# FISHing for Remote Sequence Homologues in the Midnight Zone

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

[1] Umeå Centre for Molecular Pathogenesis
[2] Department of Computing Science
[3] High Performance Computing Center North
Umeå University, S-901 87 Umeå, Sweden

**Abstract.** The FISH (**F**amily **I**dentification with **S**tructure anchored **H**idden Markov models, saHMMs) server is highly accurate in identifying family memberships of domains in a query protein sequence, even if the sequence identities to other family members are very low. Matches 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 remote homologues. A user can also search protein sequence data bases with individual saHMMs, in order to find protein sequences that harbour remote domain homologues. The core of the server is a collection of saHMMs, which are based on multiple structure superimpositions of remote homologues from which multiple sequence alignments were extracted. The FISH server can be accessed at http://soul.ucmp.umu.se/fish/.

## 1 The FISH server

For the correct characterization and annotation of newly sequenced proteins, the detection of homologous proteins with known functions and well determined three dimensional (3D) structures is crucial. 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 data base entries still contains remarks such as "hypothetical", "unidentified" or "function unknown".

FISH, Family Identification with Structure anchored HMMs, is a server for the identification of remote sequence homologues, on the basis of protein domains. The FISH server can be used as a complement to existing annotation methods.

## 2 Construction of saHMMs

At the heart of the FISH server lies a collection of 982 structure anchored hidden Markov models, saHMMs, each representing one SCOP [5], version 1.69,

domain family (manuscript in preparation). The saHMMs are built with HMMER 2.2g [6] from structure anchored multiple sequence alignments, saMSAs. The saMSAs are derived from multiple structural superimpositions of representative homologous domains. In order to maximise the sequence variability within each domain family, we superimposed only those domains which have a mutual sequence identity below the "twilight zone" curve, $pI$ [7], were determined to a resolution below 3.6 and match our requirements for good quality crystal structures. The selected domains are hereafter called the saHMM-members. Their coordinate files were obtained from the ASTRAL compendium [8] and were superimposed with STAMP [9].

Since at least two structures are needed for superimposition and because of the stringent sequence identity restrictions, our collection of saHMMs currently covers about 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.

## 3 Use of the FISH server

### 3.1 Brief description of the FISH server

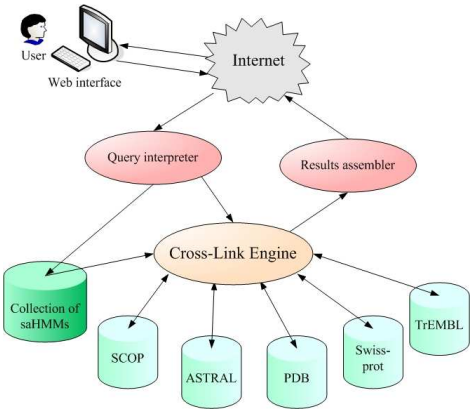


**Fig. 1.** Architecture of the FISH server.

Flat file data bases were imported into a relational data base (MySQL implemented on a Linux platform) and cross-linked. 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 analyzes the input, using the collection of saHMMs. The cross-link engine merges information from the associated data bases 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 hours.

### 3.2 Sequence vs. saHMM search

Using the FISH server, one can compare a query sequence with all saHMMs. Matches obtained in such a search provide the user with a classification on the SCOP family level and outline structurally defined, putative domain boundaries in the query sequence. For each match, the user is provided with the SCOP lineage of the matching saHMM, as well as pairwise and multiple sequence alignments of the query sequence to the saHMM-members, anchored on the saHMM. Individual saHMM-members can be investigated in more detail, for example by viewing their 3D structures in an interactive Java window. In our tests, the correct saHMM was identified for 88% of the sequences, with an accuracy of 96%.

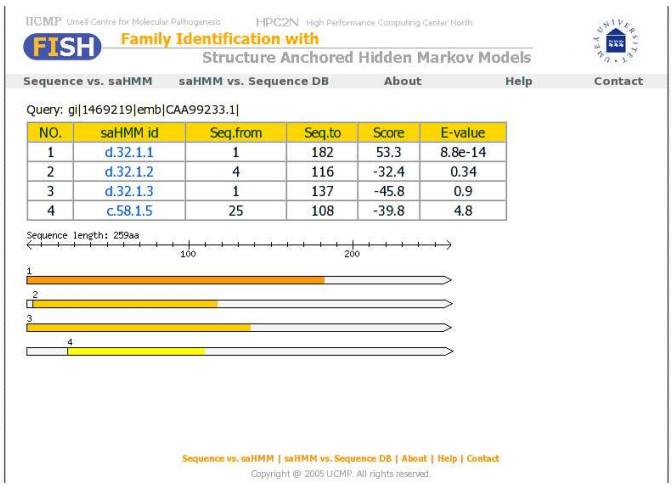


**Fig. 2.** Overview of results from a search with a query sequence versus the saHMMs.

### 3.3 saHMM searched vs. sequence database

The FISH server also allows the user to employ individual saHMMs for searching against a sequence data base (currently SwissProt and TrEMBL are available), to discover those proteins that harbour a certain domain, independent of sequence identity and annotation status. 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, below 20%. For each match, the user obtains the corresponding sequence entry, as well as pairwise and multiple sequence alignments

of the match and the saHMM-members, anchored on the saHMM. Information about the domain family used for searching is also easily available.

### 3.4 saHMM computations

In order to determine pairwise sequence identities for the selection of saHMM-members, we construct structural superimpositions for all pairs of members from each of the 2845 SCOP families. To speed up the calculations, much of the work is done in parallell, by simply running several families concurrently. Several perl programs are written in order to automate the process from raw SCOP family classification of domains to creating and testing the saHMMs, including check of consistency between the SCOP classification and the coordinate files, parsing of results to convert output from one program to input for another, etc.

A search with an saHMM vs. SwissProt can take anything from 15 minutes up to about nine hours. Searching TrEMBL, which is about ten 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 and TrEMBL using an E-value cut-off of 100. The calculations are done in parallell, using the cluster resources of High Performance Computing Center North (HPC2N). Depending on the E-value choice of the user, the results are extracted and presented up to that value.

## 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.L., Studholme, D.J., Yeats, C., Eddy, S.R.: The Pfam protein families database. Nucleic Acids Research **32** (2004) D138–D141
2. Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C., Gough, J.: The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Research **32** (2004) D235–D239
3. Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., Bork, P.: SMART 5: domains in the context of genomes and networks. Nucleic Acids Research **34** (2006) D257–D260
4. Marchler-Bauer, A., Bryant, S.H.: CD-Search: protein domain annotations on the fly. Nucleic Acids Research **32** (2004) W327–W331
5. Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology **247** (1995) 536–540
6. Eddy, S. R.: Profile Hidden Markov Models. Bioinformatics **14** (1998) 755–763
7. Rost, B.: Twilight zone of protein sequence alignments. Protein Engineering **12** (1999) 85–94
8. Chandonia, J.-M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., Brenner, S.E.: The ASTRAL Compendium in 2004. Nucleic Acids Research **32** (2004) D189–D192
9. Russell, R. B., Barton, G. J.: Multiple Protein Sequence Alignment From Tertiary Structure Comparison: Assignment of Global and Residue Confidence Levels. PROTEINS: Structure, Function and Genetics **14** (1992) 309–323