W214–W218 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl332

DOUTfinder—identification of distant domain outliers using subsignificant sequence similarity

Maria Novatchkova[1,*], Georg Schneider[1], Richard Fritz[2], Frank Eisenhaber[1] and Alexander Schleiffer[1]

[1]Research Institute of Molecular Pathology (IMP), Dr Bohr-Gasse 7, A-1030 Vienna, Austria and
[2]Institute of Virology, Medical University of Vienna, Kinderspitalgasse 15, A-1095 Vienna, Austria

Received February 15, 2006; Revised March 4, 2006; Accepted April 14, 2006

ABSTRACT

DOUTfinder is a web-based tool facilitating protein domain detection among related protein sequences in the twilight zone of sequence similarity. The sequence set required for this analysis can be provided by the user or will be collected using PSI-BLAST if a single sequence is given as an input. The obtained sequence family is analyzed for known Pfam and SMART domains, and the thereby identified subsignificant domain similarities are evaluated further. Domains with several subthreshold hits in the query set are ranked based on a sum-score function and likely homologous domains are suggested according to established cut-offs. By providing a post-filtering procedure for subsignificant domain hits DOUTfinder allows the detection of non-trivial domain relationships and can thereby lead to new insights into the function and evolution of distantly related sequence families. DOUTfinder is available at http://mendel.imp.ac.at/dout/.

INTRODUCTION

Domains are evolutionary conserved building blocks within protein sequences, which typically represent discrete structural and functional units therein. As domains have been repeatedly duplicated and reused during evolution two-third of all known proteins can be reliably assigned to at least one of several thousand already characterized domain families thereby providing an initial indication on molecular or cellular function (1).

Nonetheless the theoretically possible coverage of domain based annotation is likely not yet fully exploited. It has been suggested that around 90% of whole proteome residues are participating in globular domains (2). This is opposed to 50% of residues from all proteins that can be assigned to a known domain to date (1). The globular regions, which remain unannotated by domain based searches, are in part distantly related to known domains, and are therefore distant outliers of characterized domain families. Such domain outliers (DOUTs) represent true homologs of those families but have diverged too far away from the described consensus in order to be significantly hit in profile-based searches of a sequence against Pfam (3) or SMART (4) domain database. DOUTs are often found as false negative similarities in the twilight zone of homology searches. It is therefore common practice to analyze subsignificant domain hits individually and evaluate them on the basis of additional knowledge. DOUTfinder was developed to facilitate this latter analysis step by providing a homology-backed procedure for post-filtering of relevant subthreshold hits. In the following we demonstrate the ability of this tool to efficiently separate a fraction of potential true domain similarities from the noise in the twilight zone of similarity searches. For this purpose we introduce a scoring scheme for the evaluation of subsignificant domain hits in a group of homologs and calibrate it using a widely applied distant homology control set: the Astral SCOP 1.69 database (5).

THE METHOD: SPOTTING DISTANT DOMAIN HOMOLOGS

Commonly used domain database search facilities, such as Pfam, SMART and CDD (6) provide extremely reliable domain annotations when run with default threshold settings. In order to increase search sensitivity it is recommended to use relaxed thresholds and evaluate the obtained results individually in a consecutive step. This post-filtering process is typically performed using additional information, such as functional, contextual and taxonomic data (7). It is also customary practice to support subsignificant domain hits by likewise subthreshold matches among clear sequence homologs

*To whom correspondence should be addressed. Tel: +43 1 7973 0556; Fax: +43 1 7987 153; Email: novatchkova@imp.univie.ac.at of the initial query. We have challenged the latter approach in a test-case based on the SCOP protein classification and defined conditions for which the co-occurrence of subsignificant domain hits within a protein family is a reliable measure for a similarity to that domain.

SCOP is a database commonly used in the evaluation of distant sequence comparison tools, as it provides a hierarchy of proteins beyond obvious sequence similarities. SCOP classifies structural domains within proteins into four hierarchical levels: families, superfamilies, folds and classes. Protein families consist of closely related sequences and are further grouped into superfamilies of presumed monophyletic origin. Folds subsume superfamilies with common topology and unclear evolutionary relationship. We used the ASTRAL SCOP 1.69 dataset and supplemented sparsely populated protein families using the 30 nearest sequence neighbors identified in BLAST against a 80% non-redundant Uniref variant (8,9). Protein families were analyzed for significant ($E \leq 0.005$) as well as subsignificant ($E > 0.005$) domain hits using RPS-BLAST against the Pfam database (10). Co-occurrence of two or more subsignificant hits within a family could be either supported: by a significant domain hit in any other family of the same superfamily, or disproved: if a significant domain hit appeared in another fold and never in the fold the family belongs to. Ninety-four percent of all disproved domain similarities with an expect value between 0.005 and 20 showed a domain coverage of 0.4 or less (84% 0.3 or less), where the domain coverage is the length of the aligned domain segment versus the domains consensus length. To avoid overpopulation of the analysis by false-negative hits the default domain-coverage threshold for considering subsignificant domain hits is set to 0.4 at the DOUTfinder web server.

The *D*-sum score was introduced as a quantitative estimate rating the quality of a domain outlier prediction for a protein family F with multiple subsignificant hits of a domain D. The *D*-score interprets the sum-scores, S_r, for all occurrences, *r*, of a domain D within a family F normalized by λ (11), and penalizes for the size of the search space, *mn*, where *n* is the database length. The query length *m*, and the constants λ and κ are calculated for a concatenation product of all residues within family F. A reward proportional to the sum of the average domain coverage, C_r, and the ratio of domain instances to the number of all proteins, *N*, within the family is applied.

$$D = \lambda \sum_{i=1}^{r} S_r - \ln(\kappa mn) + \left(\sum_{i=1}^{r} C_r + \frac{r}{N} \right) r^2$$

ASTRAL dataset analysis was used to evaluate the discriminative power of the *D*-score, and to obtain recommended *D*-score thresholds corresponding to expected 5 or 10% error rates. For this purpose *D*-scores were calculated for all subsignificant domain hits appearing more than once in RPS-BLAST searches of ASTRAL families versus Pfam ($0.005 < E < 20$, $C > 0.4$). The relationships between ASTRAL families and domains were then classified as supported or disproved. Figure 1 illustrates the clear separation between supported or disproved homology predictions using the *D*-score. Based on this assignment 5 and 10% error rate cut-offs were calculated. These thresholds are used by the DOUTfinder webserver to delineate probable and potential domain outlier predictions, respectively.

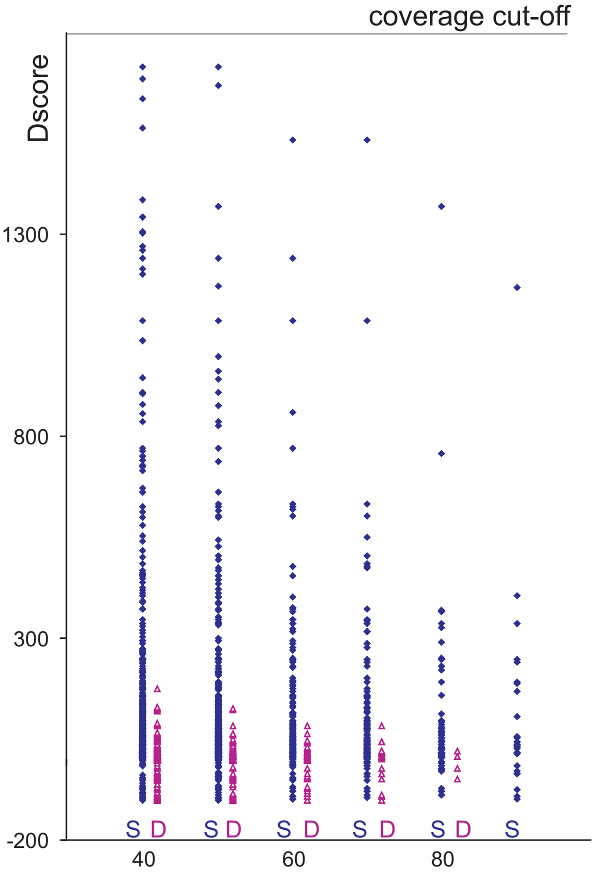
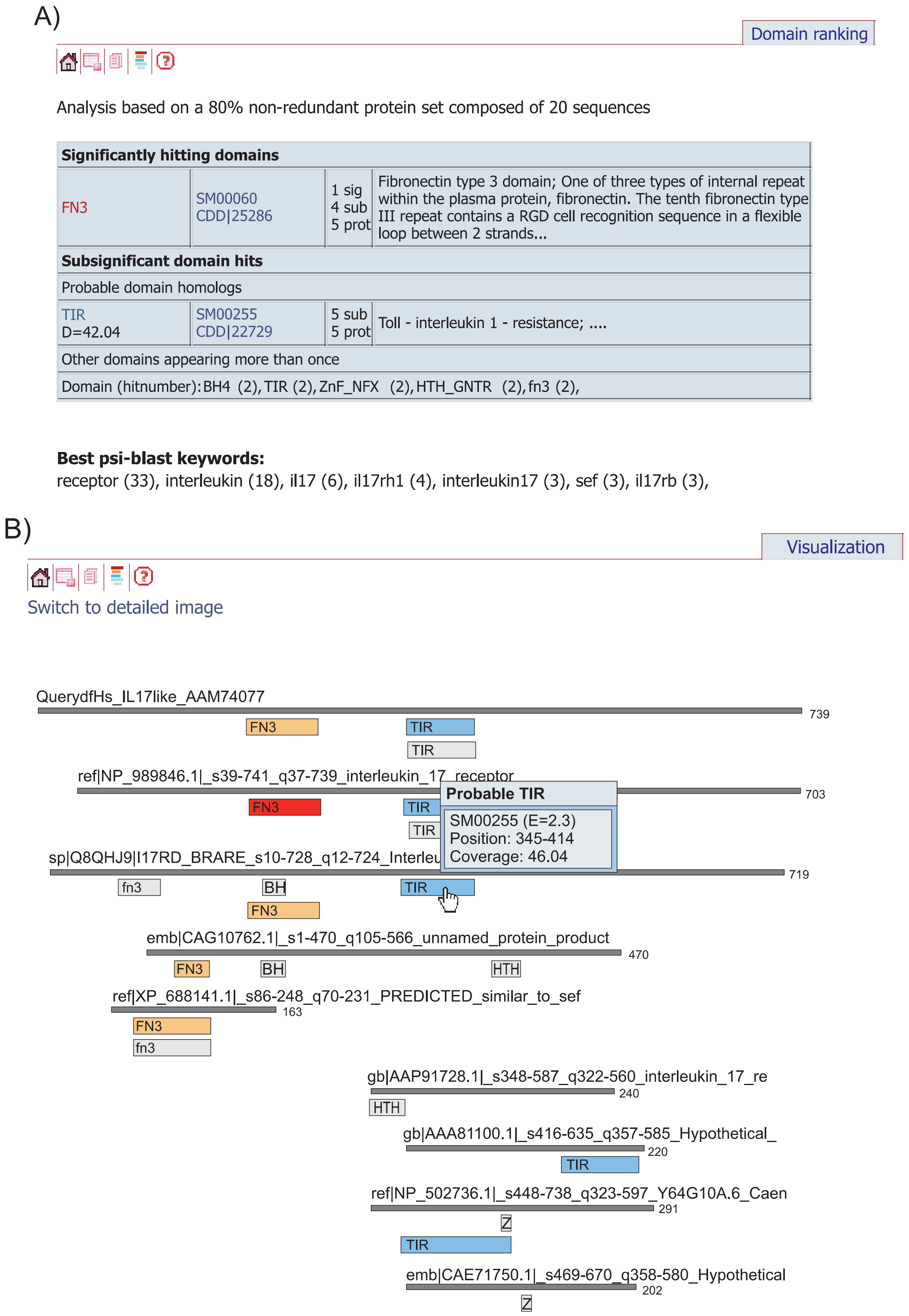


Figure 1. *D*-score distribution of potentially homologous supported (S) and non-homologous disproved (D) domains with multiple occurrences in individual ASTRAL families upon variation of the coverage cut-off ($0.005 < E < 20$). Supported and disproved domains are well separated over a wide range of *D*-scores.

The validity of the approach and of the defined cut-offs was further evaluated by a DOUTfinder analysis of 1462 domains of unknown function (DUF) derived from the Pfam18 dataset, of which 1434 retained more than one sequence after redundancy removal at a 80% identity threshold. Analysis of the subsignificant domain hits of these DUF families resulted in the suggestion of around 80 probable and around 20 potential domain outliers. In ~20% of these cases the prediction could be confirmed by a PSI-BLAST link between the domains. Approximately 35% of the DUF similarities to other Pfam domains were also detected by the profile–profile-based Clans assignment provided since Pfam19 (3). A complete listing of the suggested domain similarities can be accessed on the DOUTfinder website. The agreement of DOUTfinder predictions with Clans relationships and the even higher sensitivity in more than half of the established cases of DUF domain relationships indicates that DOUTfinder is a useful complementation to other available methods. It should be noted that as pointed out in the original Clans report, Pfam PRC profile–profile comparison (http://supfam.mrc-lmb.cam.ac.uk/PRC/) with its current settings has not yet reached its maximal sensitivity.



A)

Domain ranking

Analysis based on a 80% non-redundant protein set composed of 20 sequences

Significantly hitting domains			
FN3	SM00060 CDD|25286	1 sig 4 sub 5 prot	Fibronectin type 3 domain; One of three types of internal repeat within the plasma protein, fibronectin. The tenth fibronectin type III repeat contains a RGD cell recognition sequence in a flexible loop between 2 strands...
Subsignificant domain hits			
Probable domain homologs			
TIR D=42.04	SM00255 CDD|22729	5 sub 5 prot	Toll - interleukin 1 - resistance; ....
Other domains appearing more than once			
Domain (hitnumber): BH4 (2), TIR (2), ZnF_NFX (2), HTH_GNTR (2), fn3 (2),			

Best psi-blast keywords:
receptor (33), interleukin (18), il17 (6), il17rh1 (4), interleukin17 (3), sef (3), il17rb (3),

B)

Visualization

Switch to detailed image

QuerydfHs_IL17like_AAM74077
739
FN3
TIR
TIR
ref|NP_989846.1|_s39-741_q37-739_interleukin_17_receptor
703
FN3
TIR
TIR
Probable TIR
SM00255 (E=2.3)
Position: 345-414
Coverage: 46.04
sp|Q8QHJ9|I17RD_BRARE_s10-728_q12-724_Interleu
719
fn3
BH
TIR
FN3
emb|CAG10762.1|_s1-470_q105-566_unnamed_protein_product
470
FN3
BH
HTH
ref|XP_688141.1|_s86-248_q70-231_PREDICTED_similar_to_sef
163
FN3
fn3
gb|AAP91728.1|_s348-587_q322-560_interleukin_17_re
240
HTH
gb|AAA81100.1|_s416-635_q357-585_Hypothetical_
220
TIR
ref|NP_502736.1|_s448-738_q323-597_Y64G10A.6_Caen
291
Z
TIR
emb|CAE71750.1|_s469-670_q358-580_Hypothetical
202
Z

Figure 2. DOUTfinder output for the example analysis of human SEF receptor as a single sequence input. (**A**) tabular (**B**) graphical sections. Subsignificant domains can be either confirmed by a significant hit somewhere else in the set, as is the case with the fibronectin type three domain in the current example, or by a sufficiently high *D*-score. *D*-score evaluation is facilitated by the use of two thresholds: *D*-score predictions above the 5 and 10% false positive limit are interpreted as probable and potential domain homologs, respectively.

Example of usage: DOUTfinder single sequence analysis

The DOUTfinder web server implements the analysis of subsignificant domain hits using a *D*-score as described above. Two types of input are accepted—either a family of homologous sequences collected by the user, or a single sequence, which is used to collect homologous segments in a non-redundant protein database. For the following illustrations the full-length human IL17/SEF receptor protein (AAM74077) is used as a single sequence input to DOUTfinder. In this example the analysis of subsignificant domain hits of SEF family members can identify an intracellular region of similarity to the Toll/interleukin-1 receptor (TIR) homology domain in agreement with previous observations (12).

When a single input is provided DOUTfinder automatically and successively starts a series of steps, which do not require further user-intervention and lead to the retrieval of a homologous sequence set, its domain analysis and the domain outlier identification. According to the default parameterization the set collecting tool of DOUTfinder applies two rounds of PSI-BLAST search against a non-redundant database to obtain segments with IL17/SEF homology (13). The used non-redundant database is generated using NCBI nr (at various levels of non-redundancy) as well as Pfam and Smart domain sequences (14). Thereby the initial PSI-BLAST step can also be used to link the submitted sequence to known domains via a logically inverted profile-based search, where the query protein provides the profile against which the individual domain sequences can be matched. Upon completion of the PSI-BLAST search this initial protein set is filtered up to 80% non-redundancy—a setting which is user-adjustable (8). The obtained representative sequences are filtered using the optionally applied COILS and HMMTOP algorithms (15,16) and supplied to domain-analysis using RPS-BLAST in a search against SMART (4) and PFAM (3) databases. Domains are evaluated based on their score and graphical and textual reports are prepared.

Example output

The output of DOUTfinder domain analysis consists of a tabular and a graphical part. In the graphical part proteins are represented as bars and domains are color-coded according to the similarity category they belong to (i) significant RPS-BLAST similarity—red boxes (ii) subthreshold hits supported by a significant hit somewhere else in the homologous set—orange boxes (iii) probable domain outliers with a *D*-score above the 5% error limit—blue boxes (iv) potential domain outliers with a *D*-score above the 10% error limit—cyan boxes (v) other domains found more than once—gray boxes (vi) single occurrence domains—white boxes. Mouse-over functions provide additional information on the domains, and link to the original domain databases. In the example of SEF homologs twilight zone similarities (orange) to the fibronectin type 3 domain can be supported by a significant hit in one of the proteins in the set (red) (Figure 2B). The TIR domain is identified as a probable domain outlier with five subsignificant hits in five of the analyzed 20 sequences.

In addition to the graphical output the identified domain similarities are also presented in two types of tabular output, which are structured and colored analogously to the graphical one. The short tabular output provides comprehensive functional annotation of the domains, which support the fast evaluation of the obtained hits (Figure 2A). Further expert evaluation is assisted by the PSI-BLAST keyword assessment, a PSI-BLAST domain hit evaluation and listing of those domains within the set, which belong to the same Pfam CLAN. This information is provided below the short domain summary if applicable. An extensive listing of the obtained domain hits is provided in the second tabular output. Various links allow fast switching between the result sections.

CONCLUSIONS

Sensitive domain detection typically relies on the use of curated consensus representations of known protein families, such as PSSMs and profile HMMs (17). The indisputable advantage of these approaches compared to pairwise sequence comparison lies in the integrated description of multiple sequence information in one statistical representation. However a single domain model will likely be less sensitive in uncovering atypical homologs (18), which can arise in families with differing evolutionary speed and high diversification into a non-homogenous sequence space. The sensitivity of a profile-based search will also be hampered by domain definitions based on a small domain family with few members, represented in one taxon only, or features that are too short and therefore lead to an incorrect sequence alignment. In such cases biologically relevant twilight zone similarities can remain below recommended significance thresholds. By analyzing such subsignificant relationships DOUTfinder can identify distant sequence similarities and potentially lead to true remote homologs that could have otherwise been missed.

ACKNOWLEDGEMENTS

The authors are grateful for generous support from Boehringer Ingelheim. This project has been partly funded by the Austrian Gen-AU bioinformatics integration network. Funding to pay the Open Access publication charges for this article was provided by the Research Institute of Molecular Pathology.

Conflict of interest statement. None declared.

REFERENCES

1. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
2. Copley,R.R., Doerks,T., Letunic,I. and Bork,P. (2002) Protein domain analysis in the era of complete genomes. *FEBS Lett.*, **513**, 129–134.
3. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
4. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
5. Chandonia,J.M., Hon,G., Walker,N.S., Lo,C.L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.

6. Marchler-Bauer,A. and Bryant,S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.*, **32**, W327–W331.
7. Coin,L., Bateman,A. and Durbin,R. (2004) Enhanced protein domain discovery using taxonomy. *BMC.Bioinformatics.*, **5**, 56.
8. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics.*, **17**, 282–283.
9. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
10. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D., Ke,Z. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.*, **33**, D192–D196.
11. Altschul,S.F. (1997) Evaluating the statistical significance of multiple distinct local alignments. In Suhai,S. (ed.), *Theoretical and Computational Methods in Genome Research.* pp. 1–14.
12. Novatchkova,M., Leibbrandt,A., Werzowa,J., Neubuser,A. and Eisenhaber,F. (2003) The STIR-domain superfamily in signal transduction, development and immunity. *Trends Biochem. Sci.*, **28**, 226–229.
13. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
14. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **34**, D173–D180.
15. Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics.*, **17**, 849–850.
16. Lupas,A. (1996) Prediction and analysis of coiled-coil structures. *Meth. Enzymol.*, **266**, 513–525.
17. 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.
18. Murvai,J., Vlahovicek,K., Barta,E., Szepesvari,C., Acatrinei,C. and Pongor,S. (1999) The SBASE protein domain library, release 6.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **27**, 257–259.