta et al (2002), we define keywords as those words having many strong relations with other words. To identify which words constitute keywords, we define a score for each word w_i:

$$K_i = \sum_{j \neq i} \mu_{\tilde{I}_W}(w_j, w_i) \quad (2)$$

This is further normalised by the maximum score obtained, so that the most relevant (or important) word has a K-score of 1.00. Note that only the pairs above a certain cut-off in the value of inclusion are taken into account for the computation of K_i. We select as keywords those words that have a K-score higher than a given threshold.

The second step uses both the keywords and their associated words computed from the primary set for scoring the abstracts of the secondary set. Using the set of selected keywords, one could simply choose the abstracts in the secondary set that match any of the keywords. However, the performance of the document selection can be enhanced if one also considers the presence of other words that are highly related to (that is, included in) the keywords (Perez-Iratxeta et al 2002).

For example, if we are searching the secondary set for abstracts matching the keywords 'lipid', 'transfer' and 'protein', we would also like to retrieve the abstracts mentioning 'LTP' (the common abbreviation for lipid transfer protein), which turns out not to be a keyword, but it is nevertheless a word highly included in several keywords.

Accordingly, we score the abstracts in the secondary set with the sum of the K-scores of all the matched terms (considering keywords and words associated to them above a threshold of inclusion, typically 0.75) normalised by the number of words in the abstract. We propose such a score as a measure of the relevance of an abstract relative to the primary set.

Benchmark on the SMART database

To check the performance of the algorithm in selecting relevant references, we performed a simple benchmark in the SMART database. SMART is a database of protein domains that currently contains more than 600 entries (Letunic et al 2002). We ran our protocol on 20 randomly selected entries (with at least three references to MEDLINE) to test the selection and scoring of a new set of references.

For each of the entries in the sample, we built a small list including the five best and five worst references according to their score. All the lists were shuffled and presented to six human experts, who are all active researchers in the field of molecular biology at the European Molecular Biology Laboratory, Heidelberg (see the list of authors of this manuscript). They were asked to select the five more 'interesting' or 'relevant' papers regarding the corresponding protein domain. The agreement between the experts' opinions and the automatic classification is presented in Table 1.

Conclusions

In this work we have presented a method to score the relevance of a scientific article relative to a primary set of references, by analysis of the abstracts. Our algorithm takes into account the associations between words discovered in the abstracts of the primary reference set and scores other abstracts according to these. The computation of associations between words depends on word co-occurrences within

Table 1 Results of the benchmark on the SMART database entries compared to human experts

SMART database identifier	*Nr of abstracts in primary set*	*Nr of abstracts in secondary set*	*Overlapping nr of experts' selection and protocol selection*[a]
ARF	5	36	3
AT_hook	5	39	3
ETS	4	71	4
FRI	8	59	3
HLH	5	12	3
IQ	5	79	4
KH	6	101	2
KR	5	37	3
KU	4	42	3
LACTALBUMIN_LYSOZYME	4	79	3
NATRIURETIC_PEPTIDE	3	216	5
PLDc	7	165	4
PQQ	3	26	5
PRP	3	216	3
Sec7	4	45	4
Skp1	5	39	3
TNFR	5	299	2
VPS9	5	11	3
ZnF_C4	4	24	5
ZP	5	44	2

[a] Agreement (in number of references) between the five best scored references of the experts' selection and the protocol selection. Probabilities of having such an overlap (or greater) just by chance, according to the hypergeometric random model, are: $p(o=5)=0.0039$, $p(o\geq4)=0.1031$, $p(o\geq3)=0.5$ and $p(o\geq2)=0.8968$. According to the binomial distribution, the expected number of occurrences of an overlap equal to or greater than 3, among 20 independent trials, is 10 [$p(o\geq3)=0.5$]; the probability of its occurrence 17 or more times among 20, as in the benchmark, is 0.00128. Also, to obtain an overlap greater than or equal to 4 is a rather non-probable event (ie $p(o\geq4)=0.1031$). The probability of its occurrence 7 or more times among 20 independently repeated trials is equal to 0.00233. These small values are interpreted as the probability of performing this benchmark as well as or better than this, just by chance.

sentences; then, those words that have the most associations in the primary set are taken into account for scoring an abstract. Those words are not necessarily the most frequent ones.

In contrast, a measure such as the cosine of the word vectors derived from abstracts (Wilbur and Yang 1996) compares word frequencies. An alternative to our approach is to sort the abstracts of the secondary set according to the cosine of the angle of their associated word vector to the average word vector of the abstracts in the primary set. But such a measure cannot grasp the fine structure that the logical exposition of scientific subjects using natural language imposes in the word distribution across abstracts and sentences (see Perez-Iratxeta et al (2002) and references cited within, for discussion on this issue).

However, the use of the cosine metric between word vectors of abstracts requires much less computation than deriving associations between words. Therefore, it is appropriate for dealing with a large number of abstracts; for example, when analysing the millions of abstracts present in the MEDLINE database of scientific literature. The application of our algorithm to sets of more than a thousand abstracts becomes impractical. Accordingly, we have suggested in this paper a way of selecting the secondary set of literature from the primary one out of the whole MEDLINE database, by using a rough but quick mechanism like the cosine distance of word vectors. After that initial step, our algorithm can be used to sort and filter the results.

We note that our algorithm can be applied with all generality to the scoring of abstracts according to a primary set of references, irrespective of the method used to select the abstract to be scored. In this respect, we think that it can be useful to support any updating mechanism of MEDLINE links to any database.

Acknowledgements

We are grateful to Helmut Schmid for developing TreeTagger and making it publicly available.

References

Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P. 2002. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res*, 30:242–4.

Miyamoto S. 1990. Fuzzy sets in information and cluster analysis. Theory and decision library. Dordrecht: Kluwer Academic.

Perez-Iratxeta C, Keer HS, Bork P, Andrade MA. 2002. Computing fuzzy associations for the analysis of biological literature. *Biotechniques*, 32:1380–5.

Salton G. 1963. Associative document retrieval techniques using bibliographic information. *J Assoc Comput Machinery*, 10:440–57.

Salton G. 1978. Generation and search of clustered files. *Assoc Comput Machinery Trans Database Syst*, 3:321–46.

Wilbur WJ, Yang Y. 1996. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput Biol Med*, 26:209–22.

Xu J, Croft WB. 2000. Improving effectiveness of information retrieval with local context analysis. *Assoc Comput Machinery Trans Inf Syst*, 18:79–112.