W10–W14 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl330

FISH—family identification of sequence homologues using structure anchored hidden Markov models

Jeanette Tångrot[1,2], Lixiao Wang[1], Bo Kågström[2,3] and Uwe H. Sauer[1,*]

[1]Umeå Center for Molecular Pathogenesis, UCMP, [2]Department of Computing Science and [3]High Performance Computing Center North, HPC2N, Umeå University, Umeå, Sweden

Received February 14, 2006; Revised March 1, 2006; Accepted April 14, 2006

ABSTRACT

The FISH server is highly accurate in identifying the family membership of domains in a query protein sequence, even in the case of very low sequence identities to known homologues. A performance test using SCOP sequences and an *E*-value cut-off of 0.1 showed that 99.3% of the top hits are to the correct family saHMM. Matches to a query sequence provide the user not only with an annotation of the identified domains and hence a hint to their function, but also with probable 2D and 3D structures, as well as with pairwise and multiple sequence alignments to homologues with low sequence identity. In addition, the FISH server allows users to upload and search their own protein sequence collection or to quarry public protein sequence data bases with individual saHMMs. The FISH server can be accessed at http://babel.ucmp.umu.se/fish/.

INTRODUCTION

The detection of homologous proteins with known function and well-determined three-dimensional (3D) structures is crucial for the correct characterization and annotation of newly sequenced proteins. Since proteins are modular and can harbour many domains, it is advisable to characterize the constituent domains rather than the protein as a whole. Existing internet resources, such as Pfam (1), Superfamily (2), SMART (3), CD search (4) and others, provide the user with versatile tools for domain identification. Nevertheless, the definition field of millions of database entries still contains remarks such as 'hypothetical', 'putative', 'unidentified' or 'function unknown'.

The FISH server can be used as a complement to existing annotation methods. One can compare a query sequence with all structure anchored hidden Markov models (saHMMs) and, in case of a match, assign family membership on the domain level for such sequences even in the case of low sequence identity.

Furthermore, it is important to discover those proteins in a database that harbour a certain domain, independent of sequence identity and annotation status. The FISH server provides such a tool, where a user can employ individual saHMMs for searching against a sequence database and obtain hits even if the sequence identity is 20% or less and falls below the so called 'twilight zone' curve, *pI* (5).

METHOD

Construction of structure anchored hidden Markov models

FISH, which stands for Family Identification with Structure anchored HMMs, is a server for the identification of sequence homologues on the basis of protein domains. At the heart of the server lies a collection of 982 saHMMs, each representing one SCOP (6) domain family (Tångrot, J., Kågström, B. and Sauer, U.H., manuscript in preparation). The saHMMs are built with HMMER 2.2g (7) from structure anchored multiple sequence alignments, saMSAs. The saMSAs are derived from multiple structure superimpositions of representative homologous domains. In order to maximize the sequence variability within each domain family, we superimposed only those domains whose mutual sequence identity falls below the 'twilight zone' curve, *pI* (5). The selected domains are hereafter called the saHMM-members. Their coordinate files were obtained from the SCOP version 1.69 associated ASTRAL compendium (8) and were superimposed with STAMP (9). Only high-quality X-ray crystal structures were used. Since at least two structures are needed for superimposition and because of the stringent sequence identity restrictions, our collection of saHMMs currently covers ~35% of SCOP families belonging to true classes. We expect this number to increase due to the exponential rate at which 3D structures become available.

*To whom correspondence should be addressed. Tel: +46 90 785 6784; Fax: +46 90 77 80 07; Email: uwe@ucmp.umu.se

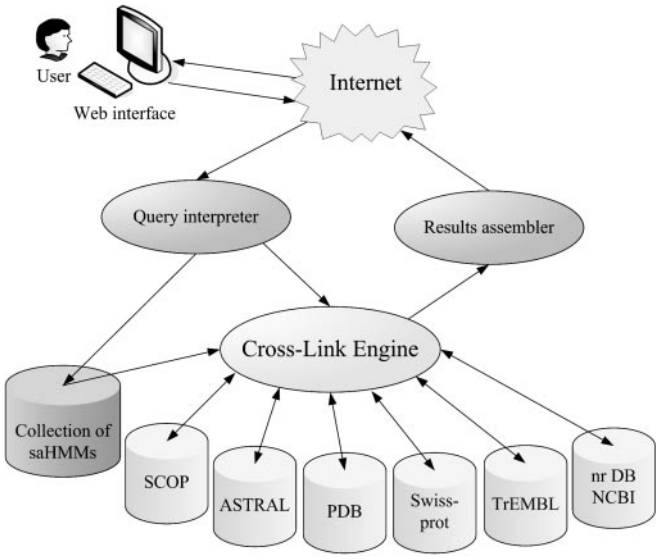


Figure 1. Schematic layout of the FISH server architecture. The user initializes a query via the web interface. The query is processed by the query interpreter, using the collection of saHMMs. The cross-link engine integrates information from the associated data bases [SCOP, ASTRAL, PDB, nr (NCBI), Swiss-Prot and TrEMBL] with the results of the query. The results assembler compiles the search results and presents them to the user via the web interface.

Brief description of the FISH server

The architecture of the FISH server is displayed in Figure 1. Flat file databases were imported into a relational data base (MySQL) and cross-linked. The MySQL database is implemented on a Linux platform. The user interface is written in Perl, PHP and JavaScript, and integrated with the Apache web server.

The user inputs a query via the web interface. The query interpreter processes the input, using the collection of saHMMs. The cross-link engine merges information from the associated databases with the results of the query. The results assembler presents the outcome of the search to the user via the web interface. The search results can be sent to the user by e-mail in the form of a www-link and are stored on the server for 24 h.

USE OF THE FISH SERVER

The organization of the FISH server input and results pages is schematically outlined in Figure 2 and described in the following.

Sequence vs. saHMM search

Using the FISH server for a sequence vs. saHMMs search is straightforward. The user is required to enter an amino acid sequence in FASTA or text format, or to upload a sequence file. The *E*-value cut-off is adjustable and determines the level of significance of the reported hits.

The FISH search results are presented in a hierarchical manner (see Figure 2). At the top of the results hierarchy is the 'overview of results' page (see Figure 3). It contains a table of all matches, sorted by ascending *E*-values up to the selected *E*-value cut-off. The lengths of the schematic arrows below the table correspond to the query sequence length. For each found domain, the position of the matching sequence interval is schematically marked by a coloured box. By following the links on the overview page the user obtains increasingly detailed information about each match.

In the table displayed in the 'overview of results' window, each saHMM identifier links to the SCOP lineage of that domain family as well as to a table listing the saHMM-members (Figure 2, left hand side, and Figure 4). Each entry in the saHMM-member field links to a saHMM-based pairwise sequence alignment of the query with that member and further to links providing coordinate information.

The chain identifier field links to a page with the sequence of the ASTRAL domain, followed by the sequence contained in the protein data bank file with the ASTRAL sequence interval marked in orange. This page also provides a link to the corresponding NCBI sequence entry.

The Coordinate icon in the table leads the user to an interactive Java window running Jmol version 10.00 (http://www.jmol.org) where the domain structure of the saHMM-member can be visualized. The user can rotate the structure and analyze it by zooming in on details or by applying a variety of colouring schemes and display options.

The coloured boxes on the sequence arrows in the 'overview of results' window lead the user to alignments of the query sequence with the saHMM consensus sequence. Links on this page lead the user to a sequence alignment of the query sequence with the saMSA used to build the saHMM (right hand side of Figure 2). The multiple sequence alignment can be viewed in different formats such as Stockholm, MSF and A2M.

It is also possible to view all pairwise sequence alignments of the query sequence with the individual saHMM-members. All alignments are anchored on the saHMM.

Using the SCOP sequences to test the performance of the server we found that in 99.3% of the cases the top hit matches the correct saHMM, choosing an *E*-value cut-off of 0.1. The matches obtained in a sequence vs. saHMM search provide the user with a classification on the SCOP family level and outline structurally defined, putative domain boundaries in the query sequence. This information can be used for sequence annotation, to design mutation sites, to identify soluble domains, to find structural templates for homology modelling and possibly for structure determination by molecular replacement.

Performance test on new sequences

In the following we assess the ability of the saHMMs to assign the correct domain family membership to newly sequenced proteins. For this purpose we used the 24 957 domain sequences that are contained in SCOP 1.69 (released July 2005) but not in SCOP 1.61 (released Nov. 2002), to quarry the collection of 682 saHMMs based on SCOP 1.61. Here and in the following two paragraphs we consider a hit only if it is the top match with an *E*-value equal to or better than 0.1.

Using the classification of SCOP 1.69 we find that 14 173 of the query sequences (57%) belong to domain families for which we have a saHMM based on SCOP 1.61. Ideally, all of these sequences should find a match to the correct family saHMM.

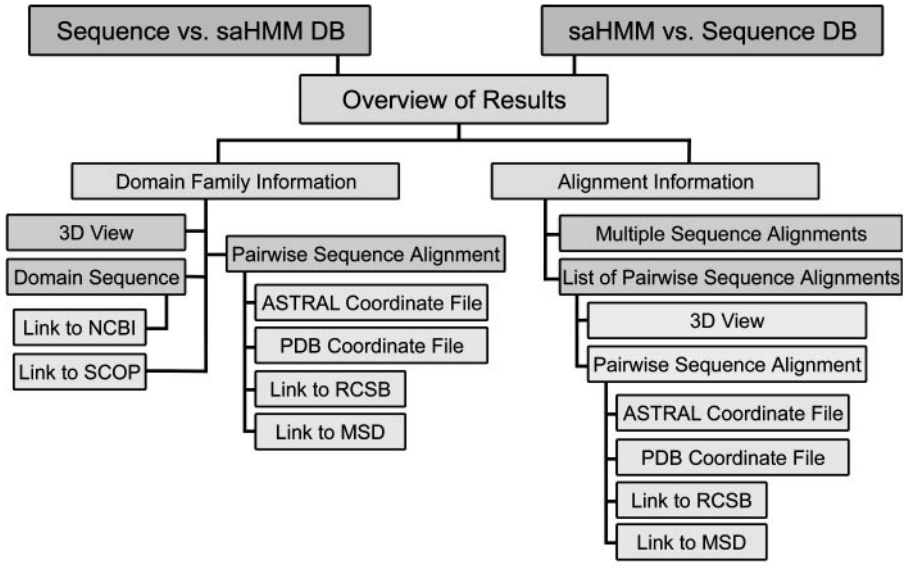


Figure 2. Organization of the FISH server input and result pages. The result pages are similar for a search of a query sequence versus the collection of saHMMs and for a search with a saHMM versus a sequence database. The information available can be roughly divided into domain family information (left branch) and alignment information (right branch). The domain family information includes SCOP classification, the sequences and 3D structures of the saHMM-members, and pairwise sequence alignments of the query to each member. The alignment information provides multiple and pairwise alignments of the query sequence to the consensus sequence extracted from the saHMM and the sequences used to build the saHMM. All alignments are anchored on the saHMM. Links to relevant data bases are provided.

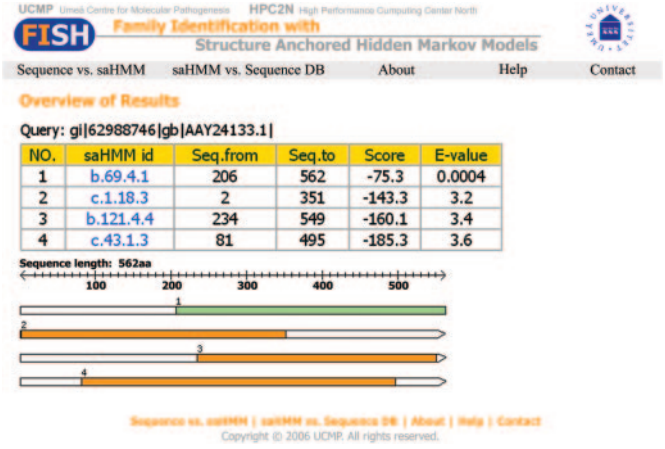


Figure 3. Overview of results page. This page contains a table of all matches, and a graphical representation of the matches mapped onto a sequence arrow. The position of the matching sequence interval is marked by a colour coded box. Green corresponds to *E*-values <0.1, yellow to an *E*-value interval between 0.1 and 1.0 and orange to an *E*-value >1.0. By following the links on the overview page the user obtains more detailed information about each match, such as the SCOP lineage, pairwise and multiple sequence alignments, and 3D structures of domain members. Shown is a search carried out with AAY24133.1, a human protein labelled 'unknown'.

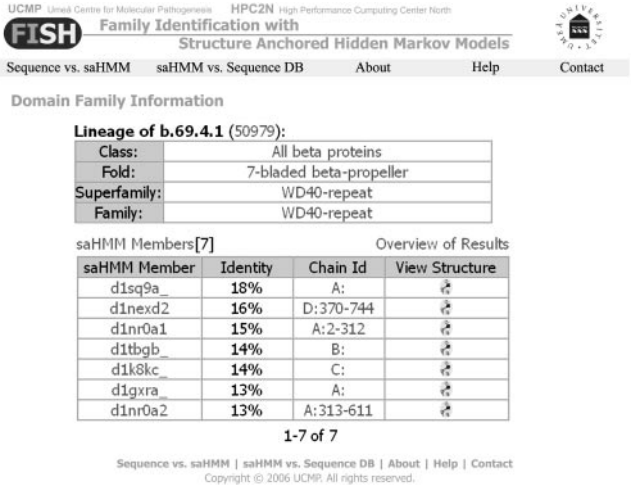


Figure 4. Domain family information page. The SCOP lineage of the domain family is shown, as well as a table listing the saHMM-members. Each saHMM-member links to a pairwise sequence alignment of the query with the member, anchored on the saHMM and to links with coordinate information. The chain entry shows the sequence of the saHMM-member. The domain structures of the saHMM-members can be visualized interactively by following the link under view structure.

Our results show, that 10 513 sequences (74%) are able to identify the correct saHMM as their top hit. This number increases to 10 737 sequences (76%) if we accept matches on the superfamily level as well. Of the 10 784 domain sequences for which we do not have a saHMM (as of version 1.61), 183 sequences (2%) found a match to a saHMM within the correct superfamily. No hit was obtained for 10 561 sequences (98%), which demonstrates that our saHMMs are very domain family specific.

The combined searches resulted in a total of 11 202 hits of which 10 513, i.e. 94% of all matches, were to the correct family saHMM. An additional 407 hits (4%) were correct on the superfamily level.

Comparing saHMMs with Pfam HMMs

To compare the performance of the FISH server with Pfam, we used saHMMs based on SCOP 1.61 and the corresponding Pfam_ls HMM release (version 7.8, released November

2002). Since the definition of a SCOP family differs from the Pfam definition, the relationships between SCOP and Pfam families were determined by finding the SCOP classification of PDB sequences that are part of the Pfam-A alignments. Of the 24 957 sequences new in SCOP 1.69 compared with version 1.61, a total of 11 592 sequences belong to families with both an HMM in Pfam and a saHMM, and are used as query sequences. In the following we consider only top hits with an *E*-value <10 as matches.

The correct family relationships were detected for 9574 of the sequences (83%) using the saHMMs and for 10 128 sequences (87%) using Pfam. It is of interest to note that 812 of the sequences with hits to the correct saHMM did not find the correct HMM in Pfam.

Detecting remote sequence homologues

We further selected, for each domain family, those sequences in the set of 11 592 query sequences that had a sequence identity below the 'twilight zone' curve compared with the saHMM-members based on SCOP 1.61. This left us with 3247 new low identity sequences, of which 2014 sequences (62%) obtained hits to the correct family saHMMs even though the sequence identity to the saHMM-members is very low. Interestingly, 79 of these relationships were not detected by Pfam, despite the possibility that some of the query sequences could have a sequence identity above *pI* to Pfam-A seed sequences.

saHMM searched vs. sequence database

By choosing a saHMM that represents a particular SCOP domain family to search a sequence database, one can identify members of that domain family within protein sequences. In this way it is possible to identify previously un-annotated sequences on the domain family level, even in case of very low sequence identities.

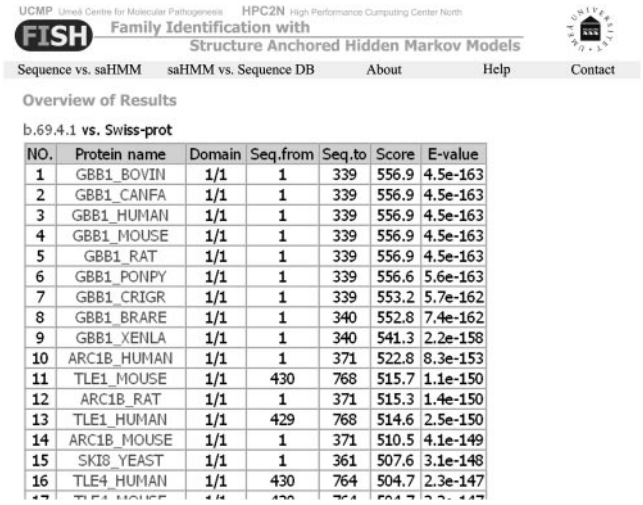
The input page of the saHMM vs. sequence database search is divided into two parts. To the left is a section with several options for selecting a saHMM to use for the search, and to the right is the actual input section.

There are several ways of choosing the saHMM to search with. If one knows which SCOP domain family to use, and how to find it in the SCOP classification, the saHMM can easily be located by browsing the classification tree. Otherwise, the saHMM can be located using the free text search option. All SCOP domain families whose description matches the text search are listed. Those with a saHMM can be selected for searching.

Alternatively, the name of the saHMM can be written directly in the input field on the right. The user can also select which sequence database to search against and input an appropriate cut-off for the *E*-value.

The results are reported in the form of a table (see Figure 5), where the matches are sorted by *E*-value with the best hit listed first. Above the results table, the user can follow a link to information about the domain family as well as sequence and structural information about the domains used to build the saHMM.

Each protein name in the results table is linked to the corresponding sequence entry, in which the matching sequence interval is marked in orange. An alignment of the matching sequence to the saHMM consensus is shown below the sequence, with the option to view both multiple and pairwise alignments anchored on the saHMM. In the pairwise alignments view, the sequence identity of the found match to each saHMM-member is displayed in a table. From there, links allow the user to view the structure of the members and to obtain coordinate information.


UCMP Umeå Centre for Molecular Pathogenesis HPC2N High Performance Computing Center North

FISH Family Identification with Structure Anchored Hidden Markov Models

Sequence vs. saHMM | saHMM vs. Sequence DB | About | Help | Contact

Overview of Results

b.69.4.1 vs. Swiss-prot

NO.	Protein name	Domain	Seq.from	Seq.to	Score	E-value
1	GBB1_BOVIN	1/1	1	339	556.9	4.5e-163
2	GBB1_CANFA	1/1	1	339	556.9	4.5e-163
3	GBB1_HUMAN	1/1	1	339	556.9	4.5e-163
4	GBB1_MOUSE	1/1	1	339	556.9	4.5e-163
5	GBB1_RAT	1/1	1	339	556.9	4.5e-163
6	GBB1_PONPY	1/1	1	339	556.6	5.6e-163
7	GBB1_CRIGR	1/1	1	339	553.2	5.7e-162
8	GBB1_BRARE	1/1	1	340	552.8	7.4e-162
9	GBB1_XENLA	1/1	1	340	541.3	2.2e-158
10	ARC1B_HUMAN	1/1	1	371	522.8	8.3e-153
11	TLE1_MOUSE	1/1	430	768	515.7	1.1e-150
12	ARC1B_RAT	1/1	1	371	515.3	1.4e-150
13	TLE1_HUMAN	1/1	429	768	514.6	2.5e-150
14	ARC1B_MOUSE	1/1	1	371	510.5	4.1e-149
15	SKI8_YEAST	1/1	1	361	507.6	3.1e-148
16	TLE4_HUMAN	1/1	430	764	504.7	2.3e-147

Figure 5. saHMM vs. sequence database search. The results for the search with the saHMM representing the SCOP family b.69.4.1 (50979) are reported in the form of a table listing the matches sorted by *E*-value. Only part of the table is shown in the figure. Above the results table is a link to information about the domain family as well as sequence and structural information about the domains used to build the saHMM. Each protein name contains a link to the corresponding sequence entry, an alignment of the matching sequence to the saHMM consensus and the option to view both multiple and pairwise alignments anchored on the saHMM.

A search with a saHMM vs. SwissProt can take anything from 15 min up to ~9 h. Searching TrEMBL, which is about 10 times larger, takes considerably longer. In order to minimize the waiting time for the user, we pre-calculated the searches of all 982 saHMMs vs. SwissProt, TrEMBL and the NCBI non-redundant database, nr, using an *E*-value cut-off of 100. Depending on the *E*-value choice of the user, the results are extracted and presented up to that value.

In addition, users can choose to upload and search their own protein sequence databases.

SUMMARY

The FISH server is a versatile tool with a dual function. On the one hand, the user can perform sensitive sequence searches versus a collection of saHMMs, which can provide matches even within the 'midnight zone' of sequence alignments. On the other hand, the user can choose one of the saHMMs to perform a search against a protein sequence data base. Since the saHMMs are based on structure anchored multiple sequence alignments, the alignment of the query to the saHMM-members can be used to draw conclusions about the probable secondary and tertiary structure of the query sequence.

A comparison of FISH saHMMs with Pfam HMMs shows that the methods are comparable in their ability to

assign family memberships. Our findings also show that each collection of HMMs can assign family memberships to sequences that are missed by the other, thus complementing each other.

Further we demonstrate that for sequences with very low sequence identity to the saHMM-members a correct assignment was made for about 62% of the sequences. This demonstrate the ability to detect remote homologues on the domain family level.

ACKNOWLEDGEMENTS

The authors thank our colleagues at UCMP for critically scrutinizing the FISH server and for suggestions for improvements. Part of this research was conducted using the resources of the High Performance Computing Center North (HPC2N). U.H.S. acknowledges the partial support for this project by a grant from the Knowledge Foundation (KK-Stiftelsen) B.K. acknowledges the partial support for this project by the Swedish Foundation for Strategic Research grant A3.02:13. Funding to pay the Open Access publication charges for this article was provided by the Knowledge Foundation (KK-Stiftelsen).

Conflict of interest statement. 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., Studholme,D.J., Yeats,C. and Eddy,S.R. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.
2. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res*., **32**, D235–D239.
3. 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.
4. Marchler-Bauer,A. and Bryant,S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res*., **32**, W327–W331.
5. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng*., **12**, 85–94.
6. 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.
7. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
8. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res*., **32**, D189–D192.
9. Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.