doi:10.1093/nar/gkn320

# SCANPS: a web server for iterative protein sequence database searching by dynamic programing, with display in a hierarchical SCOP browser

Thomas P. Walsh[1], Caleb Webber[2,3], Stephen Searle[2,4], Shane S. Sturrock[1,5] and Geoffrey J. Barton[1,*]

[1]College of Life Sciences, University of Dundee, Dundee DD1 5EH, [2]EMBL-European Bioinformatics Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, [3]Departments of Physiology, Anatomy and Genetics, MRC Functional Genetics Unit, University of Oxford, South Parks Road, Oxford OX1 3QX, [4]The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK and [5]Biomatters Ltd, Level 6 FAI Building, 220 Queen St, Auckland 1001, New Zealand

Received January 31, 2008; Revised April 27, 2008; Accepted May 7, 2008

## ABSTRACT

**SCANPS performs iterative profile searching similar to PSI-BLAST but with full dynamic programing on each cycle and on-the-fly estimation of significance. This combination gives good sensitivity and selectivity that outperforms PSI-BLAST in domain-searching benchmarks. Although computationally expensive, SCANPS exploits onchip parallelism (MMX and SSE2 instructions on Intel chips) as well as MPI parallelism to give acceptable turnround times even for large databases. A web server developed to run SCANPS searches is now available at http://www.compbio.dundee.ac.uk/www-scanps. The server interface allows a range of different protein sequence databases to be searched including the SCOP database of protein domains. The server provides the user with regularly updated versions of the main protein sequence databases and is backed up by significant computing resources which ensure that searches are performed rapidly. For SCOP searches, the results may be viewed in a new tree-based representation that reflects the structure of the SCOP hierarchy; this aids the user in placing each hit in the context of its SCOP classification and understanding its relationship to other domains in SCOP.**

## INTRODUCTION

SCANPS is a program for comparing a protein sequence to a sequence database. It performs iterative profile searching similar to PSI-BLAST (1), but with full dynamic programing on each cycle and on-the-fly estimation of significance. The SCANPS web server has been developed to simplify the running and analysis of SCANPS searches. An innovative aspect of the server is its novel tree-based presentation of results for searches against the SCOP domain database (2). A comparison of a protein domain to the domains in SCOP can be of considerable value in elucidating its structure and function. Facilities for comparing a query sequence to SCOP sequences are also provided by FPS (http://fps.sdsc.edu/), GTOP (http://spock.genes.nig.ac.jp/~genome/grpsblt.html) and CascadeBlast (http://crick.mbu.iisc.ernet.in/~CASCADE/CascadeBlast.html), all of which use PSI-BLAST as the search algorithm. However, interpreting the results of SCOP searches in the context of the classification is hampered by the fact that search methods typically produce a linear table of hits; understanding the relationships between hits to the SCOP database usually requires a manual mapping of the results table onto the SCOP hierarchy. This procedure is tedious and error prone and therefore a browser interface has been developed that maps hits onto the SCOP hierarchy for viewing using a tree-based framework. The new interface allows the user to perform a search of the complete non-redundant SCOP sequence database and view results in a form that allows the information inherent in the SCOP classification to be exploited in interpreting those results.

## MATERIALS AND METHODS

### Overview of SCANPS

Although SCANPS has been available for over 15 years and has been accessible as a service at the European

*To whom correspondence should be addressed. Tel: +01382 385860; Fax: +01382 385764; Email: geoff@compbio.dundee.ac.uk

Bioinformatics Institute (EBI) for 10 years, it has not previously been described in the literature. Accordingly, as background to the new web server, a brief overview of the motivation, novel features and performance of the program is given here.

The basic function of SCANPS is to perform a full Smith–Waterman algorithm (3) comparison of a protein sequence to a protein sequence database with either length dependent or affine gap penalties (4). The program is written in C and effort was put into coding for performance on conventional workstation hardware. This enabled the program to be used routinely for searching large databases on modest computer hardware in contrast to the belief in the early 1990s that Smith–Waterman was too CPU intensive for this task (5). Parallel processing by dynamically splitting the database across multiple processors was demonstrated on a network of five loosely coupled workstations (6), then refined to exploit Symmetric Multi Processing (SMP) hardware via OpenMP (7). The SMP implementation gave near linear speedup on a 24 processor Silicon Graphics Challenge. With the move to commodity PC hardware in the late 1990s, fourway onchip parallelism was implemented on Intel and AMD chips with a speedup over linear code of at least 3× depending on the query and database size. In addition to the onchip parallelism, multiprocessor parallelism was implemented through MPI (8,9). The MPI implementation gave parallel efficiency of over 90% on 16 processors on an Intel PIII cluster connected by 100 MB network.

The speed obtained by parallel processing, coupled with the on-the-fly statistics described below, permitted the implementation of iterative searching. In this mode, a multiple sequence alignment is constructed for sequences that score above a preset significance threshold in the initial search. The multiple alignment is built up by aligning to the query sequence as a template, but using a Position Specific Scoring Matrix (PSSM) as appropriate at each iteration. Alignment columns which contain gaps in the query sequence are deleted; as a result, the alignment is always the same length as the query sequence. A PSSM (10,11) is derived from this alignment and used to re-search the database. Since sequences that are very similar to the query sequence contribute little information to the PSSM, a percentage identity threshold (PIDT) is employed which excludes all sequences from the alignment whose similarity to the query sequence exceeds the threshold. The contributions of each sequence to the PSSM are weighted according to the method of Henikoff and Henikoff (12). The scoring matrix is then constructed at each alignment position similarly to standard log odds matrices (12,13). The process of constructing a PSSM and searching the database is repeated until convergence, or until a preset number of iterations has been completed.

In each database search, the statistical significance of a score between the query and any sequence in the database is assessed by on-the-fly modeling the distribution of query database sequence pair scores. Scores are binned according to the log of the product of the query and database sequence lengths (LPL). Within each LPL bin, an extreme value distribution is fitted to the scores. The extreme-value location and scale parameters are then fitted to exponential and linear equations respectively with respect to the LPL. The resulting extreme-value equations are applied back to all query sequence pairs and the resulting probabilities converted to *E*-values for ranking and display. This method of estimating significance is similar to those implemented in FASTA/SSEARCH (14); for a full discussion of the similarities and differences between the different fitting schemes and the effect on performance in benchmarks see (15).

## Benchmarking of SCANPS

The search performance of SCANPS was compared with that of PSI-BLAST using a benchmark based on SCOP (2). The benchmark dataset comprises 1113 sequences taken from the PDB40D-B dataset constructed by Brenner *et al.* (16). Single-segment domains whose structure had been determined by X-ray crystallography were selected, representing 479 SCOP superfamilies across 343 SCOP folds. True positives were defined as pairs of sequences belonging to the same SCOP superfamily; true negatives were defined as those pairs in which the sequences belong to different SCOP folds. The resulting set of sequence pairs contain 2528 true positives and 616 923 true negatives, resulting in a total of 618 821 pairs.

Benchmarking was performed by searching the benchmark set with each of the benchmark sequences in turn. Since SCANPS and PSI-BLAST use sequence profiles to enhance search sensitivity, it is necessary to embed the benchmark set in a larger sequence database in order to ensure that there is an adequate set of related sequences to construct the search profile for each benchmark sequence. Accordingly, the benchmark set was embedded in SWALL (17) to create the search database.

Both SCANPS version 2.3.9 and PSI-BLAST version 2.2-17 were run for a maximum of 10 iterations. Search parameters were chosen to reflect the typical use case where a low rate of false positives is acceptable in return for a high rate of true positives found. SCANPS was run with a profile *E*-value inclusion value of 0.015, a PIDT of 97% and using the BLOSUM50 scoring matrix. Gap penalties for opening and extending gaps were set to 12 and 2, respectively. These parameters have previously been established to provide the optimal combination of good sequence coverage and low error rate (15). To allow direct comparison with PSI-BLAST defaults, a further run of SCANPS was performed with BLOSUM62. PSI-BLAST scans were performed with the default BLOSUM62 scoring matrix and profile *E*-value cutoff of 0.002. The gap opening and extension penalties were set to 11 and 1, respectively. All other parameters were set to their default values. The *E*-values for the pairs in the benchmark were collected and ranked in order from lowest to highest and used to calculate coverage versus error-per-query plots.

Figure 1 shows plots of percentage true positive pairs versus percentage errors-per-query (EPQ) for SCANPS and PSI-BLAST. SCANPS offers significantly increased coverage versus PSI-BLAST for a given rate of EPQ. For example, at 1% EPQ, SCANPS using the BLOSUM 62 scoring matrix finds 27% of true positives, while

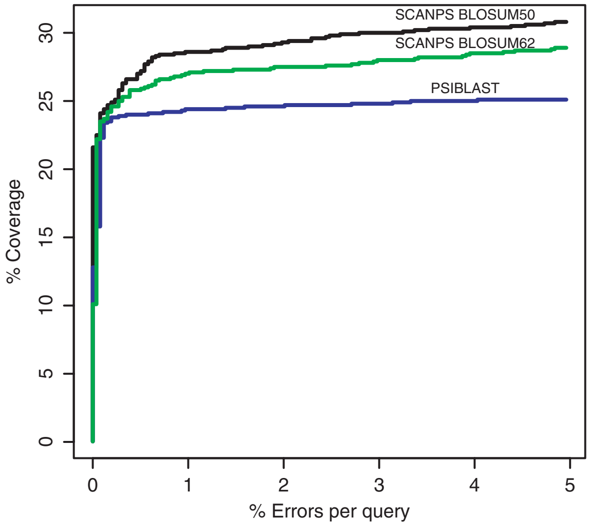


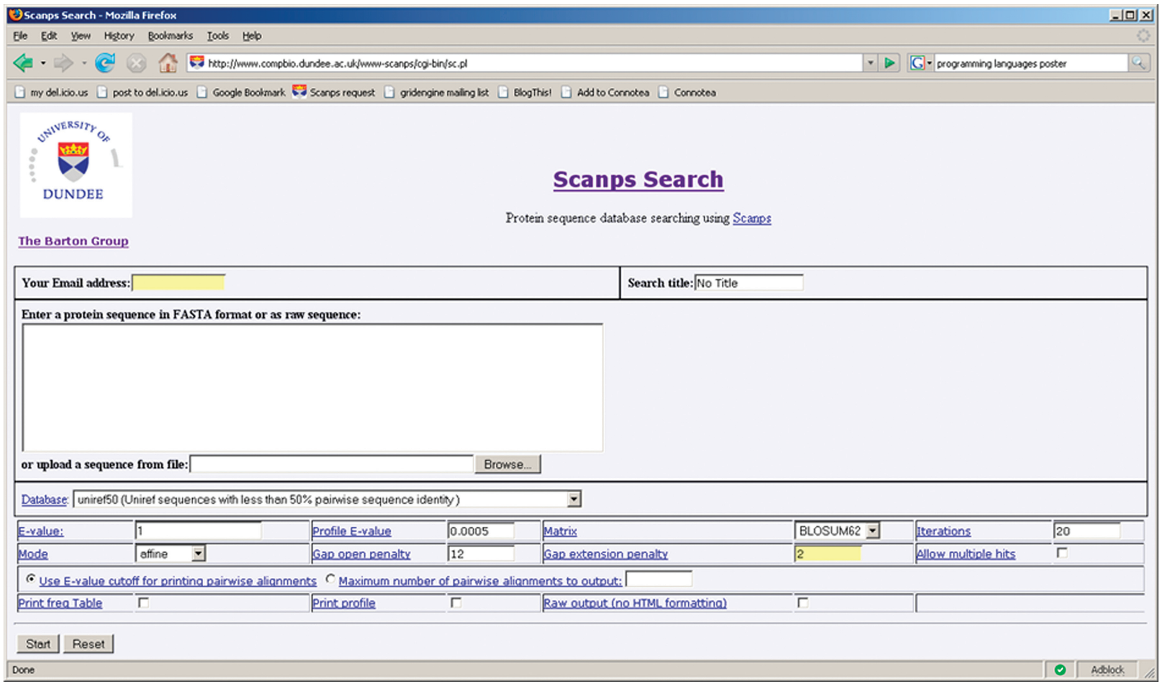
**Figure 1.** Coverage versus error plots for benchmarking of SCANPS and PSI-BLAST. The vertical axis (coverage) represents the number of true positives found divided by the total number of true positives in the benchmark, expressed as a percentage. True positives are those pairs in which the domains belong to the same SCOP superfamily. The horizontal axis (EPQ) represents the number of false positives found divided by the total number of true positives, expressed as a percentage. False positives are those pairs in which the domains belong to different SCOP folds. The black line show results for SCANPS run with the BLOSUM50 scoring matrix; profile inclusion *E*-value = 0.015, gap opening penalty = 12, gap extension penalty = 2 and profile identity threshold = 97%. The green line show results for SCANPS run with the BLOSUM62 scoring matrix; profile inclusion *E*-value = 0.015, gap opening penalty = 12, gap extension penalty = 2 and profile identity threshold = 97%. The blue line show results for PSI-BLAST run with the BLOSUM62 matrix; profile inclusion value = 0.002, gap opening penalty = 11 and gap extension penalty = 1. All other parameters for both methods were set to their default values. All runs were for a maximum of 10 iterations and pairs were collected from the final iteration of each search.

PSI-BLAST finds 24.4%. SCANPS when run with the optimal BLOSUM 50 scoring matrix finds 28.6% of true pairs at 1% EPQ.

## The SCANPS web interface

The output from SCANPS consists of tables of hits for each iteration, together with associated pairwise alignments and multiple alignments for all of the hits in a given iteration. This output is usually voluminous, so a major advantage of transforming the raw output into a browsable format is that it becomes much easier to navigate through the results. The web interface also permits crossreferencing to sequence databases and integration of the Jalview viewer (18) for viewing alignments and further analysis. The web interface allows the possibility of more sophisticated representations for SCOP search results as described below.

The server input page (Figure 2) allows the user either to upload a sequence file or paste a sequence directly into a text box. The search database is selected from a pulldown menu, with the current options of UniRef100, UniRef90, UniRef50 (19), PDB (20) and SCOP (2). The UniRef and PDB databases are updated automatically on biweekly and weekly schedules, respectively. The SCOP database is updated manually when a new version of SCOP is released. The SCOP sequences used are generated by ASTRAL (http://astral.berkeley.edu) (21) from the SEQRES records of the corresponding PDB entries and are non-redundant at the level of 100% sequence identity. The user can set all of the search parameters or use the defaults, which, with the exception of iteration number, are those found to be optimum in benchmarking on SCOP superfamily data. Each parameter is documented by a link



Scanps Search

Protein sequence database searching using Scanps

The Barton Group

Your Email address:

Search title: No Title

Enter a protein sequence in FASTA format or as raw sequence:

or upload a sequence from file:

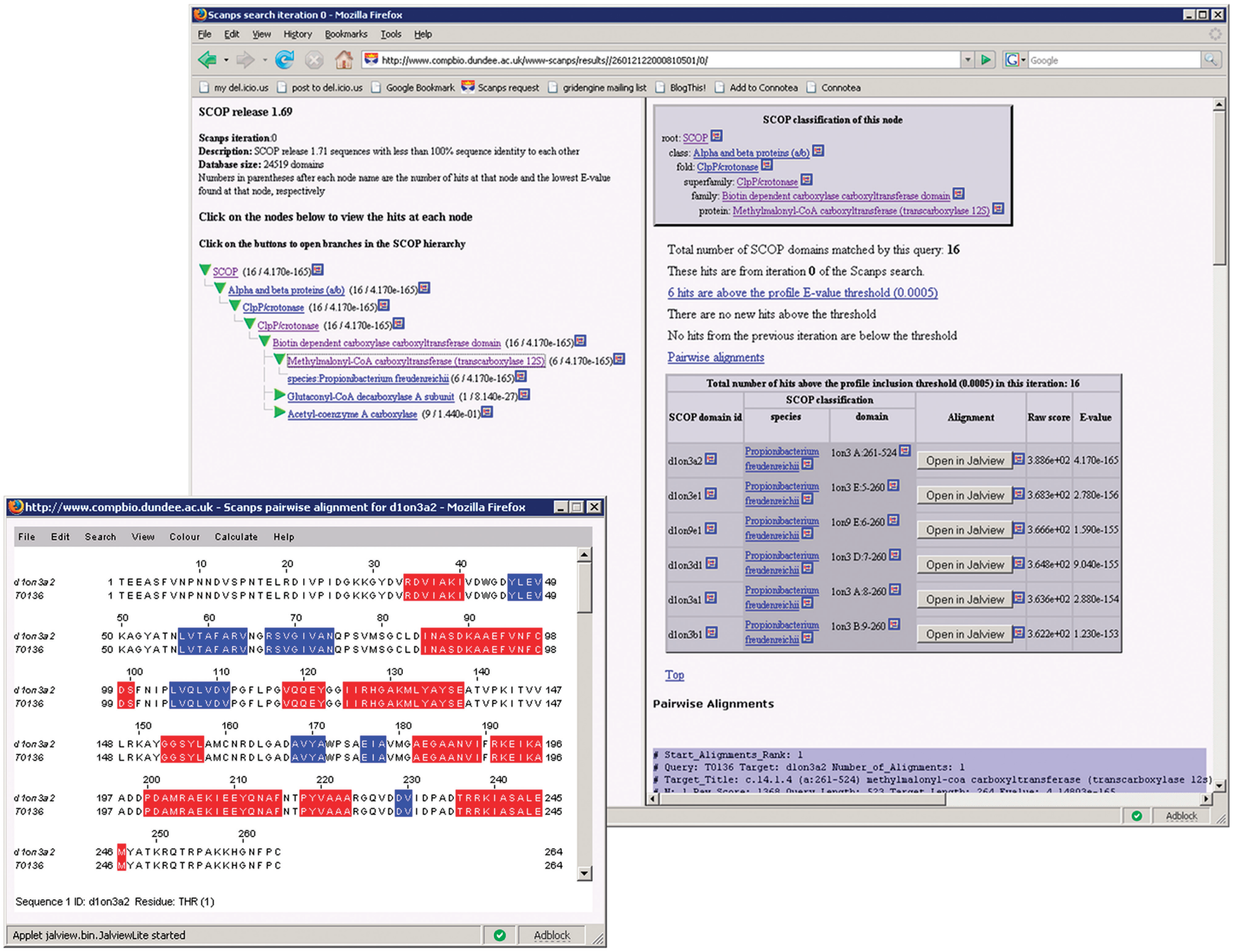
Database: uniref50 (Uniref sequences with less than 50% pairwise sequence identity)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| E-value: | 1 | Profile E-value | 0.0005 | Matrix | BLOSUM62 | Iterations | 20 |
| Mode | affine | Gap open penalty | 12 | Gap extension penalty | 2 | Allow multiple hits | □ |

● Use E-value cutoff for printing pairwise alignments ○ Maximum number of pairwise alignments to output:

| | | | | | |
|---|---|---|---|---|---|
| Print freq Table | □ | Print profile | □ | Raw output (no HTML formatting) | □ |

**Figure 2.** The SCANPS server input page. Search sequences can be pasted into the text box or uploaded from a file. Clicking on the name of each parameter displays the appropriate section of the documentation. The parameters displayed are the defaults used for searching.



Scanps search iteration 0 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.compbio.dundee.ac.uk/www-scanps/results//26012122000810501/0/
Google
my del.icio.us
post to del.icio.us
Google Bookmark
Scanps request
gridengine mailing list
BlogThis!
Add to Connotea
Connotea

**SCOP release 1.69**

**Scanps iteration**:0
**Description:** SCOP release 1.71 sequences with less than 100% sequence identity to each other
**Database size:** 24519 domains
Numbers in parentheses after each node name are the number of hits at that node and the lowest E-value found at that node, respectively

**Click on the nodes below to view the hits at each node**

**Click on the buttons to open branches in the SCOP hierarchy**

- SCOP (16 / 4.170e-165)
  - Alpha and beta proteins (a/b) (16 / 4.170e-165)
    - ClpP/crotonase (16 / 4.170e-165)
      - ClpP/crotonase (16 / 4.170e-165)
        - Biotin dependent carboxylase carboxyltransferase domain (16 / 4.170e-165)
          - Methylmalonyl-CoA carboxyltransferase (transcarboxylase 12S) (6 / 4.170e-165)
            - species:Propionibacterium freudenreichii (6 / 4.170e-165)
          - Glutaconyl-CoA decarboxylase A subunit (1 / 8.140e-27)
          - Acetyl-coenzyme A carboxylase (9 / 1.440e-01)

**SCOP classification of this node**

root: SCOP
class: Alpha and beta proteins (a/b)
fold: ClpP/crotonase
superfamily: ClpP/crotonase
family: Biotin dependent carboxylase carboxyltransferase domain
protein: Methylmalonyl-CoA carboxyltransferase (transcarboxylase 12S)

Total number of SCOP domains matched by this query: **16**

These hits are from iteration **0** of the Scanps search.

6 hits are above the profile E-value threshold (0.0005)

There are no new hits above the threshold

No hits from the previous iteration are below the threshold

Pairwise alignments

**Total number of hits above the profile inclusion threshold (0.0005) in this iteration: 16**

| SCOP domain id | SCOP classification | | Alignment | Raw score | E-value |
|---|---|---|---|---|---|
| | species | domain | | | |
| d1on3a2 | Propionibacterium freudenreichii | 1on3 A:261-524 | Open in Jalview | 3.886e+02 | 4.170e-165 |
| d1on3e1 | Propionibacterium freudenreichii | 1on3 E:5-260 | Open in Jalview | 3.683e+02 | 2.780e-156 |
| d1on9e1 | Propionibacterium freudenreichii | 1on9 E:6-260 | Open in Jalview | 3.666e+02 | 1.590e-155 |
| d1on3d1 | Propionibacterium freudenreichii | 1on3 D:7-260 | Open in Jalview | 3.648e+02 | 9.040e-155 |
| d1on3a1 | Propionibacterium freudenreichii | 1on3 A:8-260 | Open in Jalview | 3.636e+02 | 2.880e-154 |
| d1on3b1 | Propionibacterium freudenreichii | 1on3 B:9-260 | Open in Jalview | 3.622e+02 | 1.230e-153 |

Top

**Pairwise Alignments**

```
# Start_Alignments_Rank: 1
# Query: T0136 Target: d1on3a2 Number_of_Alignments: 1
# Target_Title: c.14.1.4 (a:261-524) methylmalonyl-coa carboxyltransferase (transcarboxylase 12s)
```

Adblock
http://www.compbio.dundee.ac.uk - Scanps pairwise alignment for d1on3a2 - Mozilla Firefox
File Edit Search View Colour Calculate Help

```
                10        20        30        40
d1on3a2   1 TEEASFVNPNNDVSPNTELRDIVPIDGKKGYDVRDVIAKIVDWGDYLEV 49
T0136     1 TEEASFVNPNNDVSPNTELRDIVPIDGKKGYDVRDVIAKIVDWGDYLEV 49

            50        60        70        80        90
d1on3a2  50 KAGYATNLVTAFARVNGRSVGIVANQPSVMSGCLDINASDKAAEFVNFC 98
T0136    50 KAGYATNLVTAFARVNGRSVGIVANQPSVMSGCLDINASDKAAEFVNFC 98

             100       110       120       130       140
d1on3a2  99 DSFNIPLVQLVDVPGFLPGVQQEYGGIIRHGAKMLYAYSEATVPKITVV 147
T0136    99 DSFNIPLVQLVDVPGFLPGVQQEYGGIIRHGAKMLYAYSEATVPKITVV 147

              150       160       170       180       190
d1on3a2 148 LRKAYGGSYLAMCNRDLGADAVYAWPSAEIAVMGAEGAANVIFRKEIKA 196
T0136   148 LRKAYGGSYLAMCNRDLGADAVYAWPSAEIAVMGAEGAANVIFRKEIKA 196

               200       210       220       230       240
d1on3a2 197 ADDPDAMRAEKIEEYQNAFNTPYVAAARGQVDDVIDPADTRRKIASALE 245
T0136   197 ADDPDAMRAEKIEEYQNAFNTPYVAAARGQVDDVIDPADTRRKIASALE 245

                250       260
d1on3a2 246 MYATKRQTRPAKKHGNFPC                               264
T0136   246 MYATKRQTRPAKKHGNFPC                               264
```

Sequence 1 ID: d1on3a2 Residue: THR (1)

Applet jalview.bin.JalviewLite started

**Figure 3.** Result of a SCANPS search of the SCOP database displayed in the tree-based SCOP results interface, with a pairwise alignment displayed in Jalview in the foreground. The tree frame on the left displays the SCOP hierarchy; the node frame on the right on displays the page for a particular node. The node's SCOP classification is listed at the top of the page. This is followed by summary data from SCANPS and then the domain table. For each domain, the table lists its SCOP classification below the current node, followed by the data for the domain returned by SCANPS. Pairwise alignments for each hit are listed below the domain table; these can also be displayed in Jalview by clicking the appropriate button. The alignment in Jalview is shaded to indicate the positions of strands (blue) and helices (red) in the database structure.

to the corresponding section of the help documentation. The search is run non-interactively and an Email message provides the user with the URL where the results can be viewed. The search time required depends on the inputs, parameters and the load on the server. Searches of SCOP and PDB are typically returned in within 5 min; UniRef searches may require several hours. The server does not at present use the MPI implementation of SCANPS but it is planned to do so to take full advantage of the large computing cluster that is available to the server.

The default output format is a linear presentation of the hits and corresponding pairwise and multiple alignments for each iteration. Hits for each iteration are listed in a table which displays the hit rank, the identifier assigned to the hit in the source database, a descriptive string, the *E*-value, the raw score and a button to display the pairwise alignment of the query and the hit in Jalview (18). Sequence identifiers are hyperlinked to the corresponding sequence entries on the database website. This is followed by a multiple alignment of the query with all of the hits found in the iteration and then the pairwise alignments. For each pairwise and multiple alignment, a button is provided to view the alignment in the Jalview alignment tool. Jalview provides access to a range of functions for editing and further analysis. When searching against the PDB and SCOP, Jalview uses secondary structure assignments from DSSP (22) to display the secondary structure elements in the database sequences. If a SCOP search has been performed, the results can also be viewed in a hierarchical viewer that maps the results onto the SCOP hierarchy (Figure 3). A separate mapping is generated for each search iteration and can be displayed by clicking a button in the linear interface. The 'tree frame' displays a tree-like representation of the SCOP hierarchy, the branches of which may be expanded or collapsed to display particular branches at more fine-grained levels of classification. The nodes in the tree frame are annotated with the number of SCANPS hits at that node and the lowest *E*-value found for a hit at that node, allowing the viewer to quickly identify those parts of the SCOP hierarchy in which hits are clustered. Clicking on a SCOP node in the tree frame displays the corresponding node page in the node frame. The 'node frame' displays, for a given SCOP node, a page (the 'node page'), which contains a table of the domains in the node and summary information about the node. For each domain, the table lists the domain's SCOP classification and the data returned by the search method.

## IMPLEMENTATION

The server is implemented as a set of Perl CGI scripts but most of the functionality is contained in a set of common Perl modules used by all the scripts. The SCOP interface is built using an object-oriented Perl library that is designed to facilitate building interfaces for any program that searches SCOP.

## CONCLUSION

The SCANPS web server (http://www.compbio.dundee.ac.uk/www-scanps) provides access to SCANPS in the form of a user-friendly web interface with regularly updated databases and postprocessing of results to present them in a form that facilitates analysis and interpretation. The server provides the facility to search the complete SCOP database with a query sequence and display the results in a tree view. This maps hits directly onto the SCOP hierarchy with links to the SCOP database website. It also integrates the Jalview alignment editor for viewing alignments between the hits and the query sequence. Planned enhancements to the server include deploying the MPI implementation of SCANPS to reduce search times further, as well as allowing search of SCOP embedded in UniRef to improve sensitivity for iterative searches.

## ACKNOWLEDGEMENTS

Since 1989, SCANPS development has been funded from a variety of sources including the Royal Society (G.J.B.), European Molecular Biology Laboratory (C.W., S.S. and G.J.B.) and the Scottish Funding Council (Scottish Bioinformatics Research Network grant) (T.P.W. and G.J.B.). We thank all our colleagues and SCANPS users for their help and support throughout the development of SCANPS and the new web server interface described here. Funding to pay the Open Access publication charges for this article was provided by the Scottish Funding Council.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., 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.
2. 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.
3. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
4. Barton,G.J. (1998) Protein sequence alignment techniques. *Acta Crystallogr. D*, **54**, 1139–1146.
5. Barton,G.J. (1992) Computer speed and sequence comparison. *Science*, **257**, 1609.
6. Barton,G.J. (1991) Scanning protein sequence databanks using a distributed processing workstation network. *Comput. Appl. Biosci.*, **7**, 85–88.
7. Dagum,L. and Menon,R. (1998) OpenMP: an industry-standard API for shared memory programming. *IEEE Comput. Sci. Eng.*, **5**, 46–55.
8. MPI Forum (1994) MPI: A message-passing interface standard. *Int. J. Supercomput. Appl.*, **8**, 165–416.
9. Barton,G.J., Webber,C. and Searle,S.M.J. (2000) New developments to SCANPS: high performance parallel iterated protein sequence searching with full dynamic programming and on-the-fly statistics. *Presented at the 8th International Conference on Intelligent Systems for Molecular Biology*, 19–23, August 2000, La Jolla, California. Available online at: http://www.iscb.org/ismb2000/9_8_pdfs/TuesSA/Barton.pdf.
10. Barton,G.J. and Sternberg,M.J.E. (1990) Flexible protein sequence patterns, a sensitive method to detect weak structural similarities. *J. Mol. Biol.*, **212**, 389–402.
11. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile scanning for three-dimensional structural patterns in protein sequences. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
12. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
13. Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, pp. 345–352.
14. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
15. Webber,C. (2003) Protein sequence database searching. *Ph.D. Thesis*. EMBL-European Bioinformatics Institute, University of Cambridge, Cambridge, UK.
16. Brenner,S., Chothia,C. and Hubbard,T. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–7068.
17. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
18. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
19. Suzek,B.E., Huang,H., McGarvey,P., Mazumder,R. and Wu,C.H. (2007) UniRef: comprehensive and non-redundant UniProt Reference Clusters. *Bioinformatics*, **23**, 1282–1288.
20. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
21. Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Res.*, **28**, 254–256.
22. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.