BIOINFORMATICS

Vol. 14 no. 7 1998
Pages 600–607

Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families

Miguel A. Andrade[1] and Alfonso Valencia

Protein Design Group, CNB-CSIC, Cantoblanco, E-28049 Madrid, Spain

Received on January 13, 1998; revised and accepted on May 13, 1998

Abstract

Motivation: *Annotation of the biological function of different protein sequences is a time-consuming process currently performed by human experts. Genome analysis tools encounter great difficulty in performing this task. Database curators, developers of genome analysis tools and biologists in general could benefit from access to tools able to suggest functional annotations and facilitate access to functional information.*

Approach: *We present here the first prototype of a system for the automatic annotation of protein function. The system is triggered by collections of abstracts related to a given protein, and it is able to extract biological information directly from scientific literature, i.e. MEDLINE abstracts. Relevant keywords are selected by their relative accumulation in comparison with a domain-specific background distribution. Simultaneously, the most representative sentences and MEDLINE abstracts are selected and presented to the end-user. Evolutionary information is considered as a predominant characteristic in the domain of protein function. Our system consequently extracts domain-specific information from the analysis of a set of protein families.*

Results: *The system has been tested with different protein families, of which three examples are discussed in detail here: 'ataxia-telangiectasia associated protein', 'ran GTPase' and 'carbonic anhydrase'. We found generally good correlation between the amount of information provided to the system and the quality of the annotations. Finally, the current limitations and future developments of the system are discussed.*

Availability: *The current system can be considered as a prototype system. As such, it can be accessed as a server at http://columba.ebi.ac.uk:8765/andrade/abx. The system accepts text related to the protein or proteins to be evaluated (optimally, the result of a MEDLINE search by keyword) and the results are returned in the form of Web pages for keywords, sentences and abstracts.*

Supplementary information: *Web pages containing full information on the examples mentioned in the text are available at: http://www.cnb.uam.es/~cnbprot/keywords/*

Contact: *valencia@cnb.uam.es*

Introduction

Scientific knowledge is contained in vast collections of written text. The rapid growth of these collections makes it increasingly difficult for humans to access the required information in a convenient and effective manner.

This task has been approached over the past few years from two different perspectives. On the one hand, text understanding has been based on lexical, syntactical and semantic analysis. This approach is confronted with the variability, fuzziness and complexity of human language. A number of operative methods of language analysis have emerged from this work (see Salton, 1989; Cowie and Lehnert, 1996).

On the other hand, a less ambitious attempt has also turned out to be of practical relevance, namely, the treatment of text with statistical methods. In this approach, the possible relevance of the words in a text is deduced from the comparison of the frequency of different words in this text with the frequency of the same words in reference sets of text [e.g. Berry, *et al*., 1995; or the Experimental Search System (ESS) at the US Library of Congress, http://lcweb2.loc.gov/catalog/]. This approach is clearly domain specific, since the frequency of words varies greatly between different knowledge areas, e.g. technology or poetry.

In the field of molecular biology, in particular for the annotation of protein functions, there is considerable interest in automatization. The many ongoing sequencing projects and the fast growth of the databases clearly demand this type of method. We propose here a simple approach based on word distribution statistics, and discuss the relative success of a first prototype as applied to relevant examples.

We explore the possibility of extracting biologically significant words related to protein function directly from stored text. The source of information used was the MED-

[1]*Present address: European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK*

LINE subset dealing with biological sequences. MEDLINE is a collection of abstracts from scientific journals stored and maintained at the National Library of Medicine (http://www.nlm.nih.gov/). Abstracts in this collection, when associated to proteins, normally include various aspects of protein function, such as biochemistry, cellular function, medical implications, etc. They are composed in general of short sentences of technical character using a reduced and non-ambiguous vocabulary. These characteristics render them very appropriate for statistical approaches. Extraction of functionally related significant words can be considered as the first step in the process of automatic annotation of sequences.

Annotation of biological sequences

Sequences are collected and stored systematically in different sequence repositories. In particular, SwissProt (Bairoch and Apweiler, 1997) contains a large collection of protein sequences with minimal information about their biological function. These annotations are carefully made by human experts with general biological knowledge after consultation of the relevant bibliography. Annotations are made in the form of (i) free text comments on protein function (e.g. 'Pathway: non-oxidative branch of the pentose phosphate pathway') or biological relationships (e.g. 'Similarity: to other bacterial or eukaryotic rpi'), (ii) keywords chosen from a restricted list of choices (a dictionary of 800 words in SwissProt) and (iii) the common name of the protein.

Different uses of database annotations

Database annotations are commonly used by human experts as a first indication of protein function. Perhaps the most frequent applications that rely on database annotations are database similarity searches. A typical similarity search starts with the inspection of the output of common sequence search tools such as BLAST (Altschul *et al*., 1990) and FASTA (Pearson and Lipman, 1988). The first information available is the protein name as described in the sequence database entry, e.g. the DE line in SwissProt: 'rash_human, ras-p21 oncogene'. A more detailed inspection requires manual access to other biological information annotated in the database [e.g. using a retrieval system such as SRS (Etzold *et al*., 1996)]. Finally, deeper inspection of the available information would require the retrieval of the linked MEDLINE abstracts or direct consultation of written bibliography. The successful identification of the putative function of a protein often depends on the first steps of the search and, consequently, on the quality of the database annotations.

In large-scale sequencing projects, the process of annotation is carried out automatically, fulfilling only the first step of the process described above (Casari *et al*., 1995; Gaasterland and Sensen, 1996; Koonin *et al*., 1996). Tools for automatic genome analysis therefore depend completely on the description of protein function provided in the database, and often only on the annotation of the first similar sequence found. New tools for accessing functional information will be important for the annotation of the overwhelming number of sequences derived from different sequencing projects, as well as for the annotation of protein sequences directly translated from DNA sequences (e.g. TREMBL; Bairoch and Apweiler, 1997). The process of database annotation is generally accurate, as it is performed by human experts, but it is time consuming and highly idiosyncratic.

A more systematic exploitation of the database annotations, and in particular of the SwissProt keywords, has also been attempted by Guigó *et al*. (1991) who used them for automatic discovery of new functional relationships between sequence families. We have also used keywords to address the problem of classifying sequences in functional groups, i.e. cellular functions, using statistics about the relationships between keywords and classes of cellular function (Tamames *et al*., 1996).

The system presented here could constitute a first-aid tool for retrieving information and suggesting functional annotations, could also be useful for the investigation of functional relationships between proteins, and could be integrated into genome analysis systems.

Methods

This method estimates the significance of words by comparing the abundance of words in a given set of abstracts related to a protein family with their abundance and distribution in a background set of abstracts associated to a wide range of different protein families.

Selecting a background distribution of abstracts associated to a diverse set of protein families

To obtain a representative set of words (and their abundance) in protein families, we selected a subset of distinct non-overlapping protein families. These were taken from PDBSELECT, which contains proteins with <25% sequence similarity between them (Hobohm and Sander, 1994) (ftp://ftp.sander.heidelberg-de/pub/databases/protein_extras/pdb_select/pdb_select.mar_1994).

Protein families were taken from the HSSP database (Schneider *et al*., 1997), with each family corresponding to one of the PDBSELECT proteins. To ensure that the proteins contained in each family perform only one function, we selected only those proteins with >40% of sequence similarity to the master sequence of the family. The set of abstracts corresponding to each of the families was assembled with the MEDLINE pointers in the corresponding SwissProt entry of each protein. Very small protein families were excluded, i.e. those with less than five proteins linked to MEDLINE. This

set of protein families was used as the domain-specific background distribution.

Selecting sets of abstracts corresponding to protein families or protein functions

To collect the abstracts, we used the SRS system (Etzold *et al*., 1996), which provides convenient access to MEDLINE through the WWW. For example, a search for the MEDLINE files containing a word beginning with 'plastocyanin' can be performed through http://www.embl-heidelberg.de/srs/srsc?[MEDLINE-AllText:plastocyanin*].

Even if the process can be accelerated when the database is locally available, this information retrieval is the most computationally demanding task for large protein families.

Statistical analysis of the words in the background distribution

The first step in the procedure is building the dictionary of all the words used in the abstracts linked to the background distribution of families. For this dictionary and for all the statistical analyzes that follow, the words were stemmed, considering variant forms of the same word as identical.

For this prototype system, we have implemented a simple stemming algorithm. First, we applied a set of simple rules for choosing words: (i) any hyphen followed by a carriage return is removed, assuming it splits a word; (ii) any other hyphen is replaced by a space; (iii) characters other than letters and numbers are removed; (iv) words composed exclusively of numbers are removed. The stem of the words is obtained by selecting their common beginnings. Two words are considered to have the same stem if they have the same beginnings and their endings differ in one or two characters. This includes most plurals (e.g. 'kinase-' and 'kinase-s') and verb tenses (e.g. 'transcript-s' and 'transcript-ed'). We do not apply this rule when the stem has less than five characters to avoid unwanted situations like mistaking 'acti-n' with 'active'. Other, more elaborated methods already developed could be included in the system at a later phase (e.g. Ulmschneider and Doszkocs, 1983) and are used in database search engines such as PubMed (http://www.ncbi.nlm.nih.gov/PubMed/) or the Library of Congress (http://lcweb2.loc.gov/catalog).

Once the dictionary has been compiled, two statistical parameters are computed for each of the words of this dictionary: their frequency in each family and the deviation of the distribution of their frequencies in the set of families.

The frequency of a word *a* in the family *i*, can be measured as:

$$F_i^a = \frac{W_i^a}{S_i} \qquad (1)$$

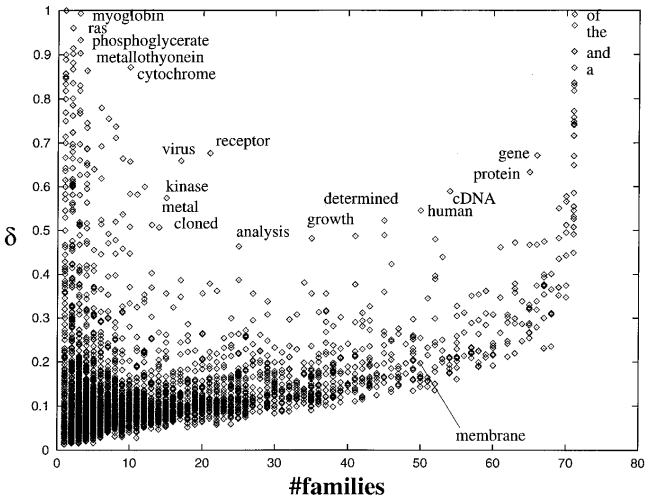


Fig. 1. Distribution of the words in the background set of protein families. The mean frequency of a word in the families in which it is present (δ) versus the number of families in which it is present. Some of the word names are depicted in the figure. The right side of the graph shows words that are present in most families (up to a total of 71 families). The left part contains words that are present in only a few families. Those at the upper left are found with high frequency, indicating that they are likely to be good indicators of the function of the families from which they have been extracted. The full list of keywords generated for the 71 families is available as additional material.

where W_i^a is the number of sequences of family *i* for which the word *a* was found in a linked MEDLINE and S_i is the number of sequences in family *i*.

With the following graph, we illustrate the complexity of the data that will be used as the background set. Figure 1 displays the frequency of a word in the background set of 71 protein families. The *x*-axis of the graph contains the number of families in which a word *a* is present and the *y*-axis contains the average frequency of the word (δ^a) in the families in which it is present:

$$\delta^a = \frac{\sum_{i=1}^{n} F_i^a}{\sum_{i=1}^{n} x_i} \qquad (2)$$

where $x_i^a = 0$ if $W_i^a = 0$ and $x_i^a = 1$ if $W_i^a > 0$, and n is the number of families. The upper right portion of the graph shows words that are present in most of the families and in most of the sequences of those families. There we find common English words like 'the', 'and' or 'of'. It is interesting to observe that other words like 'gene', 'protein' or 'cDNA', very common in this specific knowledge domain, are close to the common words in frequency and number of families.

The upper left part of the graph contains words present in a small number of families, but in high frequency (present in most of the proteins of the corresponding family). In this case, we find words like 'myoglobin', 'ras', 'phosphoglycerate', 'metallothionein' or 'cytochrome', which correspond to the common name of one or a few of the families included in the analysis. The detection of other possibly meaningful words like 'kinase' and 'receptor' requires a more precise analysis.

To discriminate high-information-content words relevant to the knowledge domain of protein function, it is essential to combine the frequency of a word in the families with the number of families in which it is present. For example, 'kinase' is an extreme case since it has a high frequency only in a few families, three in this case, and it occurs very seldom in other families. Common words like 'the' are at the other extreme; they are also very frequent, but are found in all protein families. Intermediate cases between these extremes are the interesting ones; for example, the word 'membrane' is found in almost all families, but in very different frequencies, ranging from 0 to 100%. It is in these cases where it is important to be able to annotate 'membrane' as a keyword only in those families where its frequency is significant.

The deviation of the distribution of word frequencies in the set of families is indicative of whether or not the word is strongly associated with particular families, and is thus an indication of the functional informational content of the word.

The deviation of word distribution is given as:

$$\sigma^a = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(F_i^a - \overline{F^a})^2} \qquad (3)$$

where $\overline{F^a}$ is the mean frequency of word a in the background distribution:

$$\overline{F^a} = \frac{\sum_{i=1}^{N} F_i^a}{n} \qquad (4)$$

and n is the total number of families.

The dictionary of the background set of protein families, and the frequency and deviation of its words to evaluate new protein families, can now be used.

Evaluation of a query family

Provided with a query family with an associated set of MEDLINE abstracts, we can now evaluate the words that are likely to be important for the family (putative keywords) by comparison with the background set.

First we compute the dictionary of the words used in the MEDLINE abstracts provided for the query family. The frequency of each of these words is then calculated as:

$$F^a = \frac{W^a}{S} \qquad (5)$$

where W^a is the number of abstracts in which the word a is found and S is the number of abstracts supplied for the query family.

To evaluate whether a word constitutes a keyword in the particular query family, we compute the z-score for each word. The z-score of a word a is defined as the difference between the frequency of a word in the query family minus the mean frequency of word in the background distribution, divided by the deviation of the word background distribution:

$$z^a = \frac{F^a - \overline{F^a}}{\sigma_a} \qquad (6)$$

This score gives an idea of the distance of the frequency of a word in a query family from the general distribution of this word in the background set of families.

Note that our procedure to obtain word frequencies is different from a simple word counting and, therefore, the kinds of distribution we are analyzing are also different from the typical Zipfian distribution (Zipf, 1935).

It is conceivable that some words present in the query protein family will not be found in the small background distribution used here. At this point, we consider these new words as significant ones, giving them a symbolic value of 'new' instead of a numerical z-score.

Selection of the most significant sentences

Sentence boundaries were identified by inspection of the punctuation marks (i.e. '.') of the abstracts associated to the query family. Cases in which punctuation marks are used for abbreviations or for numerical annotations were ignored. This simple process was found to perform well in most cases, given the simplicity of the language used in scientific abstracts. Other more elaborate methods have been attempted for non-scientific text (e.g. Reynar and Ratnaparkhi, 1997).

The extracted sentences are scored by averaging the z-score values of the individual words. During the analysis, it becomes evident that the sentences are very useful for interpretation of the keywords since they provide the appropriate context for their interpretation.

Web interface

The algorithm has been implemented in a Web server that accepts text from multiple abstracts, e.g. a concatenation of abstracts proceeding from a search in MEDLINE. The results are given as Web pages with relative links to facilitate information retrieval. The information provided includes keywords, sentences and evaluated abstracts with their corresponding scores. The keywords are linked to the sentences containing them, and the sentences themselves are

linked to the abstracts in which they were found. If the abstracts have a MEDLINE identifier, they will be linked to the NCBI MEDLINE. The abstracts are presented highlighted with the more significant sentences and keywords.

Results

Analysis of the background distribution of MEDLINE abstracts related to protein families

The performance of this system depends critically on the composition of the background distribution used as a reference. For this reason, we first assessed the quality and quantity of the information contained in the background set designed for the prototype system. The algorithm was applied to the families used for the background distribution. The keywords obtained for each of 71 protein families were compared with the words selected by human experts for annotating the corresponding entries in the database. For the purpose of this comparison, the words included in the SwissProt (Bairoch and Apweiler, 1997) KW and DE fields were treated in the same way as the words extracted from the MEDLINE abstracts: they were computed when they appeared associated to >50% of the proteins of a family and showed a *z*-score > 0.10. The number and quality of both sets of keywords were then compared by eye for each family. The results obtained with this system and with the extraction of SwissProt keywords are available as additional material.

Our assessment is that, for 16 families, the background distribution contains more information than the equivalent SwissProt keywords, for 12 families the performance was similar, and for 31 families the background contains substantial information, but not as much as the SwissProt entries. Only in the remaining 12 cases does the system perform quite poorly and assigns only one keyword. This is due mainly to the heterogeneity and small number of abstracts that are linked directly in SwissProt. Better results are expected for the analysis of more extensive and coherent sets of abstracts.

In any case, the background distribution contains a large and diverse set of words associated with different protein families, providing an adequate reference set for the analysis of the information contained in other protein families, at least at the level of the prototype system.

Automatic annotation of protein families

In the first example, we analyze abstracts obtained by querying MEDLINE with the word ‘ataxia telangiectasia’, which refers to a human disease associated with a disorder of a particular protein. The most significant words obtained (Figure 2) are directly related to the function of the protein and to its genetic origin, such us ‘recessive’, ‘disorder’, ‘atm’ (the gene name), ‘predisposition’ to the disease, ‘heterozygote’ mode of ‘inheritance’. In many cases, the *z*-scoring scheme used is able to highlight significant words that are not very frequent, but differ significantly from the background, e.g. ‘heterozygote’ has the highest *z*-score (57.35), even though its frequency is only 22%. Inversely, some very abundant words are considered less significant, since they differ little from the composition found in the background distribution, e.g. ‘genetic’ has a low *z*-score (3.53) even though its frequency is 39%. It is also interesting that some words, such us ‘cancer’, given their connotation, seem to be a very attractive choice for most authors and they end up scoring high in our system. With a *z*-score of 8.93 and a frequency of 44%, ‘cancer’ has the highest frequency among the words with high *z*-score.

The most significant sentences selected by the system are also shown in Figure 2. Sentences are easier to follow than single words, since they include fundamental contextual information. A clear, descriptive sentence could be the sixth one in Figure 2, with an average *z*-score of 6.77: ‘ataxia telangiectasia is a genetic disorder with an autosomic recessive transmission’ (keywords are underlined). Many of the other high-scoring sentences are also informative, and a human expert would have no difficulty in choosing the most appropriate one by simple browsing.

For the purpose of comparison, it can be said that the corresponding database entry (HS24551 in EMBL) is described as ‘human phosphatidylinositol 3-kinase homolog (ATM) mRNA’ and ‘Ataxia-Telangiectasia mutated’, whose meaning is not immediately obvious to the reader. The second information provided by the database entry is a pointer to a single MEDLINE reference that corresponds to a paper entitled ‘A single gene with a product similar to PI-3 kinase’. It seems clear that in this example our system brings more easily interpretable functional information.

A second example was obtained by querying the system with abstracts containing the words ‘ran’ and ‘GTPase’, related in principle to the function of the small GTPase ‘ran’. This protein is implicated in nuclear transport and belongs to the large superfamily of ras-related proteins. This example is presented to illustrate some of the performances of the current system, in particular, how useful the sentence analysis is. This case also shows how the system is tolerant to some errors in the selection of the initial set of texts.

In this case, the words with higher *z*-scores were ‘binding’, ‘hydrolysis’, ‘GTPase’, ‘GTP’, ‘GDP’, ‘exchange’, ‘ras’, ‘tc4’, ‘binding’, ‘nuclear’, ‘import’. Many of these words are clearly descriptive of the function of the protein family and could be used directly as keywords. For example, ‘ras’ refers to the protein superfamily to which ran belongs, and ‘GTP’ and ‘GDP’ are the cofactors bound by this protein family, while ‘binding’, ‘hydrolysis’, ‘GTPase’ and ‘exchange’ refer to the differentiated enzymatic activity of the protein that uses GTP and afterwards replaces the used GDP for a new GTP. Finally, ‘nuclear ’ and ‘import’ are related to the

(a)

Keywords extracted by AbXtract for query ataxia

Text Sentences

keyword	freq.	z	found at
disorder	0.37	12.30	[95312868] [96255945] [96018070] [89070677] [96391788] [97288735] [96105020] [97262209] [96154672] [94141360] [91171851] [90174198] [92097021] [96254972] [90235178] [92058549] [95150036] [93233707] [93206979] [95357487] [83107389] [91169545] [95048379] [95381456] [95364870] [82029766] [89151839] [89139395] [96390593] [87164160] [89250669] [97041722] [83259005] [96404417] [94221101] [97294602] [95138623] [90316537] [97123513]
autosomal	0.33	9.02	[95312868] [96255945] [89070677] [97288735] [97262209] [96154672] [95372371] [90321562] [91171851] [96038263] [92097021] [96254972] [90235178] [92058549] [93233707] [86061765] [93206979] [95357487] [91169545] [96338579] [95048379] [88314587] [95213012] [82029766] [89003946] [92035738] [89139395] [83259005] [96404417] [94221101] [92072632] [90316537] [90259001] [97123513]
disease	0.30	5.29	[95312868] [96255945] [93104606] [97288735] [95372371] [96038263] [92194830] [96254972] [91288572] [92298322] [92058549] [86061765] [93144253] [95357487] [83107389] [91169545] [96338579] [95048379] [88314587] [96105008] [95213012] [95187960] [82029766] [92035738] [87239245] [96390593] [96081682] [86061784] [96377133] [94221101] [92072632]

generated by **AbXtract** (Fri Nov 21 19:36:47 GMT 1997)

(b)

Sentences extracted by AbXtract for query ataxia

Text Keywords

medline	sentence	score
90235178	**cancer predisposition** of **ataxia telangiectasia heterozygotes**	14.70
85254461	workshop on **ataxia telangiectasia heterozygotes** and **cancer**.	12.31
96154672	a **t heterozygotes** are moderately **cancer** prone	9.47
92298322	enhanced levels of **radiation** induced g2 phase delay in **ataxia telangiectasia heterozygotes**	8.36
86061784	cellular hypersensitivity to chronic gamma **radiation** in cultured fibroblasts from **ataxia telangiectasia heterozygotes**	7.72
96255945	**ataxia telangiectasia** is a **genetic disorder** with an **autosomic recessive** transmission	6.77
96154672	**ataxia telangiectasia** is an **autosomal recessive disorder** involving **cerebellar** degeneration, immunodeficiency **radiation** sensitivity, and **cancer predisposition**	6.69
92072632	**patients** with **ataxia telangiectasia** and cells derived from homozygotes and **heterozygotes** are unusually sensitive to **ionizing radiation**	6.62

(c)

Formatted text by AbXtract for query ataxia

Sentences Keywords

95312868

> a single **ataxia telangiectasia** gene with **a product** similar to pi-3 kinase a gene, **atm**, that is mutated in the **autosomal recessive disorder ataxia telangiectasia** was identified by positional cloning on chromosome 11q22-23

> *at is **characterized** by **cerebellar** degeneration, **immunodeficiency**, **chromosomal** instability, **cancer predisposition**, **radiation** sensitivity, and cell **cycle** abnormalities*

> *the **disease** is genetically heterogeneous, **with four complementation** groups that have been suspected to represent different genes*

> **atm**, which has a transcript of 12 kilobases, **was found** to be mutated in at **patients** from all **complementation** groups, indicating that **it is probably** the sole gene **responsible** for this **disorder**

> a partial **atm** complementary dna clone **of 5.9 kilobases** encoded a putative protein that is similar to several yeast and mammalian phosphatidylinositol-3' kinases that are involved in mitogenic signal transduction, meiotic recombination, and cell cycle control

> *the discovery of **atm** should enhance understanding of at and related **syndromes** and may allow the identification of at **heterozygotes**, who are at **increased** risk of **cancer**.*

Fig. 2. Results of the analysis of 100 abstracts containing 'ataxia telangiectasia'. The results are shown in the original Web interface format. Words, abstracts and sentences are cross-linked, and it is possible to travel between them. (**a**) Keywords and the MEDLINE identifiers of the abstracts in which they were found, scored and selected by our algorithm. Five words, including 'atm', which is the gene name, or those used for the query, 'ataxia' and 'telangiectasia', were not present in the background distribution, but were selected due to their high frequency (>20%). A total of 21 words with *z*-scores > 0.20 were reported. Only some are shown in the screen dump of this figure. Best *z*-scoring words (with frequency indicated in percentage): 'heterozygote' (57.35, 22%), 'recessive' (29.73, 37%), 'radiation' (23.08, 41%), 'disorder' (12.30, 37%), 'patient' (9.79, 39%), 'autosomal' (9.02, 33%), 'cancer' (8.93, 44%), 'familial' (8.28, 23%). (**b**) Best sentences with the highlighted significant words (in bold face). Note that 'at' is an abbreviation of 'ataxia telangiectasia', as in the third sentence, and not the preposition 'at'. (**c**) One of the abstracts used. Best sentences and words are highlighted (in italics and bold face, respectively).

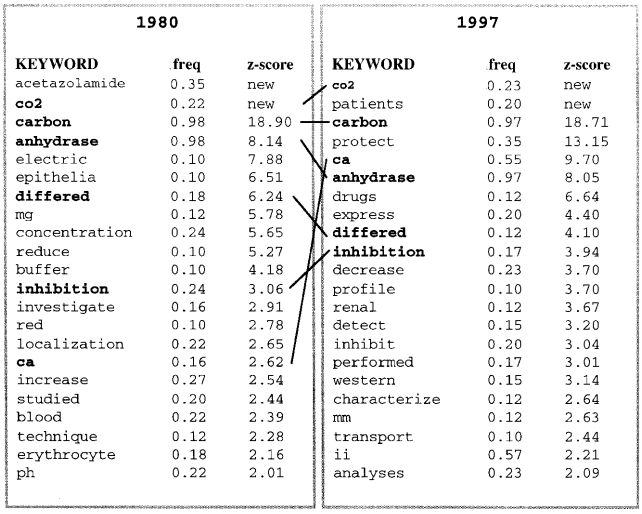
cellular function of this protein, and 'rtc4' is the name of the gene coding for the ran protein.

The meaning of single words becomes clearer when they are embedded in the context of the corresponding sentences, such as 'like all ras related GTP binding proteins, gsp1p undergoes cycles of GTP hydrolysis and GDP GTP exchange', which is the best scoring sentence (10.09). An interesting phrase is 'ran, a GTPase involved in nuclear processes' (9.40, fourth-highest scoring).

The importance of the usage of sentence scoring schemes is highlighted by the results obtained with sets of abstracts selected by keywords and shows the superiority of our algorithm to this kind of simplistic search. For example, in the set of abstracts for the 'ran' family, one completely unrelated abstract was included in the analysis. This abstract contained the word 'ran', in this case with the meaning 'ran counter'. None of the sentences associated to this abstract were selected as relevant due to the absence of other significant words. The system appears able to tolerate a certain amount of noise in the selection of the input set of abstracts.

A last example is chosen to illustrate the dependency of the information contained in the literature analyzed. The results obtained by triggering the system with the words 'carbonic' and 'anhydrase', restricted to abstracts of articles published during two years (1980 and 1997), are compared in Figure 3. There are some significant words common to the two sets, like 'carbon' or 'CO_2', the cofactor of the protein family. Interestingly, there are some significant words that are very different between the two sets. For example, 'acetazolamide' was highly significant in the first year, but not later, and 'ii' for a second isoform of the enzyme is only significant in recent years. This is because the initial studies on carbonic anhydrase put much emphasis on discovering inhibitors that allow the manipulation of the enzymatic activity by 'acetazolamide', which later became standard knowledge and was no longer mentioned in the later abstracts. By the time of the second period, a second isoform had been discovered (isoform 'ii'), triggering significant interest and a corresponding increase in published papers.

The system depends on the input information. As in most protein families, functional research is still in active progress



1980

KEYWORD	freq	z-score
acetazolamide	0.35	new
co2	0.22	new
carbon	0.98	18.90
anhydrase	0.98	8.14
electric	0.10	7.88
epithelia	0.10	6.51
differed	0.18	6.24
mg	0.12	5.78
concentration	0.24	5.65
reduce	0.10	5.27
buffer	0.10	4.18
inhibition	0.24	3.06
investigate	0.16	2.91
red	0.10	2.78
localization	0.22	2.65
ca	0.16	2.62
increase	0.27	2.54
studied	0.20	2.44
blood	0.22	2.39
technique	0.12	2.28
erythrocyte	0.18	2.16
ph	0.22	2.01

1997

KEYWORD	freq	z-score
co2	0.23	new
patients	0.20	new
carbon	0.97	18.71
protect	0.35	13.15
ca	0.55	9.70
anhydrase	0.97	8.05
drugs	0.12	6.64
express	0.20	4.40
differed	0.12	4.10
inhibition	0.17	3.94
decrease	0.23	3.70
profile	0.10	3.70
renal	0.12	3.67
detect	0.15	3.20
inhibit	0.20	3.04
performed	0.17	3.01
western	0.15	3.14
characterize	0.12	2.64
mm	0.12	2.63
transport	0.10	2.44
ii	0.57	2.21
analyses	0.23	2.09

Fig. 3. Best keywords extracted for two queries with abstracts containing 'carbonic' and 'anhydrase' in the years 1980 and 1997. The significant words common to both queries are bold faced and linked. The tag 'new' in the *z*-score column indicates that a given word was not found in the background distribution, indicating its exclusive relationship with the protein family under study.

and the extracted descriptions are necessarily a product of their time.

Discussion

The analysis of a set of abstracts related to protein families is carried out comparing word frequency with background distributions in broad sets of protein families. The words selected are found in most cases to be good indicators of different aspects of protein function and can be used as a guide for database annotations, sentences help to understand the function of the protein family under study in a summarized way, and the highlighted abstracts save time during bibliographic searches. Three different examples have been presented to illustrate how keywords, sentences and abstracts are selected by the system.

Relationship to other systems

The proposed system is based on simple word statistics with techniques similar to those used in statistical approaches to language understanding (Jacobs, 1992; Allen, 1994; Wilbur and Coffee, 1994). The difficulties are, therefore, those that arise from the interpretation of free-style text (human-written text) by computer programs.

A vectorial system for comparing text has been used in other domains, and a variant based on neighbor relation has been especially useful in the biological domain, where MEDLINE abstracts are already scanned by word in the ENTREZ system [http://www3.ncbi.nlm.nih.gov/Entrez/index.html (Schuler *et al*., 1996)]. The vector information is used by the NEIGHBOUR system (Wilbur and Coffee, 1994) to compare individual abstracts with the background distribution of words in MEDLINE, and individual abstracts are compared to discover similarities.

Our approach attempts to extract relevant information contained in the abstracts. We are therefore interested in analyzing the set of abstracts related to protein families rather than in comparing individual abstracts.

Our system is based on the consideration of protein families (proteins related by an evolutionary link), and it evaluates the information contained in MEDLINE abstracts grouped by protein families rather than segregated in individual abstracts. The discrimination of keywords from non-informative words thus occurs within and between protein families. In our system, protein families are used to build a background distribution of words specific to the knowledge domain of protein functionality. The significance of word frequency in sets of abstracts is estimated by comparing them with a background distribution obtained from a selected set of protein families.

An example can illustrate the difference between considering individual abstracts and protein family-related information. A word full of biological meaning, such as 'membrane', can be found in many abstracts of articles describing proteins. A particular abstract may have the word many times, but we cannot ascertain whether it refers to the protein function, or to a technique (e.g. 'membrane' can be used either in the context of a transmembrane protein or it can refer to a dialysis membrane used to extract the protein). If we consider protein families, the situation is clearly different. The word 'membrane' will appear in most of the abstracts associated with some specific protein families, but very seldom in abstracts concerning others. It should be possible to conclude that 'membrane' is a keyword for some protein families and not for others.

It is conceivable that more complex approaches would lead to further improvements. Part of our intention with this first communication is to trigger the interest of researchers in the area of language understanding applied to the annotation of biological function.

Future prospects

At least five important features are required to extend the current prototype to a fully operational system.

In the first place, correlation between words should be considered. In this case, words that do not score high themselves can be discovered by their association with other words. The problem with negative sentences can also be addressed through the study of short-range correlations between words.

Second, a larger background distribution should replace the one derived from the 71 protein families used in this study. They can be derived from other definitions of protein families, e.g. PROSITE (Bairoch *et al*., 1997), pfam (Sonnhammer *et al*., 1997), or recently derived sets of protein clusters (Holm and Sander, 1998). Any of these sets include more functional diversity than the HSSP database used here.

Third, a larger corpus of textual information should be analyzed for each query. This can be achieved by scanning MEDLINE abstracts for long periods of time or including full papers rather than abstracts alone.

Fourth, the selection of sentences can be optimized to avoid spurious short or partial sentences by implementing a sentence size-dependent weighting scheme.

Fifth, the sentences selected in some cases are similar. It would be interesting to select sets of sentences with complementary meaning. The analysis of word overlap between sentences could be implemented to achieve this goal.

Unfortunately, some of these enhancements will significantly increase computation costs. On a typical UNIX workstation, it took 6 min to extract and analyze a protein family of 62 abstracts against the background of 71 families. We estimate that roughly 100 times more effort will be required for a system working with a larger reference set, hundreds of abstracts for the family to be analyzed and scoring single words and pairs of words. At that point, more sophisticated computational techniques will have to be used.

Acknowledgements

We thank Chris Sander for various encouraging discussions on this work. This work was supported under EC-TMR grant 'GeneQuiz'. M.A.A. holds a postdoctoral fellowship from the same program.

References

Allen,J. (1994) *Natural Language Understanding*. Benjamin/Cummings, Redwood City, CA.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bairoch,A. and Apweiler,R. (1997) The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Res.*, **25**, 31–36.

Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.*, **25**, 217–221.

Berry,M.W., Dumais,S.T. and O'Brien,G.W. (1995) Using linear algebra for intelligent information retrieval. *SIAM Rev.*, **37**, 573–595.

Casari,G. *et al*. (1995) Challenging times for bioinformatics. *Nature*, **376**, 647–648.

Cowie,J. and Lehnert,W. (1996) Information extraction. *Commun. ACM*, **39**, 80–91.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.

Gaasterland,T. and Sensen,C.W. (1996) Fully automated genome analysis that reflects user needs and preferences—a detailed introduction to the MAGPIE system architecture. *Biochimie*, **78**, 302–310.

Guigó,R., Johansson,A. and Smith,T.F. (1991) Automatic evaluation of protein sequence functional patterns. *Comput. Applic. Biosci.*, **7**, 309–315.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.

Jacobs,P.S. (1992) *Text-based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*. Lawrence Erlbaum Associates, Hillsdale, USA.

Koonin,E.V., Tatusov,R.L. and Rudd,K.E. (1996) Protein sequence comparison at genome scale. *Methods Enzymol.*, **266**, 295–322.

Pearson,W. and Lipman,D. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Reynar,J.C. and Ratnaparkhi,A. (1997) A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the 5th Conference on Applications of Natural Language Processing*, Washington, DC, pp. 16–19.

Salton,G. (1989) *Automatic Text Processing. Addison-Wesley Series in Computer Science*. Addison-Wesley, Reading, MA.

Schneider,R., de Daruvar,A. and Sander,C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.

Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.

Sonnhammer,E., Eddy,S. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.

Tamames,J., Ouzounis,C., Sander,C. and Valencia,A. (1996) Genomes with distinct functional composition. *FEBS Lett.*, **389**, 96–101.

Ulmschneider,J.E. and Doszkocs,T. (1983) A practical stemming algorithm for online search assistance. *Online Rev.*, **7**, 301–315.

Wilbur,W.J. and Coffee,L. (1994) The effectiveness of document neighboring in search enhancement. *Inf. Process. Manage.*, **30**, 253–266.

Zipf,G.K. (1935) *The Psycho-Biology of Language*. MIT Press (1968), Cambridge, MA.