Design, Construction and Use of the FISH Server

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 (HPC2N),
Umeå University, SE-901 87 Umeå, Sweden
{jeanette, bokg}@cs.umu.se, {lixiao, uwe}@ucmp.umu.se

Abstract. At the core of the FISH (**F**amily **I**dentification with **S**tructure anchored **H**idden Markov models, saHMMs) server lies the midnight ASTRAL set. It is a collection of protein domains with low mutual sequence identity within homologous families, according to the structural classification of proteins, SCOP. Here, we evaluate two algorithms for creating the midnight ASTRAL set. The algorithm that limits the number of structural comparisons is about an order of magnitude faster than the all-against-all algorithm. We therefore choose the faster algorithm, although it produces slightly fewer domains in the set. We use the midnight ASTRAL set to construct the structure-anchored Hidden Markov Model data base, saHMM-db, where each saHMM represents one family. Sequence searches using saHMMs provide information about protein function, domain organization, the probable 2D and 3D structure, and can lead to the discovery of homologous domains in remotely related sequences.

The FISH server is accessible at `http://babel.ucmp.umu.se/fish/`.

1 Introduction

Genome sequencing projects contribute to an exponential increase of available DNA and protein sequences in data bases. Millions of sequence entries contain remarks such as "hypothetical", "unidentified", or "unknown". It is therefore crucial to develop accurate automated sequence annotation methods. For proper characterization of newly sequenced proteins it is important to associate them with homologous proteins of well characterized functions and possibly high quality three dimensional (3D) structures. Proteins are modular and can harbour many domains. Consequently, it is advisable to characterize the constituent domains rather than the protein as a whole. Existing resources, such as Pfam [1], Superfamily [7], SMART [6] and others, provide the user with versatile tools for domain identification. Common for these methods is that they use protein sequence alignments that include as many sequences as possible, even with high sequence identity of up to 95%, to construct hidden Markov models, HMMs. At the core of our approach lies a data base of structure-anchored hidden Markov

B. Kågström et al. (Eds.): PARA 2006, LNCS 4699, pp. 647–657, 2007.

models, saHMMs. In contrast to the other methods, we derive structure anchored multiple sequence alignments, saMSAs, exclusively from multiple structure superimpositions of protein domains within SCOP families [9]. Only spatial distance criteria are considered to find matching residues and to deduce the multiple sequence alignments from which the saHMMs are built. Great care is taken to ensure sequence diversity among the domains by including only such members with a mutual sequence identity below a certain cut-off value. We call the data set containing the low mutual sequence identity domains the "midnight ASTRAL set", since it was derived using the ASTRAL compendium [2]. We have made the saHMM data base, saHMM-db, publicly available through the FISH server, which has been introduced and briefly described earlier [13]. FISH, which stands for Family Identification with Structure-anchored HMMs, is a versatile server for the identification of domains in protein sequences. Here, we describe the algorithms behind the server in more detail, in particular the creation of the midnight ASTRAL set. In addition, we present a layout of the cross-linking of the underlying data bases and describe in more detail how to use the server.

2 The Midnight ASTRAL Set and Selection Algorithms

The midnight ASTRAL set is the non-redundant collection of representative domains used to construct the saHMMs. In order to maximize the sequence variability within each SCOP domain family [9], we included only domains with low mutual sequence identities, below the "twilight zone" curve, $p^I(L,0)$ [10],[8]:

$$p^I(L,n) = n + \begin{cases} 100 & for \;\; L \leq 11, \\ 480 \cdot L^{-0.32 \cdot \left(1+e^{-L/1000}\right)} & for \;\; 11 < L \leq 450, \\ 19.5 & for \;\; L > 450. \end{cases} \qquad (1)$$

The function $p^I(L,0)$ defines the limit of percent sequence identity for clearly homologous protein sequences, as a function of the alignment length L.

To construct the midnight ASTRAL set, representative domains must be selected for each of the 2845 SCOP families belonging to true classes. Individual families can harbour as few as one domain and as many as 1927 domains. We have evaluated two methods for selecting saHMM-members into the midnight ASTRAL set. Both methods are modified versions of the algorithms described by Hobohm *et al.* [4]. The algorithms select, for each SCOP family, only those domains that were determined by X-ray crystallography to a resolution of 3.6 Å or better, and have mutual sequence identities equal to or less than $p^I(L,0)$.

Within each family we construct pairwise structural superimpositions in order to obtain the percent sequence identities. The coordinate files of the domains are obtained from the ASTRAL compendium [2] corresponding to SCOP version 1.69 [9]. We have evaluated several structure alignment programs, and found that, currently, MUSTANG [5] results in the best performing saHMMs (to be published elsewhere). In case the program fails to align two structures, the pair of domains is treated like a pair with too high sequence identity. As a

minimum requirement for building an saHMM, the SCOP domain family must be represented by at least two structures. Therefore, all families with only one representative were excluded from the midnight ASTRAL set.

All computations were done in parallel, using up to 20 processors on the HPC2N Linux cluster Seth. The compute nodes on Seth are AMD Athlon MP2000+ with 1GB of memory per dual node, connected in a high-speed SCALI network.

2.1 Algorithm 1 for Selecting saHMM-Members

Algorithm 1 is designed to limit the number of structural comparisons. It works by removing one of the domains in a pair from further consideration, if the mutual sequence identity falls above $p^I(L, 0)$.

Outline of Algorithm 1

1. Collect all family members with < 3.6 Å resolution into to-be-checked set.
2. Take domain `d1` from to-be-checked set, place in select set.
3. For each other domain `d2` in to-be-checked set.
   (a) Pairwise structural alignment of `d1` and `d2` to determine sequence identity sI and alignment length L.
   (b) If $sI > p^I(L, 0)$ then `dToRemove` = `selectOne(d1,d2)`.
      i. place `dToRemove` in to-remove set.
      ii. if `dToRemove` = `d1` repeat from 2.
4. Repeat from 2 until no more domains remain in to-be-checked set.

In order to retain the highest quality structures for constructing optimal structure superimpositions as the basis for the saHMMs, the algorithm selects the domain with the better resolution. In cases where the resolution values of the structures to be compared are too similar, i.e., they differ by less than 10% of their average, we exclude the domain with the higher mean thermal factor, B-factor. This rule applies in particular to domains extracted from the same PDB (Protein Data Bank) file. The mean B-factor reflects the data quality and is here calculated as the arithmetic mean of the B-factors for all C_α atoms within the domain. The function `selectOne` is used to select which domain to remove in case of high sequence identity.

Outline of function `selectOne`

1. Read in domains to compare: `d1` and `d2`
2. if |resolution(`d1`) - resolution(`d2`)| $< 0.1 \cdot$mean(resolution(`d1`), resolution(`d2`))
   (a) if the mean B-factor for `d1` is smaller than the mean B-factor of `d2`, then set `dToRemove = d2`
   (b) else set `dToRemove = d1`
3. else if resolution of `d2` is poorer than that of `d1`, then set `dToRemove = d2`
4. else set `dToRemove = d1`

After the first round of selection, all the preliminary discarded protein domains stored in the to-remove set are again compared to all domains in the select set, in order to assure that only domains with sequence identities above $p^I(L,0)$ are permanently discarded. The rationale behind this additional step is that in the process of removing domains, it is possible that a domain A is removed due to high sequence identity to domain B. If B is later removed due to high sequence identity to domain C, it could be that A and C have low mutual sequence identity. Thus A must be compared with C, and in case the identity is equal to or less than $p^I(L,0)$ both A and C must be kept.

2.2 Algorithm 2 for Selecting saHMM-Members

We evaluated a second algorithm, called Algorithm 2, which is designed to maximize the number of representative domains. Using Algorithm 2, one first fills an $n \times n$ score matrix M based on all-against-all structural comparisons of all n members within a particular SCOP family. An entry M_{ij} is a measure of the level of sequence identity and the relative data quality of domains d_i and d_j, and is defined as:

$$M_{ij} = \begin{cases} 1 & \text{if} \quad i = j, \\ 0 & \text{if} \quad sI \leq p^I(L,0), \\ 1 + 1/n & \text{if} \quad d_j = \texttt{dToRemove}, \\ 1 - 1/n & \text{if} \quad d_i = \texttt{dToRemove}. \end{cases} \tag{2}$$

Which domain to remove in case of too great sequence identity is determined using the same procedure `selectOne` as described for Algorithm 1. To select representative domains using M, we remove in each step the domain similar to most other domains, until no more similarities can be detected. The domain d_k, corresponding to row index k in M, which is similar to most other domains is the one with the highest row sum:

$$k = \text{argmax}_i(\sum_j M_{ij}). \tag{3}$$

Removing the domain d_k from the set corresponds to setting elements $M_{ki} = 0$ and $M_{ik} = 0$ for all i, including the diagonal element M_{kk}. The process is finished when $\max_i(\sum_j M_{ij}) = 1$. The representative domains are those with 1 on the diagonal ($M_{yy} = 1$ for all representatives y). For reasons described in Algorithm 1, all removed domains are checked once more against all selected domains to make sure that no representatives were mistakenly discarded.

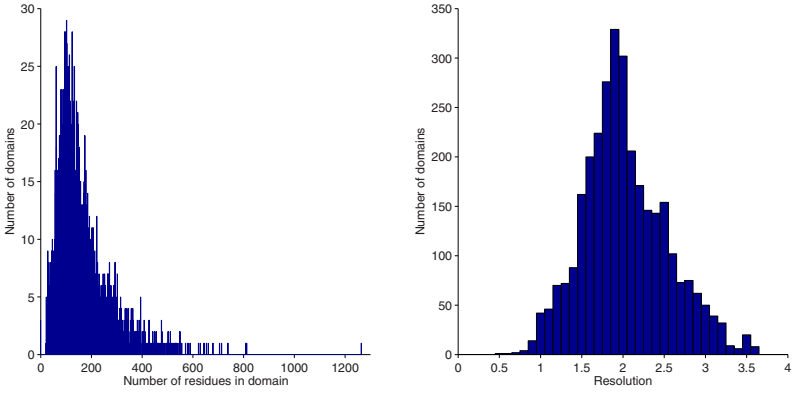
2.3 Comparing Algorithm 1 and Algorithm 2

Calculations using Algorithm 1 result in 3129 domains in the midnight ASTRAL set, representing 850 different SCOP domain families. These families cover 65% of the SCOP domains and correspond to 30% of the SCOP families belonging to true classes. Algorithm 2 gives 3293 domains in the midnight ASTRAL set, which represent 894 SCOP domain families. These families cover about 60% of SCOP domains and correspond to 31% of the true class SCOP families.

The advantage of Algorithm 2 is that it produces more saHMM-members for the midnight ASTRAL set. However, it is time expensive due to the all-against-all structural comparisons, which cause the problem to scale quadratically with the number of domains. It was not practical to use Algorithm 2 for the four very largest families, each harbouring more than 600 domains. Even so, the computing time used to select representative domains with Algorithm 2 exceeded the total time used by Algorithm 1 by an order of magnitude. We therefore decided against Algorithm 2, and will from now on use Algorithm 1 to select saHMM-members, even though Algorithm 1 results in a slightly reduced coverage of SCOP families.

2.4 Analysis of the Midnight ASTRAL Set

In Fig. 1(a) the distribution of lengths of domains within the midnight ASTRAL set selected with Algorithm 1 is displayed. The sharp peak shows that the most common sequence length of the saHMM-members is about 100 residues. The length varies from 21 amino acids for the shortest domain up to 1264 residues for the longest. In Fig. 1(b) the distribution of resolutions at which the structures of the domains were determined is displayed. The majority of the crystal structures from which the domains are extracted fall into the resolution range between 1.5 to 2.5Å. This assures a high confidence in the determined structures.



(a) (b)

Fig. 1. Distribution of (a) sequence lengths and (b) resolutions among domains in the midnight ASTRAL set

3 The saHMM Data Base

The construction of structure-anchored Hidden Markov Models, saHMMs, requires three major steps. First, the non-redundant midnight ASTRAL set must be generated as was described above. Then a multiple 3D superimposition of the peptide chains of these domains, called the saHMM-members, is constructed. By

using only spatial criteria to compare their structures, it is possible to match those amino acids that are from different chains and in close spatial vicinity, into a structure anchored multiple sequence alignment (see also [12]). The final step involves building the saHMMs from the deduced structure-anchored multiple sequence alignment.

The coordinate files of the saHMM-members are obtained from the ASTRAL compendium corresponding to SCOP version 1.69. The domains are superimposed with MUSTANG [5] and the saHMMs are built using HMMER 2.2g [3].

We implemented several Perl programs in order to automate the process from raw SCOP family classification of domains, through the construction of the midnight ASTRAL set, to the creation and testing of the saHMMs. The programs perform tasks such as detecting and correcting inconsistencies between the notations used in SCOP and the ASTRAL coordinate files, standardizing the notation used in the coordinate files and parsing of results to convert output from one program to input for another.

3.1 Coverage of SCOP

Since at least two structures are needed for superimposition, and because of the stringent sequence identity restrictions, our collection of saHMMs currently includes 850 saHMMs, which cover about 30% of the 2845 SCOP families belonging to true classes and 65% of the 67210 domain sequences. We expect these numbers to improve due to the exponential increase of deposited 3D structures.

4 The FISH Server

4.1 Design of the FISH Server

Flat file data bases were imported into a relational data base (MySQL implemented on a Linux platform) and cross-linked (Fig. 2). The user interface is written in Perl, PHP, and JavaScript and integrated with the Apache web server.

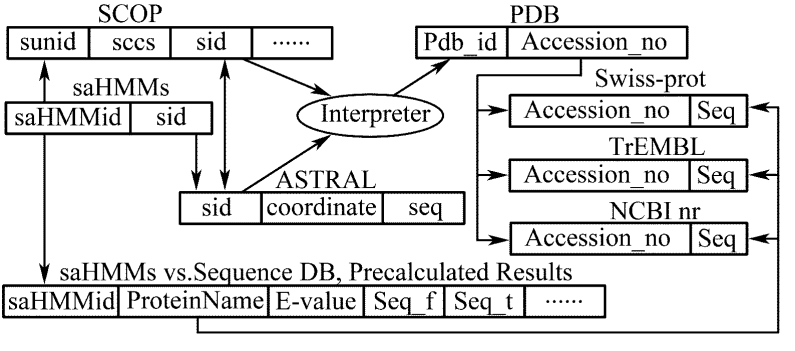


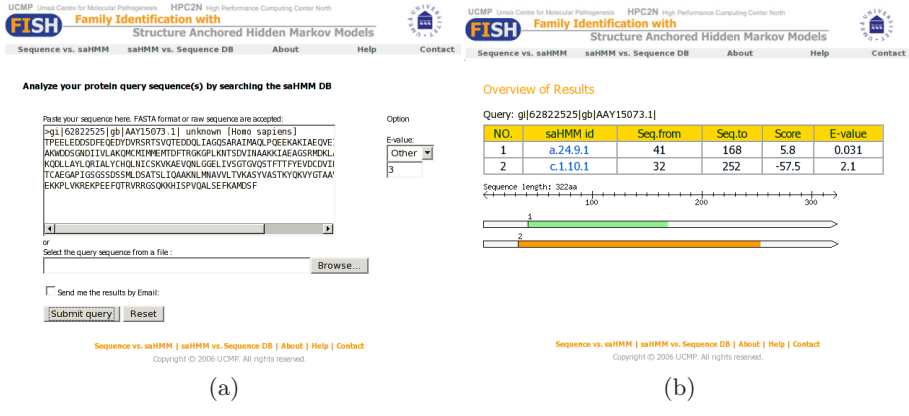
Fig. 2. Schematic view of the data base cross-linking used in the FISH server

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 also be sent to the user by e-mail in the form of a www-link and are stored on the server for 24 hours.

4.2 How to Use the FISH Server

Sequence Searches vs. the saHMM-db

Using the FISH server, a user can compare a query sequence with all models in the saHMM-db. 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. This information is useful for sequence annotation, to design mutations, to identify soluble domains, to find structural templates for homology modelling and possibly for structure determination by molecular replacement.



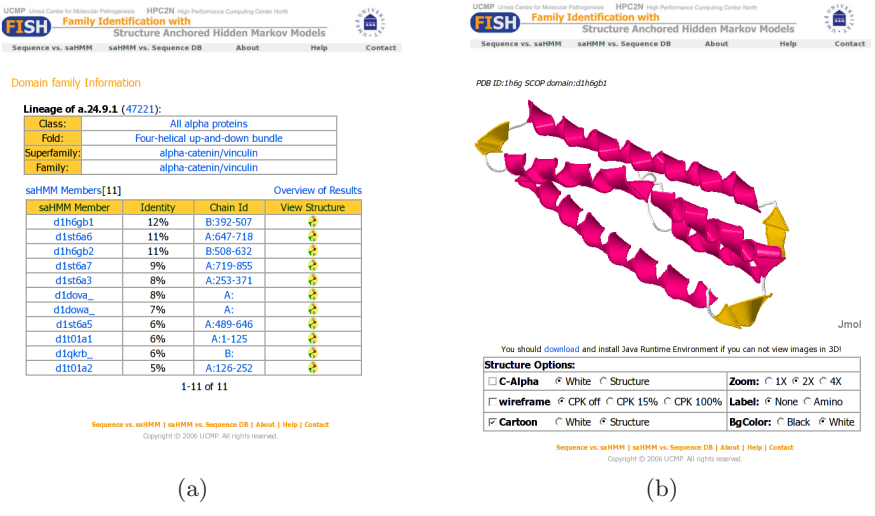
(a) (b)

Fig. 3. Sample (a) input and (b) results pages from a sequences vs saHMMs search

Fig. 3(a) displays an example of the input page. The user enters one or more query sequences and can select an E-value cut-off for the results. The E-value of a hit is the expected number of false matches having at least the same score as the hit, and hence is a measure of the confidence one can have in the hit. The closer the E-value is to zero, the more the match can be trusted. In the 'overview of results' page (Fig. 3(b)) the list of matches is sorted in increasing order with respect to the E-value, up to the chosen cut-off. When selecting one entry from the list, the family specific information for that match is displayed (Fig. 4(a)). The top table provides information about the SCOP classification. It is followed by a table listing all saHMM-members of this family together with details about, for example, the percent sequence identity of the query sequence aligned to the

member. For each saHMM-member, it is possible to view the structure of the selected domain in an interactive Java window, as shown in Fig. 4(b).

Below the list of matches in the 'overview of results' page (Fig.3(b)) is a horizontal bar graph representation of the query sequence, where matches are marked as coloured ranges. A light green range corresponds to an E-value of 0.1 or less, a yellow range to $0.1 \leq$ E-value ≤ 1.0 and an orange range for E-values above 1.0. Each coloured range links to a pairwise alignment of the query sequence and the saHMM consensus. The user has the option to display a multiple sequence alignment of the query sequence and the saHMM-member sequences in different formats. In addition, it is possible to reach a list with pairwise comparisons of the query and each saHMM-member. All alignments are anchored on the saHMM.



(a) (b)

Fig. 4. Example pages displaying (a) the domain family information of the top hit from Fig. 3(b) and (b) the structure view of the domain with highest sequence identity compared to the query sequence

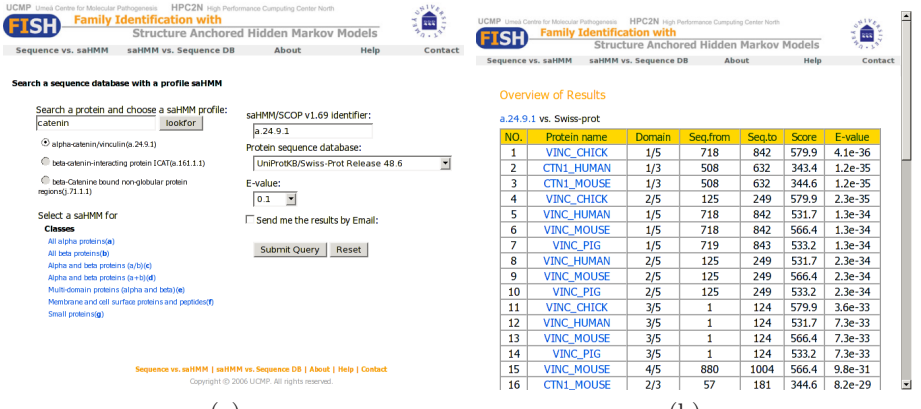
saHMM Searches vs. a Sequence Database

Furthermore, the FISH server allows the user to employ individual saHMMs for searching against a sequence data base to find those proteins that harbour a certain domain, independent of sequence identity and annotation status. For this purpose, the user can choose a particular saHMM from a list of available models and specify against which data base to perform the search. Currently, the Swiss-Prot, TrEMBL and the non-redundant data base, nr, from NCBI are available for searching. In addition, a user has the option to upload his/her own sequence database, as long as its size does not exceed 2 MB. 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 $p^I(L, 0)$. For each match, the user

obtains the corresponding sequence entry, as well as pairwise and multiple sequence alignments of the matched sequence and the saHMM-members, anchored on the saHMM. Information about the domain family used for searching is also easily available.

A search with a single saHMM vs. SwissProt can take from 15 minutes up to about nine hours. Searching TrEMBL, which is about ten times larger, takes considerably longer. In order to minimize the time a user has to wait for the results, we pre-calculated the searches of all 850 saHMMs vs. SwissProt, TrEMBL and 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. The computations were done in parallel, by searching the databases with several saHMMs concurrently, using up to 20 processors on the HPC2N Linux cluster Seth.

Fig. 5 shows an example of (a) the input page and (b) the results page of a search with an saHMM versus a sequence database. In the example, SwissProt was used. The results of the search are represented in form of a list sorted by E-value up to the user-specified cut-off.



NO.	Protein name	Domain	Seq.from	Seq.to	Score	E-value
1	VINC_CHICK	1/5	718	842	579.9	4.1e-36
2	CTN1_HUMAN	1/3	508	632	343.4	1.2e-35
3	CTN1_MOUSE	1/3	508	632	344.6	1.2e-35
4	VINC_CHICK	2/5	125	249	579.9	2.3e-35
5	VINC_HUMAN	1/5	718	842	531.7	1.3e-34
6	VINC_MOUSE	1/5	718	842	566.4	1.3e-34
7	VINC_PIG	1/5	719	843	533.2	1.3e-34
8	VINC_HUMAN	2/5	125	249	531.7	2.3e-34
9	VINC_MOUSE	2/5	125	249	566.4	2.3e-34
10	VINC_PIG	2/5	125	249	533.2	2.3e-34
11	VINC_CHICK	3/5	1	124	579.9	3.6e-33
12	VINC_HUMAN	3/5	1	124	531.7	7.3e-33
13	VINC_MOUSE	3/5	1	124	566.4	7.3e-33
14	VINC_PIG	3/5	1	124	533.2	7.3e-33
15	VINC_MOUSE	4/5	880	1004	566.4	9.8e-31
16	CTN1_MOUSE	2/3	57	181	344.6	8.2e-29

(a) (b)

Fig. 5. Example of (a) input and (b) results of a search with the catenin saHMM (a.24.9.1) vs SwissProt version 1.69. Only the top part of the results page is shown.

5 Conclusions

The foundation of the structure-anchored hidden Markov model method is the 3D superimposition of carefully chosen domains representing the SCOP domain family to be modelled. For the selection of the representative domains, called the saHMM-members, we evaluated two algorithms, Algorithm 1 and Algorithm 2. Even though the use of Algorithm 2 results in 164 more saHMM-members in the midnight ASTRAL set, which leads to 44 more saHMMs, we prefer Algorithm 1 since it is more than an order of magnitude faster and can handle even the largest families in a reasonable amount of time. The resulting saHMMs together constitute the saHMM-db, which covers 30% of the SCOP families and

65% of the domains belonging to true classes. So far, every new SCOP release has lead to new saHMMs and has increased the number of saHMM-members for many families. As the number of deposited structures grows, we anticipate that the saHMM-db will cover more of SCOP. In addition, we expect that new domain sequences will be added to families, which in turn increases the number of saHMM-members and improve saHMMs with only few saHMM-members. The saHMM-db is publicly available through the FISH server, which is a powerful and versatile tool with dual function. On the one hand, the user can perform sequence searches versus the saHMM-db, and possibly obtain matches even for remote homologues, 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 sequence alignments and the structures of all representatives are known, the alignment of a sequence to the saHMM-members can be used to draw conclusions about the secondary and tertiary structures of the sequence.

Acknowledgements. We thank Åke Sandgren for assistance with parallel computing and Marek Wilczynski for support with the FISH server. Part of this research was conducted using the HPC2N cluster resources. B.K. acknowledges the partial support for this project by the Swedish Foundation for Strategic Research grant A3.02:13. U.H.S. acknowledges the partial support for this project by a grant from the Knowledge Foundation (KK-Stiftelsen).

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, D138–D141 (2004)
2. 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, D189–D192 (2004)
3. Eddy, S.R.: Profile Hidden Markov Models. Bioinformatics 14, 755–763 (1998)
4. Hobohm, U., Scharf, M., Schneider, R., Sander, C.: Selection of representative protein data sets. Protein Science I, 409–417 (1992)
5. Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., Lesk, A.M.: MUSTANG: A multiple structural alignment algorithm. PROTEINS: Structure, Function, and Bioinformatics 64, 559–574 (2006)
6. 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, D257–D260 (2006)
7. Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C., Gough, J.: The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Research 32, D235–D239 (2004)
8. Mika, S., Rost, B.: UniqueProt: creating representative protein sequence sets. Nucleic Acids Research 31, 3789–3791 (2003)

9. 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, 536–540 (1995)
10. Rost, B.: Twilight zone of protein sequence alignments. Protein Engineering 12, 85–94 (1999)
11. 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, 309–323 (1992)
12. Tångrot, J.: The Use of Structural Information to Improve Biological Sequence Searches. Lic. Thesis, UMINF-03.19. Dept. of Comput. Sci., Umeå Univ. (2003)
13. Tångrot, J., Wang, L., Kågström, B., Sauer, U.H.: FISH – family identification of sequence homologues using structure anchored hidden Markov models. Nucleic Acids Research 34, W10–W14 (2006)