BIOINFORMATICS

Vol. 16 no. 2 2000
Pages 125–129

SAWTED: Structure Assignment With Text Description—Enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons

*Robert M. MacCallum, Lawrence A. Kelley and Michael J. E. Sternberg**

Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

Received on June 19, 1999; revised on September 2, 1999; accepted on September 10, 1999

Abstract

Motivation: *Sequence database search methods often identify putative sub-threshold hits of known function or structure for a given query sequence. It is widespread practice to filter these hits by hand using knowledge of function and other factors; to the expert, some hits may appear more sensible than others. SAWTED (Structure Assignment With Text Description) is an automated solution to this post-filtering problem which will be applicable to large scale genome assignments.*

Results: *A standard document comparison algorithm is applied to text descriptions extracted from SWISS-PROT annotations. The added value of SAWTED in combination with PSI-BLAST has been shown with a benchmark of difficult remote homologues taken from the SCOP structure database.*

Availability: *A SAWTED PSI-BLAST Web server is available to perform sensitive searches against the protein structure database (http://www.bmm.icnet.uk/servers/sawted).*

Contact: *R.MacCallum@icrf.icnet.uk*

Introduction

Increasingly there is a need to perform automated database searches for large numbers of protein sequences. The goal is to assign function and/or structure to a query sequence by homology to database proteins where this is already known. Of paramount importance to database search methods is a reliable means of distinguishing true hits (similarity due to homology) from false hits (similarity due to noise). In addition, the broadest possible coverage is desirable, in order that the maximum number of sequences can be assigned.

*To whom correspondence should be addressed; E-mail: M.Sternberg@icrf.icnet.uk

The widespread use of the BLAST family of programs (Altschul *et al.*, 1990), SSEARCH and FASTA (Pearson, 1990) can be attributed to their implementation of reliable confidence scores which estimate the probability of observing a false hit by chance. Recently, PSI-BLAST (Altschul *et al.*, 1997) has superseded these older programs in both reliability and coverage through the use of multiple alignment based profiles and an iterated approach (Park *et al.*, 1998).

Experience with blind predictions in the fold recognition section of the CASP3 experiment (Koehl and Levitt, 1999, http://PredictionCenter.llnl.gov/casp3) with both sequence-based and threading-type approaches has shown that hits with poor confidence scores and/or ranking can in fact be correct. In the realm of remote homology detection the signal and noise can be difficult to separate by automated means. With expert human intervention it can sometimes be a relatively simple (if time-consuming) task to compare information known about the query with the literature and database annotations associated with each potential hit. In the light of this comparison some hits may seem more promising than others. In practice at CASP3 this approach was often successful and widely used. However, to the non-expert or investigator with many sequences and too little time it is not an option.

This paper documents a straightforward method to automate the process described above by comparison of SWISS-PROT annotations. Thorough benchmarking of text comparison in combination with PSI-BLAST sequence retrieval shows the added value of this new method in the 'twilight zone'. The algorithm, known as SAWTED (Structure Assignment With Text Description), combined with PSI-BLAST is available on the Web for searches against the protein structure database.

Algorithm

The basic vector-cosine model of text retrieval described in Wilbur and Yang (1996) is used. This is also part of the algorithm used to calculate 'related articles' for PubMed at the NCBI (http://www.ncbi.nlm.nih.gov/PubMed). In summary, the similarity (c) between two texts (A and B) is calculated as the cosine of the angle between two vectors representing each text (v^A and v^B).

$$c = \cos\theta = \frac{v^A \cdot v^B}{|v^A| \times |v^B|}$$

The dimensions of the vectors (v_w) correspond to words in common between the two texts. Their magnitudes are the product of a local weight (l_w) and a global weight (g_w). The local weight reflects the frequency of the word in that particular text. The global weight reflects the information content of the word in the universe of texts (in this case the whole SWISS-PROT database) and is based inversely on the number of texts containing that word.

$$v_w = l_w \times g_w$$

$$l_w = 0.5 + 0.5(n_w/m)$$

$$g_w = \ln(T/t_w)$$

where n_w is the number of times word w occurs in the text, m is the maximum of n_w over all words w, T is the total number of texts (SWISS-PROT entries), and t_w is the number of texts containing word w. Thus high values for c are the result of matching rare words which occur frequently within each text.

Each entry of SWISS-PROT (Bairoch and Boeckmann, 1991) release 37.0 was pre-processed into 'words' as follows. Text from one or more of the RT (Reference Title), CC (Comments) and KW (Keywords) fields was extracted. The DE (description) line was not used because it contains very specific nomenclature unlikely to be shared between remote homologues. Non-text characters were removed except for dash characters within words. The following words and their plurals were then removed: protein, similarity, function, chain, domain, 3d-structure, immunoglobulin, fold, and, the, in, to, of, a, is, are, or, at, from. Text from the RT and CC fields was lowercased, while KW text was uppercased. The algorithm is case-sensitive therefore keywords are treated as separate words. Taxonomic words are not related to functionality but are often found in the processed text. They were collected from the OS and OC fields (Organism Species and Classification) and then removed from the processed text.

Implementation

SCOP benchmark

The SCOP structural database (Murzin *et al.*, 1995) provides a valuable resource for benchmarking sequence searching methods (Brenner *et al.*, 1998; Park *et al.*, 1998). The methods presented here have been tested on a set of query sequences (probes) also used in fold recognition experiments in our laboratory. The library against which they are searched (and from which the probes are also selected) is a representative set of 1560 protein domain sequences taken from SCOP of which no pair within the same SCOP family has more than 40% sequence identity (known as SCOP1560).

The set of probes is generated by selecting at random one or two domains from each superfamily. A second domain is permitted when it is not related to the first by sharing the same SCOP family or by a PSI-BLAST E-value (E_P) better than 0.1.

Many of the probes have 'easy' homologues in the library. In the benchmark we wish to concentrate on only the 'hard' probe–library pairs, so the easy pairs are simply ignored in all stages of the analysis. Easy sequence pairs are defined as those that can find each other using PSI-BLAST (in forward and/or reverse directions) with $E_P < 0.1$ or as sequences which have one or more SWISS-PROT homologues in common. As a result there are 152 probes which have at least one 'hard' homologue in the library to recognize, and a total of 786 possible pairs (on average five possible hits per probe).

Throughout this work (unless otherwise stated) PSI-BLAST has been run for 20 iterations against a non-redundant (at the level of identity) protein sequence database composed of the PDB, SWISS-PROT, TREMBL, PIR and the 1560 SCOP library sequences. The E-value cutoff for inclusion in the iterative model was 0.0005 and a maximum of 2000 hits were collected with $E_P < 500$. The low complexity filter SEG option was used.

Optimal use of descriptive information

To benchmark, a SWISS-PROT entry is assigned where possible to each library sequence (and hence for each probe) using the best scoring PSI-BLAST hit with $E_P < 0.1$. The library is then searched for each probe using the text comparison score as the sole discriminant, analogous to ranking by PSI-BLAST E-value or threading Z-score. In order to compare the effectiveness of different combinations of SWISS-PROT derived information (keywords, reference title and comments), errors vs coverage plots (Brenner *et al.*, 1998; Park *et al.*, 1998) are used. These plots are generated by sorting all probe–library pairs in reverse order of their cosine score (best first), then descending through the list plotting the number of erroneous pairs (non-homologues) vs the fraction of possible correct pairs (786) accumulated thus far. The errors are actually plotted as errors per query [dividing the number of errors by the number of probes (152)]. It is also possible to think of these plots as errors per query vs coverage at varying score thresholds.

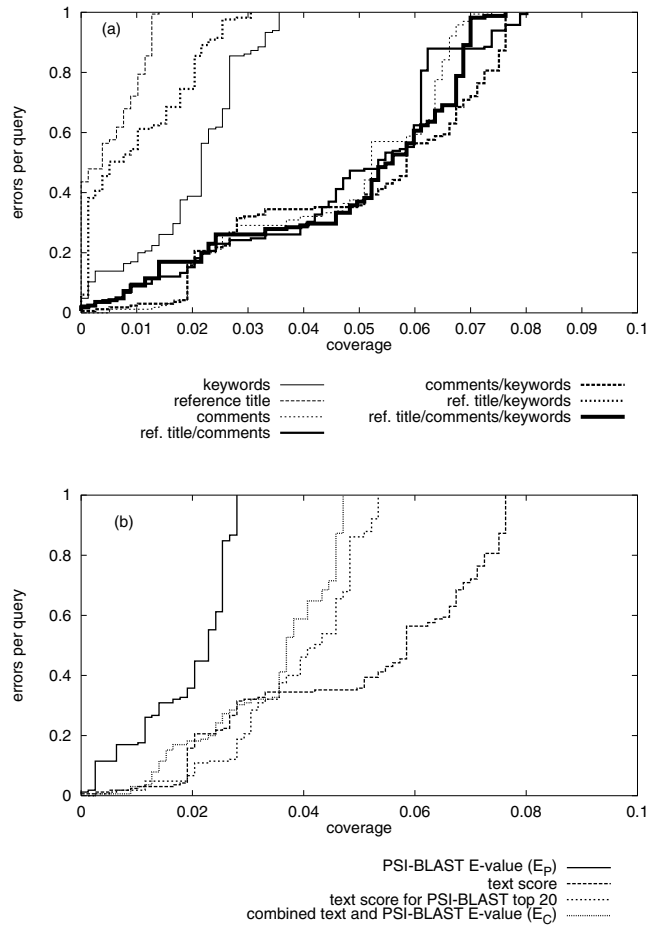


Fig. 1. Errors per query vs coverage plots for detection of remote homologues in the benchmark. Errors per query are calculated as the number of erroneous probe–library pairs with scores above a given threshold divided by the total number of queries (152). Coverage is the fraction of correct pairs found above a given score threshold. Each data point is generated using a different threshold. Lines to the right of the graph indicate better results. (a) Text comparison algorithm using various combinations of SWISS-PROT information. (b) Text comparison (comments/keywords) and PSI-BLAST in combination.

Figure 1a shows the relative performance of the seven combinations of text information. Used singly, SWISS-PROT comments are the most effective means of detecting remote homologues with the text comparison algorithm. The best multiple combination, by a small margin, is comments/keywords, although this is only seen at higher error rates.

SAWTED PSI-BLAST post-filter

How does text-based retrieval of remote homologues compare to sequence-based retrieval with PSI-BLAST? A direct comparison of errors vs coverage, shown in Figure 1b, favours the text-based approach (using comments/keywords). However, this result is not surprising given that the homologous pairs on which PSI-BLAST performs best have been excluded. Furthermore, text-based retrieval is totally dependent on the quality of the annotations; in the benchmark of structurally characterized proteins the annotation quality is generally high (but with little emphasis on structure, see below). More interesting and relevant is the performance of the text-based scoring for the putative homologues already identified by PSI-BLAST, but with poor E-values. Here, agreement between two methods using different sources of information (annotation and sequence) provides much stronger evidence for homology.

If the top 20 hits from the PSI-BLAST benchmark are re-scored using the text-based score there are fewer errors at greater coverage than with PSI-BLAST alone (Figure 1b). In other words, given a poor scoring PSI-BLAST hit, its text score is a more reliable indication of homology than E_P. Only three out of 200 matching words from these comparisons appear to contain explicit structural information (greek (key), plastocyanin-like, ZINC-FINGER).

How can the text score (c) be combined with E_P into a single number? One solution is to convert the text score into an E-value (E_T), and multiply the two E-values together ($E_C = E_P E_T$). When c is zero (because no words match or because SWISS-PROT annotations are not available) the most practical solution is for E_T to equal 1. In this way, the text score can only improve the combined E-value. Since the term E-value is synonymous with 'theoretical errors per query', c can be converted into E_T from the empirical relationship between errors per query and c, as shown in Figure 2. The relationship $E_T = e^{-12c}$ was chosen to fit part of this data (by eye) and also to satisfy $E_T = 1$ when $c = 0$. The best possible E_T is 6×10^{-6}, but a typical 'good' value for E_T is around 1×10^{-2}.

Figure 1b shows that the combined score, E_C, does not perform quite as well as the simple use of text scores for the PSI-BLAST hits. On a practical level however, E_C should be easier to use than the two separate scores.

Web server and CASP3 example

The combination of PSI-BLAST and text scoring described above has been tested on a benchmark of sequences with known structure. It is also made available on the Web (http://www.bmm.icnet.uk/servers/sawted) in a form which allows the user to search a sequence against the current PDB. The restriction to the PDB is warranted for two reasons. Firstly, if the database search is undertaken in order to find a possible function for the query sequence, then the closest SWISS-PROT homo-

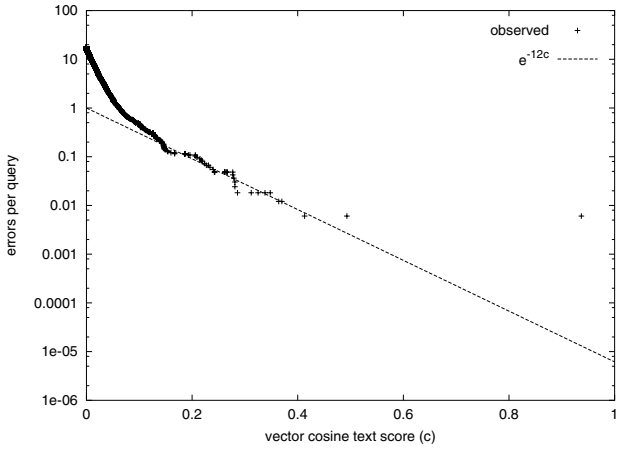
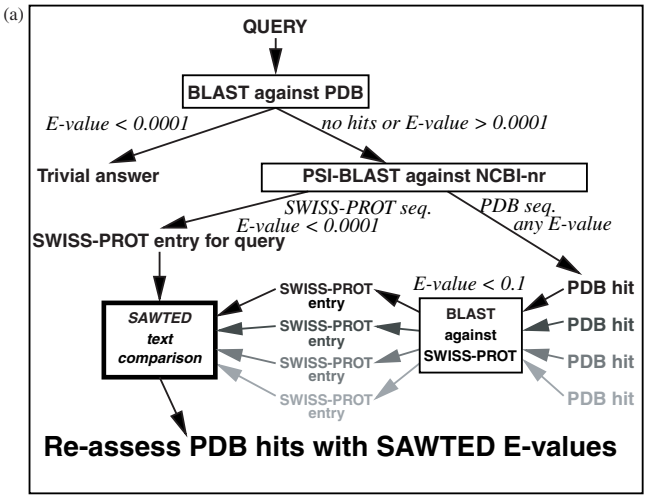


Fig. 2. Observed errors per query vs text score c and the theoretical E-value calculated as e^{-12c}, which approximates the observed data in the region of interest (errors per query <0.1).



(b) **Example: CASP3 query T0085 - C554_NITEU - cytochrome c554**
PSI-BLAST converges at 1 iteration
E_T and E_C discriminate correct answer (1FGJ-A) most clearly

rank	PDB	SWISS-PROT	E_P	E_T	E_C	best five matching words	correct
1	1FGJ-A	HAO_NITEU	4.0	0.0029	0.012	hydroxylamine hao per oxidation electrons	√
2	2CDV	CYC3_DESVM	6.8	0.054	0.37	per electrons heme molecule groups	×

Sentences from SWISS-PROT comments containing matching words:

C554_NITEU	INVOLVED IN AMMONIA OXIDATION; ACCEPTS ELECTRONS DIRECTLY FROM HYDROXYLAMINE OXIDOREDUCTASE (HAO)
HAO_NITEU	CATALYZES THE OXIDATION OF HYDROXYLAMINE TO NITRITE. THE ELECTRONS RELEASED IN THE REACTION ARE PARTITIONED TO AMMONIUM MONOOXYGENASE AND TO THE RESPIRATORY CHAIN. THE IMMEDIATE ACCEPTOR OF ELECTRONS FROM HAO IS CYTOCHROME C-554
CYC3_DESVM	PARTICIPATES IN SULFATE RESPIRATION COUPLED WITH PHOSPHORYLATION BY TRANSFERRING ELECTRONS FROM THE ENZYME DEHYDROGENASE TO FERREDOXIN. BINDS FOUR NONPARALLEL HEME GROUPS PER MOLECULE

Fig. 3. SAWTED PSI-BLAST Web server. (a) Flow diagram of internal workings. NCBI-nr is the non-redundant sequence database from the National Center for Biotechnology Information, USA. Note that the SWISS-PROT entry assigned to the query is the one with the greatest number of words (not the best E-value, as in the benchmark). (b) Results obtained from server at time of writing the first draft for CASP3 target T0085 (the revised server has regular updates, so now the 'answer' is trivially found in the PDB). E_P, E_T and E_C are the PSI-BLAST, SAWTED and combined E-values, respectively.

logue to the query (required for the text comparison) will already hold the functional information which is sought after. Secondly, although the approach might identify interesting remote homologies between sequences of unknown structure, the potentially heavy load on our server cannot be justified.

The implementation of the Web server is shown in Figure 3a and described briefly here. If no obvious homologue exists in the PDB, the query sequence is searched against a large sequence database (NCBI-nr) using 10 iterations of PSI-BLAST. If PDB hits occur in the output then the text comparison is performed between SWISS-PROT entries corresponding to each of them (found using BLAST) and a SWISS-PROT entry homologous to the query (from the original PSI-BLAST output). Text descriptions provided by the user will be used if no such homologue exists. The results, including the text comparison E-values and a pseudo-multiple alignment formatted with MView (Brown *et al.*, 1998) from the PSI-BLAST output are given to the user on a Web page (alerted by email). Databases are updated weekly from the NCBI.

The results obtained by submitting CASP3 target T0085 (cytochrome c554 from *Nitrosomonas europae*, http://PredictionCenter.llnl.gov/casp3/targets/templates/t0085.doc.html) to the server (as it existed before revision of the manuscript) are summarized in Figure 3b. SAWTED was developed after the CASP3 experiment, and it should be stressed that the example shown was not submitted as a blind prediction. The PSI-BLAST hits for T0085 are all cytochrome c-type proteins. The most correct hit is the rank one hit to 1FGJ-A (hydroxylamine oxidoreductase, HAO). The arrangement of the haem groups and their surrounding secondary structures is similar enough to suggest common ancestry (Iverson *et al.*, 1998). The other hits are to cytochrome c_3 structures. While PSI-BLAST E-values do not clearly distinguish the correct answer, there is approximately a 20-fold difference in SAWTED E-values. The five most significant words in common between the SWISS-PROT annotations for c554 and HAO are hydroxylamine, hao, per, oxidation, electrons. Examination of the SWISS-PROT annotations (see Figure 3b), shows that the words hydroxylamine and hao are found in comments about the electron transport

chain in which c554 and HAO are neighbours. This is an interesting result: sharing weak sequence similarity and the same pathway would appear to be sufficient evidence for gene duplication. Our benchmark shows that false positives from pathway neighbours are not a significant problem.

Discussion

SAWTED text scoring is already implemented in the 3D-PSSM fold recognition Web server offered by our laboratory (http://www.bmm.icnet.uk). We consider that SAWTED scores would substantially improve fully automated fold recognition which was proposed at CASP3 (Fischer *et al.*, 1999). Anyone wishing to apply SAWTED comparisons in their work should contact the first author (R.MacCallum@icrf.icnet.uk) for the relevant programs.

The next generation of SAWTED would take advantage of information available in MEDLINE abstracts for characterized proteins where SWISS-PROT annotations are unavailable. Along these lines, Andrade and Valencia (1998) have recently developed a system for extracting SWISS-PROT-like annotations from multiple MEDLINE abstracts.

It should be stressed that both text and sequence comparisons have the potential to proliferate fictitious annotations throughout the databases (Doerks *et al.*, 1998). The use of SAWTED is expected to alleviate rather than worsen this problem because the probability of detecting a false relationship simultaneously with text and sequence information is low. Furthermore, annotations made with SAWTED can be explicit in stating, via *E*-values, the relative contributions of text and sequence information.

There are, of course, limitations in the SAWTED approach. The text-matching algorithm is simplistic and discards word order information. Remote homologies may also be missed if the functions have become too different or if the proteins have acquired completely new functions. Conversely, text comparisons may produce false hits where unrelated proteins have similar textual descriptions as the result of convergent evolution. However because SAWTED is used as a post-filter for database searches, these problems do not have any serious practical implications. As with all database search methods, some homologies will be missed, or a small fraction of assignments (relating to the *E*-value threshold chosen) will be incorrect. In genomic and other large-scale automated database searches, the modest fractional increase in coverage given by SAWTED post-filtering will lead to the discovery of numerous remote homologies giving functional insight to previously uncharacterized proteins.

Acknowledgements

R.M.M. is supported by the Imperial Cancer Research Fund. L.A.K. is supported by Glaxo-Wellcome.

References

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.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

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.

Bairoch,A. and Boeckmann,B. (1991) The SWISS-PROT protein-sequence data-bank. *Nucleic Acids Res.*, **19**, 2247–2248.

Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.

Doerks,T., Bairoch,A. and Bork,P. (1998) Protein annotation: detective work for function prediction. *Trends Genet.*, **14**, 248–250.

Fischer,D., Barret,C., Bryson,K., Elofsson,A., Godzik,A., Jones,D., Karplus,K.J., Kelley,L.A., MacCallum,R.M., Pawowski,K., Rost,B., Rychlewski,L. and Sternberg,M. (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Protein Struct. Funct. Genet.*, **S3**, 209–217.

Iverson,T.M., Arciero,D.M., Hsu,B.T., Logan,M.S.P., Hooper,A.B. and Rees,D.C. (1998) Heme packing motifs revealed by the crystal structure of the tetra-heme cytochrome c554 from *Nitrosomonas europaea*. *Nat. Struct. Biol.*, **5**, 1005–1012.

Koehl,P. and Levitt,M. (1999) A brighter future for protein structure prediction. *Nat. Struct. Biol.*, **6**, 108–111.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP—a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.

Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.*, **183**, 63–98.

Wilbur,W.J. and Yang,Y.M. (1996) An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput. Biol. Med.*, **26**, 209–222.