BIOINFORMATICS

Vol. 20 no. 13 2004, pages 2084–2091
doi:10.1093/bioinformatics/bth207

Gene annotation from scientific literature using mappings between keyword systems

Antonio J. Pérez[1], Carolina Perez-Iratxeta[2,3,†], Peer Bork[2,3], Guillermo Thode[1] and Miguel A. Andrade[2,3,*,†]

[1]University of Málaga, Facultad de Ciencias, Departmento de Genetica, Group of Bioinformatics, Campus Universitario de Teatinos, 29071 Málaga, Spain, [2]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany and [3]Max Delbrück Center for Molecular Medicine, Department of Bioinformatics, PO Box 740238, 13092 Berlin-Buch, Germany

Received on October 17, 2003; revised on January 7, 2004; accepted on January 19, 2004
Advance Access publication April 1, 2004

ABSTRACT

Motivation: The description of genes in databases by keywords helps the non-specialist to quickly grasp the properties of a gene and increases the efficiency of computational tools that are applied to gene data (e.g. searching a gene database for sequences related to a particular biological process). However, the association of keywords to genes or protein sequences is a difficult process that ultimately implies examination of the literature related to a gene.

Results: To support this task, we present a procedure to derive keywords from the set of scientific abstracts related to a gene. Our system is based on the automated extraction of mappings between related terms from different databases using a model of fuzzy associations that can be applied with all generality to any pair of linked databases. We tested the system by annotating genes of the SWISS-PROT database with keywords derived from the abstracts linked to their entries (stored in the MEDLINE database of scientific references). The performance of the annotation procedure was much better for SWISS-PROT keywords (recall of 47%, precision of 68%) than for Gene Ontology terms (recall of 8%, precision of 67%).

Availability: The algorithm can be publicly accessed and used for the annotation of sequences through a web server at www.bork.embl.de/kat

Contact: mandrade@ohri.ca

INTRODUCTION

Since their inception, the protein sequences stored in the SWISS-PROT us.expasy.org/sprot/ and TrEMBL databases have been annotated with keywords chosen from a list of controlled terms (Boeckmann *et al*., 2003). Typically, keywords are manually chosen by a database curator from a controlled vocabulary, possibly after examination of the scientific literature related to the gene. Given the fact that the number of genes and data about these are growing faster than the amount of annotators, it is obvious that automatic methods are needed to support the task. Kretschmann *et al*. (2001) approached this problem by deducing rules for the association of protein domains and taxonomy to SWISS-PROT keywords. An alternative keyword set, Gene Ontology (GO), a system of keywords hierarchically organized as a directed graph with three main categories (‘biological process’, ‘cellular component’ and ‘molecular function’) (Ashburner *et al*., 2000), provided a unified set of terms for the annotation of proteins in different organisms. Although the scheme was initially set up for eukaryotic organisms and each annotation had to have a link to a scientific reference, the system was quickly expanded for the annotation of genes from all organisms. Among the approaches for automated assignment of GO terms to sequences that are recently flourishing, one of the most pragmatic is used by the Gene Ontology Annotation (GOA, www.ebi.ac.uk/GOA/) project (Camon *et al*., 2003). The authors of GOA developed manually mappings between protein domains and GO terms, and between SWISS-PROT keywords and GO terms, so that a sequence can automatically receive certain GO terms if it contains a domain or if it is already annotated in SWISS-PROT with a certain keyword.

Another possibility for the extraction of keywords related to a gene is the analysis of the scientific literature, ultimately the richest and most accurate source of functional information related to genes. In this respect, literature data have been used for detection of words related to function (Andrade and Valencia, 1998; Shatkay *et al*., 2000), GO terms (Pouliot *et al*., 2001; Xie *et al*., 2002) and for the extraction anew of a whole ontology of gene function (Valencia and Blaschke, 2002). However, as far as we know, there is no method yet to assist the annotation of a gene with keywords using as source

*To whom correspondence should be addressed.

†Present address: Ottawa Health Research Centre, 501 Smyth Road, Ottawa, Ontario K1H 8L6, Canada.

of data just a small set of manually selected related references (from 1 to 10), a typical situation that annotators face everyday.

To this end, we approached the problem of establishing automatically mappings between terms from the MEDLINE database of scientific literature and keyword systems such as SWISS-PROT keywords and GO. For this, we expanded a method previously used for the association of GO terms to human inherited diseases (Perez-Iratxeta *et al*., 2002a) and for the annotation of sequences in SWISS-PROT with keywords according to the MeSH terms present in the scientific references linked to the entry (Perez-Iratxeta *et al*., 2003). The MeSH terms are an ontology developed at the National Library of Medicine for the annotation of the entries in the MEDLINE database (NLM, www.nlm.nih.gov/mesh/). In this work, we used both the MeSH terms and words from the abstracts in MEDLINE as source of literature data. We developed an annotation system that was applied and tested for the annotation of sequences with both SWISS-PROT keywords and GO terms.

SYSTEMS AND METHODS

Mapping keyword systems from cross-links between databases

It is usual to find links between entries in different molecular biology databases. Here, we propose a method to define mappings between databases from these links. These mappings can be used to check a database internal coherence (Perez-Iratxeta *et al*., 2003) to synchronize databases upon the update of one of them, or, as in this work, to generate new database annotations.

Here, we focus on deriving mappings between the database of SWISS-PROT sequences (Boeckmann *et al*., 2003) and the MEDLINE database of scientific references. The links between these databases consist of the references to articles indexed in MEDLINE that are associated with protein entries in SWISS-PROT. The central idea underlying our method is to establish a mapping between MEDLINE and SWISS-PROT from such cross-links, so that a sequence can be annotated from a few related scientific references according to the precomputed mapping.

Given a protein in SWISS-PROT already annotated with a small subset of keywords and one or more links to MEDLINE references, it may be assumed that most of the keywords will summarize some of the scientific references linked to the entry. The analysis of the links in the whole database can be used to map the contents of the articles (e.g. the MeSH terms as annotated at the National Library of Medicine) to the keywords. For example, a strong association between the keyword 'Fatty acid biosynthesis' (a metabolic pathway) and the MeSH term 'Stearoyl-CoA Desaturase' (a protein that participates in that pathway) can be detected by counting in how many SWISS-PROT entries annotated with the keyword 'Fatty acid biosynthesis', one of the linked references, is annotated in MEDLINE with the MeSH term 'Stearoyl-CoA Desaturase'. If this occurs almost always, then we can learn automatically an association between the SWISS-PROT keyword and the MeSH term. Once such association is established, if a protein to be annotated is linked to a MEDLINE reference that contains the MeSH term 'Stearoyl-CoA Desaturase', the keyword 'Fatty acid biosynthesis' will be suggested by the system.

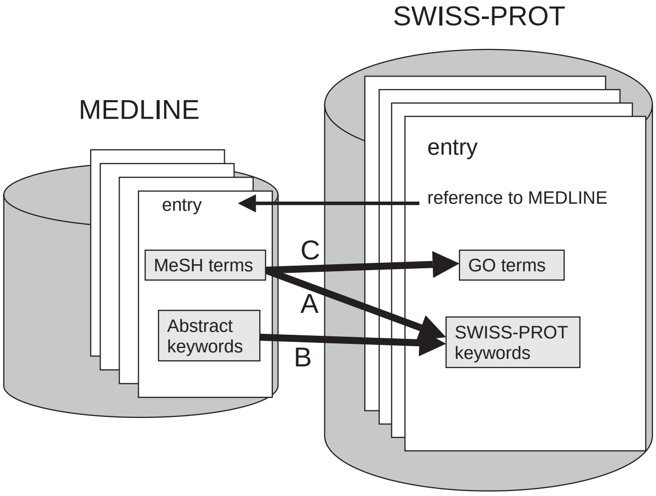


Fig. 1. We computed three different mappings. Each SWISS-PROT entry was considered as a transaction that puts in relation the keywords with which it is annotated (SWISS-PROT keywords and GO terms) with the terms associated with the abstracts linked in that entry (MeSH terms or relevant words extracted from the abstract). (**A**): Mapping between MeSH terms and SWISS-PROT keywords. (**B**): Mapping between words extracted from the abstract and SWISS-PROT keywords. (**C**): Mapping between MeSH terms and GO terms.

Following this outline, we derived three different mappings between MEDLINE entries and protein keyword systems (Fig. 1). The first one associated MeSH terms with SWISS-PROT keywords. The second one associated keywords that we extracted from the text of the abstracts to SWISS-PROT keywords. A third mapping was computed between MeSH terms and GO terms.

To derive a mapping we adapted a fuzzy thesaurus model (Miyamoto, 1990). A fuzzy thesaurus comprises collections of related words and of the type and 'strength' of the associations between them. We can consider two types of associations, namely 'related to' and 'including'. In a given context, two terms that co-occur frequently can be considered as highly related. However, always given a context, words can also have an asymmetrical relation: some words refer to broader concepts and are including others that are more particular. For example, the keyword 'Fatty acid biosynthesis' would be a broad concept including narrower terms such as the MeSH term 'Stearoyl-CoA Desaturase', because that protein is one of the several involved in the fatty acid biosynthesis. The relation of inclusion can be easily detected because if the narrow concept occurs, then the broader term tends to occur too, and the contrary does not necessarily happen; for example, we can find SWISS-PROT entries annotated with the keyword ‘Fatty acid biosynthesis’ but without any link to articles speaking of the ‘Stearoyl-CoA Desaturase’: they just might be dealing with another protein involved in fatty acid biosynthesis.

In summary, each mapping is nothing more than all possible ordered pairs formed with the elements of the two sets put in relation (e.g. MeSH terms and SWISS-PROT keywords) with a value attached to each pair that reflects the strength of the inclusion of the first element in the second element. In our work, we selected pairs with high values of inclusion of the element used to derive the annotation (e.g. the MeSH term) in the element to be used for annotation (e.g. the SWISS-PROT keyword) to ensure that the term used for annotation is implied by the term used to derive the annotation. For example, we wanted to be sure that ‘Stearoyl-CoA Desaturase’ was included in ‘Fatty acid biosynthesis’, so that if a reference linked to a sequence mentioned ‘Stearoyl-CoA Desaturase’ then we could safely annotate the sequence with the keyword ‘Fatty acid biosynthesis’. Pairs of equivalent terms, e.g. the keyword ‘Disease mutation’ and the MeSH term ‘Inherited disease’ would also be present in the mapping because equivalent terms are completely included one in the other. See the APPENDIX for details.

Database terms that were mapped

As was mentioned earlier, we computed mappings between the SWISS-PROT database and the MEDLINE database. The terms that were mapped in the SWISS-PROT database were the SWISS-PROT keywords and GO terms associated with the entries. The terms mapped from the MEDLINE database were the MeSH terms associated with the references and relevant words that we selected from the abstracts with an automated procedure (see below for details; Fig. 1). Here follows the description of these four data sets.

The MeSH terms conform to an ontology developed at the National Library of Medicine (www.nlm.nih.gov/mesh/). References to scientific literature are extensively annotated with several MeSH terms, typically around a dozen, which describe the contents of the article. MeSH terms are organized hierarchically, the top of that hierarchy being arranged in eight main categories. In this work we used only three such categories, namely ‘Diseases’ (MeSH C), ‘Chemicals & Drugs’ (MeSH D) and ‘Biological Sciences’ (MeSH G), because these were the categories that fit better to keywords describing gene and protein function. The inclusion of other MeSH main categories, as MeSH A terms (‘Anatomy’), did not yield better results (data not shown). We removed several non-informative terms from the set (the list can be accessed at the KAT web server, www.bork.embl.de/kat/).

The keywords used in SWISS-PROT (Boeckmann *et al*., 2003) belong to a controlled vocabulary but they are not hierarchically organized. This means that some incoherences in the annotation are possible (e.g. Perez-Iratxeta *et al*., 2003). All SWISS-PROT keywords were mapped with the exception of ‘Complete proteome’, ‘Hypothetical protein’, ‘Multigene family’ and ‘3D-structure’, which do not relate to a positive description of the protein function. The set of GO terms was taken from the current release of GO without any filtering (www.geneontology.org; December 2002). The annotation of SWISS-PROT sequences with GO terms was obtained from the GOA project [current release, December 2002; Camon *et al*. (2003)].

Finally, to select the relevant words (keywords) from abstracts in MEDLINE we adapted a procedure based on the fuzzy model explained earlier (Perez-Iratxeta *et al*., 2002b). Here, we considered only the nouns of the abstract. The sentences in the abstract were the transactions relating those nouns, so that we registered the co-occurrences of words in the same sentence. Then, given one abstract, we computed the strength of the association for every pair of words in that abstract and, finally, we selected as keywords the ones that were making more and stronger relations with others. See the Appendix for details.

We used only nouns as a source of keywords because they are better in this respect than other parts of speech such as adjectives or verbs. The reason is that the objects of molecular biology are normally defined by unique nouns, whereas synonymous adjectives and verbs can be found in scientific text.

Databases and links

The other components for our computation were the links from the SWISS-PROT database to the MEDLINE database. We used the version 40 of the SWISS-PROT database (containing 121 532 sequences). We selected the entries that had both associated keywords and at least one reference with associated MeSH terms. In order to be sure of recording links to the literature that were specific for the corresponding protein, we considered only those MEDLINE references linked to less than 12 sequences. In this way, we removed references to a total of 647 scientific articles (a small fraction of the total of 81 626 articles) most of them dealing with complete genome sequencing, like the one describing the sequencing of the K-12 strain of *Escherichia coli* (associated to a total of 3405 proteins of this organism). The resulting set of sequences consisted of 65 263 sequences with 171 687 links to MEDLINE abstracts (an average of 2.7 links per sequence). The number of abstracts linked was 116 482 (which is sensibly lower than the number of links because some of the abstracts in MEDLINE can be linked from multiple entries in SWISS-PROT).

The set of abstracts was annotated with an average of 0.19 MeSH C terms, 3.7 MeSH D terms (excluding

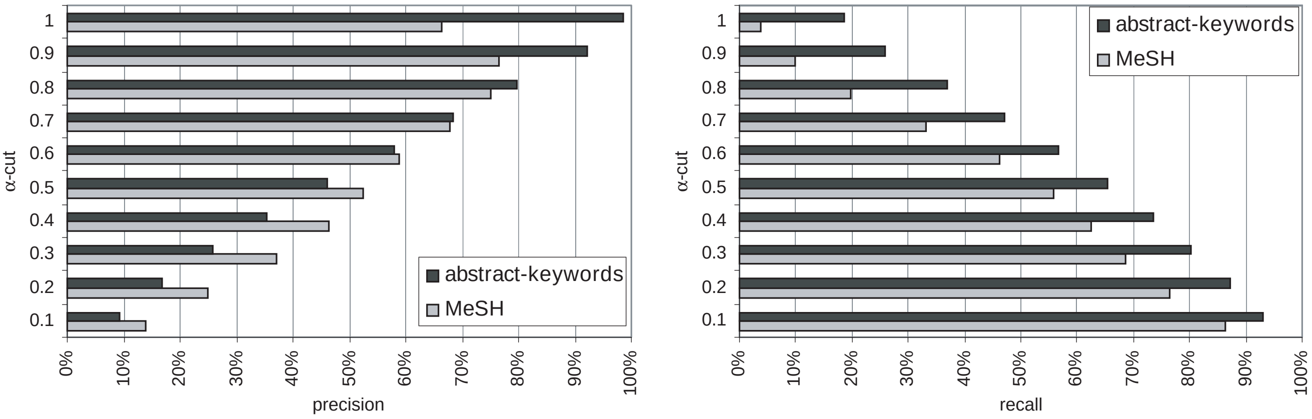


Fig. 2. Recall and precision in the prediction of SWISS-PROT keywords for varying cuts of the inclusion relation of the mapping between both MeSH terms (excluding non-informative MeSH D terms, see text for details) and abstract-keywords to SWISS-PROT keywords.

non-informative terms) and 5.6 MeSH G terms. The set of sequences was annotated with an average of 6.2 GO terms and 3.9 SWISS-PROT keywords.

BENCHMARK AND RESULTS

We performed a benchmark of the annotation protocol on the SWISS-PROT database (version 40) to test the validity of the approach. From the 65 263 entries considered (see Systems and methods section for details) we separated randomly 6526 entries (one-tenth). The three mappings were computed using the remaining set of entries and links. We annotated the 6526 excluded sequences using each of the mappings and compared the automated annotations with the actual annotations of the entries. The evaluation of the performance was carried out in terms of recall and precision. Recall is the fraction of the actual annotations that were automatically predicted. Precision is the fraction of automated predictions that were correct.

Derivation of SWISS-PROT keywords from MeSH terms

We computed a mapping that associated MeSH terms (C, D and G) with SWISS-PROT keywords. Each relation received a score (inclusion value, α) according to the strength of inclusion of the SWISS-PROT keyword into the MeSH term. The set of relations in the 0.1-cut (with $\alpha \geq 0.1$) and with a support of five or more (i.e. the pair was observed for a minimum of five SWISS-PROT entries) comprises 17 472, 44 093 and 10 877 pairs of MeSH terms (C, D and G, respectively) to SWISS-PROT keywords. See the Appendix for details about the computation of α.

The performance of the benchmark for different α-cuts of the relation is graphically displayed in Figure 2. For an α-cut 0.7 the precision was 68% and the recall 33%. The removal of non-informative MeSH D terms improved the results greatly (data not shown).

One of the examples of perfect automated annotation was obtained for the SWISS-PROT sequence ACOD_RAT that was predicted to be associated to the keywords ‘Endoplasmic reticulum’, ‘Fatty acid biosynthesis’, ‘Iron’, ‘Oxidoreductase’ and ‘Transmembrane’, when using an α-cut of 0.6. Those five keywords are the ones annotated in SWISS-PROT and no other keyword was predicted, meaning a recall and precision of one for this particular case. All of them originated from a MEDLINE entry linked to the sequence (PubMed Identifier, PMID: 2428815) that contained the MeSH D term ‘Stearoyl-CoA Desaturase’. The mapping related this term to the five SWISS-PROT keywords with an inclusion value of 0.89 and a support of 8.

We note that a wrong prediction in the benchmark did not necessarily mean that the system was making a mistake. For example, another SWISS-PROT sequence, VSP1_TRIST, was assigned the keyword ‘Plasminogen activation’ because one article linked (PMID: 7730329) that characterizes the function of the sequence is annotated with the equivalent MeSH term ‘Plasminogen Activators’. However trivial, this annotation was not present in the version of SWISS-PROT used for the analysis and the automated assignment of the keyword by our system was counted as a false prediction.

Of course, the system is error prone and can assign wrong keywords. For example, the sequence MYS2_DICDI was assigned the keyword ‘Muscle protein’ because in the mapping the MeSH term ‘Myosin Subfragments’ is included in that keyword with an inclusion value of 0.80, and the term was found in an article linked to MYS2_DICDI (PMID: 8611530). However, the sequence is clearly annotated by SWISS-PROT with the description ‘Myosin II heavy chain, non muscle’ meaning that it is not a muscle protein, and the assignment of the keyword was wrong.

Derivation of SWISS-PROT keywords from abstract-keywords

The mapping that associates abstract-keywords to SWISS-PROT keywords was considerably larger: the 0.1-cut of the inclusion relation contained 466 661 pairs. The results of the benchmark are displayed in Figure 2. The performance was better than that corresponding to the MeSH/SWISS-PROT mapping. For the 0.7-cut the precision was of 68% with a recall of 47%.

For example, the SWISS-PROT entry AIP_CERAE was annotated only with the two SWISS-PROT keywords, ‘Repeat’ and ‘TPR repeat’, both of them related with an inclusion value higher than 0.8 to the word ‘tetratricopeptide’, which was extracted as a keyword from one of the abstracts linked to the sequence (PMID: 9447995).

An example of a correct prediction that was not present in SWISS-PROT was given by the annotation of the sequence HS7S_CUCMA with the keyword ‘Chaperone’ because of the related abstract-keyword ‘chaperonin’ (from the linked article PMID: 8096466, that describes the detection of this heat-shock protein).

An example of wrong prediction was the annotation of FER_HUMAN, that was assigned the keyword ‘Transmembrane’ because in one of the related papers (PMID: 2725517) one sentence contains the noun ‘transmembrane’, but it is used in a negative sense by describing that the translation product of the cDNA encoding the corresponding gene ‘lacks a clear transmembrane region’. This kind of problem can be corrected by ignoring negative sentences (see Xie *et al*., 2002) but this increases computational time, the fraction of negative sentences is low, and there is never the security of detecting all negative sentences. For these reasons we decided not to apply any sentence filtering. Moreover, the application of the algorithm through our web server allows the user to examine the evidence used for the automatic assignment (linked terms, abstracts of articles, etc.), so that a mistake like this one can be easily spotted.

Derivation of GO terms from MeSH terms

The mapping that associated C, D and G MeSH terms with GO terms comprises 79 107, 1 016 973, and 297 832 pairs, respectively, in the α-cut of 0.1 of the relation of inclusion. A total of 6043 entries from the test set of 6526 were annotated with at least one GO term. Unexpectedly, the hierarchical structure of the GO terms, which is very appropriate to allow a flexible annotation process (Ashburner *et al*., 2000), made the prediction very complicated. In order to test this effect, we analyzed the distance in the GO hierarchy between the GO terms predicted and those of the benchmark set of 6043 database entries. We considered a match at a distance n in the hierarchy if, given a prediction with a common node in the GO hierarchy to the benchmark term, the number of steps from the predicted term to the benchmark term is n. For example, if the predicted term is ‘large ribosomal subunit’ (which depends on ‘ribosome’ which depends on ‘ribonucleoprotein complex’) and the benchmark term is ‘ribonuclease MRP complex’ (which depends on ‘ribonucleoprotein complex’), we consider a match with a distance of 3 in the hierarchy. Considering matches over longer scopes improved the precision of the results but not the recall for large values of α (Fig. 3). Considering only perfect matches and matches at distance 1 in the hierarchy, the 0.7-cut produced a precision of 67% and a recall of only 8% (Fig. 4). This is in agreement with the poor results obtained by previous approaches for the extraction of GO terms from the literature associated with genes (Pouliot *et al*., 2001; Raychaudhuri *et al*., 2002; Xie *et al*., 2002). See the Discussion section for a comparison of methods and results.

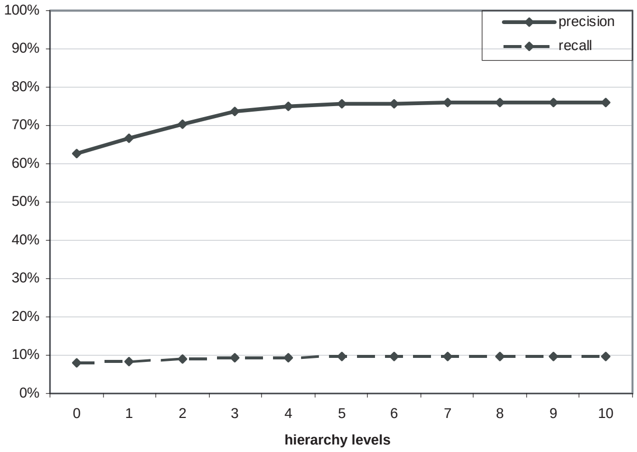


Fig. 3. Recall and precision in the prediction of GO terms for a 0.7-cut of the inclusion relation of the mapping between MeSH terms and GO terms considering several distances in the GO hierarchy between the predicted term and the benchmark term.

The annotation of the sequence GLTD_ECOLI is a good example of the variety of problems that the flexible structure of GO produces for automated annotation. This sequence is annotated with the following six GO terms: ‘glutamate synthase (NADPH)’, ‘electron transport’, ‘glutamate biosynthesis’, ‘disulfide oxidoreductase’, ‘oxidoreductase’ and ‘oxidoreductase, acting on the CH–NH_2 group of donors, NAD or NADP as acceptor’. Five GO terms were predicted, three of them were perfect matches: ‘electron transport’, ‘glutamate biosynthesis’ and ‘oxidoreductase’. Another one was ‘monooxygenase’ that depends on ‘oxidoreductase’ and, therefore, was considered to match at a distance of 1 in the hierarchy. The fifth prediction was ‘transaminase’ that is the grandchild of ‘enzyme’; since ‘oxidoreductase’ is the child of ‘transmaminase’ this was considered a match at distance 4.

Web server

The procedure was implemented on a public web server named KAT (Keyword Annotation Tool) that is accessible

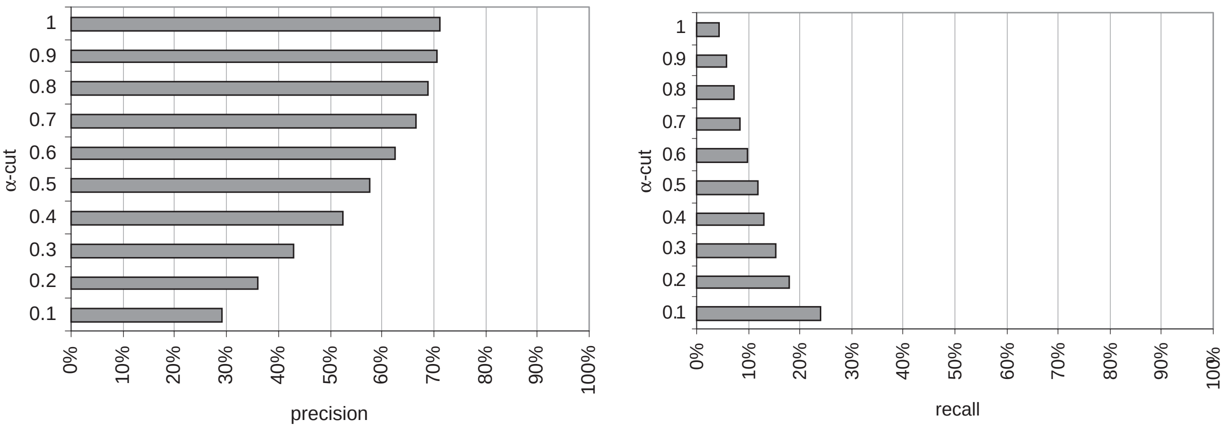
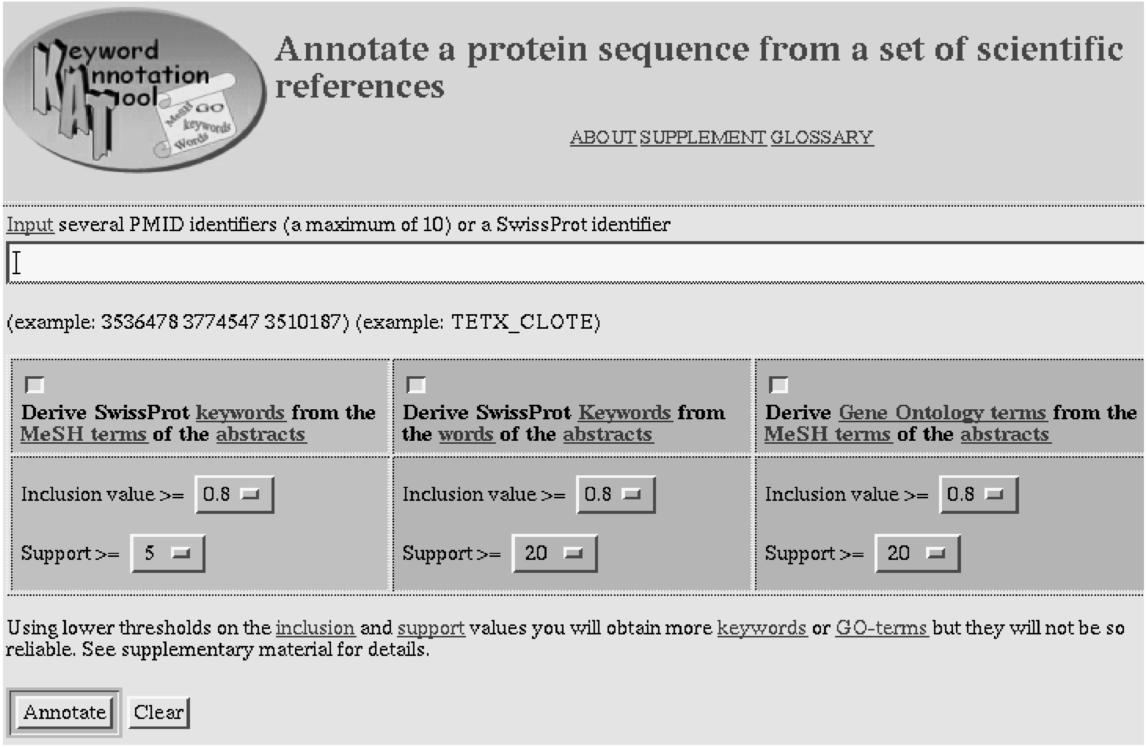


Fig. 4. Recall and precision in the prediction of GO terms for varying cuts of the inclusion relation of the mapping between MeSH terms to GO terms considering matches at distance 1 in the GO hierarchy.



Annotate a protein sequence from a set of scientific references

ABOUT SUPPLEMENT GLOSSARY

Input several PMID identifiers (a maximum of 10) or a SwissProt identifier

(example: 3536478 3774547 3510187) (example: TETX_CLOTE)

□ Derive SwissProt keywords from the MeSH terms of the abstracts	□ Derive SwissProt Keywords from the words of the abstracts	□ Derive Gene Ontology terms from the MeSH terms of the abstracts
Inclusion value >= 0.8 Support >= 5	Inclusion value >= 0.8 Support >= 20	Inclusion value >= 0.8 Support >= 20

Using lower thresholds on the inclusion and support values you will obtain more keywords or GO-terms but they will not be so reliable. See supplementary material for details.

Fig. 5. Home page of the KAT web server.

at www.bork.embl.de/kat. It annotates a sequence with both SWISS-PROT keywords and GO terms, based on MeSH terms and words extracted from a list of MEDLINE entries. The user can provide either the PMIDs (from PubMed identifiers) of the MEDLINE entries or a SWISS-PROT identifier. In the latter case, the references that are linked to the corresponding SWISS-PROT entry in the current SWISS-PROT version are considered. Just below the input entry there are three differently colored boxes that indicate the possible output options, that is SWISS-PROT keywords or GO terms deduced from the different mappings (Fig. 5). They allow the user to control the α-cut and minimum support of the relations used for the annotation.

DISCUSSION

In this work, we presented an algorithm that maps related terms between biological databases using the cross-links between two databases. We have developed, tested and distributed its application to the assignment of keywords (either

SWISS-PROT or GO terms) to molecular sequences from their association to one or more MEDLINE references. The benchmark of the system with SWISS-PROT entries indicates a very good performance for the annotation of SWISS-PROT keywords, with a recall of 47% and a precision of 68%. For the sake of comparison, we note that previous approaches that predicted keywords using as input the protein domains of the sequence and the taxonomy of the organism (Kretschmann *et al*., 2001; Pérez *et al*., 2002) reach a precision of $>90\%$ for a recall of 60%. The better performance of such algorithms is not surprising because it is easier to relate unequivocally protein domains to protein features—such as function (Camon *et al*., 2003) or protein cellular localization (Mott *et al*., 2002)—than to related literature. However, an approach based on the literature like ours is complementary to those because a protein sequence may not have detectable domains.

The performance of our system for the annotation of GO terms was much poorer than that of SWISS-PROT keywords, with a recall of 8% and a precision of 67% for an α-cut value of 0.7 of the relation of inclusion, even when considering the matching of the parent or the child of the GO term. However, as far as we know, we have presented, for the first time, a method for annotation of GO terms exclusively based on the literature and with a benchmark that is fully unbiased both in the set of sequences used and in the terms evaluated. For these reasons, the comparison with previous approaches for the annotation of GO terms from the literature, which use different source data and lack clean benchmarks, was complicated. For example, Xie *et al*. (2002) use a combination of literature, homology, domain mapping [taken from GOA, Camon *et al*. (2003)] and cellular localization of the protein for the prediction. Given an automated prediction of several GO terms for a sequence, the evaluation of the prediction considers only the correctness of the best predicted GO term (considering also correct the prediction of the parent or child) and they offer values by GO category instead of global. Values vary between 96–99% coverage and 65–80% reproducibility.

Raychaudhuri *et al*. (2002) use only the literature for the annotations but make a benchmark exclusively in yeast proteins and only for a restricted set of 12 GO terms with uneven results. Different to our approach, they extract GO terms from the consensus of several abstracts, and that generates problems when many of the associated references are uninformative. A similar problem was observed in the prediction of GO terms for groups of genes derived from microarray data from their associated literature Shatkay *et al*. (2000).

Pouliot *et al*. (2001) use a combination of manual mappings from protein domains and words from the SWISS-PROT entries (probably SWISS-PROT keywords but this is not explained in that work) to a manually constructed ontology. The system is far from automatic and its evaluation is also separated in different categories of their ontology that do not fully map to GO.

In summary, one advantage of our method is that it can be used to annotate a sequence with keywords using a very few abstracts, even just one, which is in fact the usual need of database curators. The time consumed in producing the annotation for a particular gene is negligible, because it is just a look up in a pre-computed table of pairs, and it is very easy to evaluate the annotation by examining the elements of the pairs in the mapping that are pointing to the keyword automatically selected. The annotator can discover very quickly if the suggested terms are appropriate or not, or, in the case of GO, select a more suitable level of the hierarchy if a too specific one has been produced. For this purpose we have made the system fully accessible through a web server that includes information about the pairs, scoring, and links to the databases used.

We found that the prediction of SWISS-PROT keywords was very much better than that of GO terms. It is difficult to predict GO terms with the precise specificity of the annotations in the database, a problem already mentioned (Raychaudhuri *et al*., 2002). We think that the different outcome in the benchmarks relies on intrinsic differences between a controlled vocabulary (SWISS-PROT keywords) and an ontology (GO terms). In principle, an ontology goes one step further than a simple controlled vocabulary by adding a structure, and therefore it should be preferred. However, it may be that the SWISS-PROT keywords are focused on the detail level of the information (or hierarchy level), to which the biologists are used (and that is found in the literature), whereas the GO terms cover exhaustively all levels of a hierarchy of biological concepts. As a result, the SWISS-PROT keywords are easier to fit to human annotations in other databases such as MeSH terms or words from abstracts in MEDLINE than the GO terms.

Although the obvious advantage of the GO terms is their exhaustive nature, sometimes the fact of choosing one level or another in the hierarchy does not follow an objective way of decision but obeys the subjectivity of the annotator. This fact imposes intrinsic limitations for automated annotation of GO terms using merely the literature: how can we make a correct prediction if two human annotators could easily annotate the same gene with different GO terms? The lesson is that the evaluation of methods for annotation of GO terms should account for the distance in the GO hierarchy from the predicted term to the term to be matched.

Our concluding point is that the system presented here can be used to pre-compute mappings between any pair of cross-linked databases. Once the map is obtained, it is very inexpensive to produce new annotations and the properties of the map allow the scoring of the annotations predicted and the tracking of their origin. We think that this approach will be very helpful since the number of databases and their size keeps increasing. This should encourage database developers to keep the communication between databases in the form of cross-links: the more the databases that are linked, and the more the links that exist, the easier it will be to keep all biological databases synchronized, using the most of their annotation capability.

ACKNOWLEDGEMENTS

We are also grateful to the developers and maintainers of SWISS-PROT, MeSH terms and Gene Ontology terms, to Helmut Schmid (IMS, Stuttgart University) for the implementation and public distribution of TreeTagger, and to the members of our group at EMBL-Heidelberg for fruitful discussions. We acknowledge the European Molecular Biology Organization, the European Molecular Biology Laboratory Staff association and the Junta de Andalucía for funding support.

REFERENCES

Andrade,M.A. and Valencia,A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J., Cox,A. and Apweiler,R. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.*, **13**, 1–11.

Kretschmann,E., Fleischmann,W. and Apweiler,R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17**, 920–926.

Miyamoto,S. (1990) *Fuzzy Sets in Information and Cluster Analysis*. *Theory and Decision Library*. Kluwer Academic Publishers, Dordrecht, Germany

Mott,R., Schultz,J., Bork,P. and Ponting,C.P. (2002) Predicting protein cellular localization using a domain projection method. *Genome Res.*, **12**, 1168–1174.

Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002a) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.

Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002b) Computing fuzzy associations for the analysis of biological literature. *Biotechniques*, **32**, 1380–1385.

Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2003) Mapping words for genome data integration. In Andrade,M.A. (ed.), *Bioinformatics and Genomes: Current Approaches*. Horizon Scientific Press, UK, pp. 141–152.

Pérez,A.J., Rodríguez,A., Trelles,O. and Thode,G. (2002) A computational strategy for protein function assignment which addresses the multidomain problem. *Comp. Funct. Genet.*, **3**, 423–440.

Pouliot,Y., Gao,J., Su,Q.J., Liu,G.G. and Ling,X.B. (2001) DIAN: a novel algorithm for genome ontological classification. *Genome Res.*, **11**, 1766–1779.

Raychaudhuri,S., Chang,J.T., Sutphin,P.D. and Altman,R.B. (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.*, **12**, 203–214.

Shatkay,H., Edwards,S., Wilbur,W.J. and Boguski,M. (2000) Genes, themes, and microarrays: using information retrieval for large-scale gene analysis. *Intell. Syst. Mol. Biol.*, **8**, 317–328.

Valencia,A. and Blaschke,C. (2002) Automatic ontology construction from the literature. *Genome Informatics*, **13**, 201–213.

Xie,H., Wasserman,A., Levine,Z., Novik,A., Grebinskiy,V., Shoshan,A. and Mintz,L. (2002) Large-scale protein annotation through gene ontology. *Genome Res.*, **12**, 785–794.

APPENDIX

Derivation of associations between words

The association between two words (w_i, w_j) can be modeled as the degree of inclusion of one word into the other which can be defined as the fuzzy binary relation, $\tilde{I}_W$, whose membership function is estimated as $\mu_{\tilde{I}_W}(w_i, w_j) = |W_i \cap W_j|/|W_i|$, i.e. the ratio of the number of transactions (in this case, a single SWISS-PROT entry) where both words w_i and w_j co-occur divided by the number of transactions where the word w_i occurs. This is an asymmetric relation very appropriate to model hierarchical relations between terms from thesauri (Miyamoto, 1990).

Given a fuzzy binary relation, the α-cut, where α is positive real and smaller than 1, is the subset composed from all pairs whose membership function value is equal or greater than α. The support of a pair is defined as the number of occurrences across all the transactions.

Computation of abstract-keywords

First, the text of the abstract is preprocessed with a part-of-speech tagger (Tree tagger, from Helmut Schmid, IMS, Stuttgart University, www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/) and nouns are selected. We define a fuzzy binary relation as described above with the only difference being that, here, we consider the sentence as the transaction unit putting pairs of words in relation. We compute the strength of the relations as the degree of inclusion (see above). Next, we identify a word as relevant for the text analyzed if it establishes many and strong relations to other words (Perez-Iratxeta *et al*., 2002b). Accordingly, we define a score for a word $\mathrm{w_i}$ that is equal to, $K_i = \sum_{j \neq i} \mu_{\tilde{I}_W}(w_j, w_i)$ normalized to the maximum value found for K of any word in that abstract. Finally, the keywords of the abstract are defined as those words that have a K score above an arbitrary value.