PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Accurate domain identification with structure-anchored hidden Markov models, saHMMs

Jeanette E. Tångrot,[1,2,3,5] Bo Kågström,[3,4] and Uwe H. Sauer[1,2,5]*

[1] Umeå Centre for Molecular Pathogenesis, UCMP, Umeå University, Umeå SE-901 87, Sweden

[2] Department of Chemistry, Umeå University, Umeå SE-901 87, Sweden

[3] Department of Computing Science, Umeå University, Umeå SE-901 87, Sweden

[4] High Performance Computing Center North, Umeå University, HPC2N, Umeå SE-901 87, Sweden

[5] Computational Life Science Cluster, CLiC, Umeå University, Umeå SE-901 87, Sweden

ABSTRACT

The ever increasing speed of DNA sequencing widens the discrepancy between the number of known gene products, and the knowledge of their function and structure. Proper annotation of protein sequences is therefore crucial if the missing information is to be deduced from sequence-based similarity comparisons. These comparisons become exceedingly difficult as the pairwise identities drop to very low values. To improve the accuracy of domain identification, we exploit the fact that the three-dimensional structures of domains are much more conserved than their sequences. Based on structure-anchored multiple sequence alignments of low identity homologues we constructed 850 structure-anchored hidden Markov models (saHMMs), each representing one domain family. Since the saHMMs are highly family specific, they can be used to assign a domain to its correct family and clearly distinguish it from domains belonging to other families, even within the same superfamily. This task is not trivial and becomes particularly difficult if the unknown domain is distantly related to the rest of the domain sequences within the family. In a search with full length protein sequences, harbouring at least one domain as defined by the structural classification of proteins database (SCOP), version 1.71, versus the saHMM database based on SCOP version 1.69, we achieve an accuracy of 99.0%. All of the few hits outside the family fall within the correct superfamily. Compared to Pfam_ls HMMs, the saHMMs obtain about 11% higher coverage. A comparison with BLAST and PSI-BLAST demonstrates that the saHMMs have consistently fewer errors per query at a given coverage. Within our recommended *E*-value range, the same is true for a comparison with SUPERFAMILY. Furthermore, we are able to annotate 232 proteins with 530 nonoverlapping domains belonging to 102 different domain families among human proteins labelled "unknown" in the NCBI protein database. Our results demonstrate that the saHMM database represents a versatile and reliable tool for identification of domains in protein sequences. With the aid of saHMMs, homology on the family level can be assigned, even for distantly related sequences. Due to the construction of the saHMMs, the hits they provide are always associated with high quality crystal structures. The saHMM database can be accessed via the FISH server at http://babel.ucmp.umu.se/fish/.

Proteins 2009; 76:343–352.
© 2008 Wiley-Liss, Inc.

Key words: protein domain; structure alignment; remote homologue; sequence annotation; protein family; protein superfamily.

INTRODUCTION

Ongoing genome sequencing projects produce an exponentially increasing number of new sequences. It is common to deduce information about their function and possibly structure from already characterized homologues[1] instead of by means of experiments. Due to the modularity of proteins it is advisable to characterize their constituent domains rather than the protein as a whole. The aim of our approach is to accurately assign domains to their proper structural family, in order to support the assignment of function.

Finding a needle in a haystack is a relatively simple task compared to finding a particular pin in a stack of needles. A similar situation arises when one attempts to

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: Swedish Foundation for Strategic Research; Grant number: A3.02:13; Grant sponsor: Knowledge Foundation (KK-stiftelsen).
*Correspondence to: Uwe H. Sauer, Department of Chemistry, KBC-building, Linneus väg 6, Umeå University, Umeå SE-901 87, Sweden. E-mail: uwe.sauer@ucmp.umu.se.
Received 21 April 2008; Revised 13 November 2008; Accepted 18 November 2008
Published online 11 December 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22349

find pairs of homologous sequences sharing very low sequence identity. The task is difficult in view of the fact that for a certain alignment length, L, the sequence identities of homologous pairs are virtually undistinguishable from the sequence identities of randomly picked sequences as the pairwise sequence identities decrease from about 20% towards zero.[2,3]

Park *et al.* showed that profile hidden Markov models, HMMs, outperform other methods in detecting remote homologues, in particular methods based on pair-wise alignments.[4] It is crucial that the HMMs are built from reliable multiple sequence alignments, MSAs, in order to best represent the families of sequences they model. Most MSA programs neglect structural information and base their alignment steps solely on sequence information, on certain evolutionary models and on statistical analysis. Sequences with mutual identities above 20–30%, depending on alignment length L, can be aligned by standard alignment tools. However, as the mutual sequence identities fall below a soft boundary at roughly 20%, often referred to as the "twilight zone",[2,3,5,6] existing methods might not be able to produce reliable alignments.

In the twilight zone one can no longer determine whether two aligned protein sequences are homologous or not, in case the decision is solely based on the percent sequence identity after optimal alignment. As the level of sequence identity drops below the twilight zone and into the midnight zone, this task becomes very challenging or even impossible. This indicates the need for a sequence search tool that is capable of recognizing similarities in proteins even at very low levels of sequence identity.

To overcome the difficulties of sequence-only alignments, our method makes use of the fact that the three-dimensional (3D), structures of homologous protein domains are evolutionary more conserved than their amino acid sequences.[7–11] It is quite usual that the peptide chains of two domains with a very low sequence identity, clearly in the midnight zone, adopt almost identical 3D-structures, which means that their main-chain atoms are superimposable with a low root mean square distance (RMSD). The inclusion of structural information has in many cases improved the ability to find remote relationships. Secondary structure information was used in addition to the sequences to construct so called ssHMMs.[12] Tertiary structure superimpositions were used to generate substitution matrices,[13–15] to construct sequence profiles,[16,17] and to build hidden Markov models.[18–23]

Hidden Markov models are the most powerful of the profile methods, and have been used in a variety of ways. Gough *et al.*[24] used all individual structural classification of proteins database (SCOP)[25] superfamily sequences, with less than 95% mutual sequence identity, as seeds to construct one HMM from each seed. Their library of HMMs, called SUPERFAMILY, represent essentially all proteins of known structure. In a similar manner, Buchan *et al.* generated sets of HMMs to represent each CATH superfamily.[26] Others have constructed HMMs by explicitly including structural information.[18,19,21–23] Except for Al-Lazikani *et al.*[18] and Griffiths-Jones and Bateman,[19] all studies were carried out on the superfamily level. However, the family level provides much more detailed domain specific information for accurate annotation.

One conclusion that can be drawn from the body of work mentioned above is that the inclusion of structural information positively affects the accuracy of sequence alignments for remote homologues[21] and that HMMs built from structure linked alignments complement sequence only methods, in particular at the edge of the twilight zone.[23]

Even though the inclusion of many sequences into a HMM will improve its statistics, the large number of sequences might not be essential if one instead ensures that the HMMs are built from MSAs that contain a well balanced distribution of very diverse sequences within a particular family. At the core of our method lie multiple structure superimpositions of homologous domains belonging to the same family. To maximize the sequence diversity of the alignments, we only include domains whose mutual sequence identities fall below a curve defined by the function $p^I(L,0)$ (see Materials and Methods section), which is related to the HSSP-curve.[3,6] These domain sequences, which we refer to as saHMM-members, are collected into our "midnight ASTRAL set"[20,27] and are used for multiple structure superimpositions. Based on structural criteria, structure-anchored multiple sequence alignments, saMSAs, are assembled and used to build structure-anchored hidden Markov models (saHMMs), each representing one SCOP family. We assume that the saMSAs provide a less biased indicator of the evolutionary variability at each aligned position as compared to MSAs based on statistical methods. Our results demonstrate that using the saMSAs of only a few distantly related homologues is sufficient to capture the essence of an entire domain family.

The main steps involved in constructing the database of saHMMs are displayed in Figure 1. The database can be used in two ways: (i) A query sequence can be searched against the saHMMs in order to find which of the saHMMs gives the highest score, that is describes the sequence best, thus identifying the domain family the sequence most likely belongs to. In case the query sequence comprises two or more domains, one can expect one hit with low E-value for each domain. (ii) The saHMM describing a particular family can be used to search protein sequence databases or translations of newly sequenced genomes for hitherto unidentified members of that family. In either case, saHMMs are able to identify domains in proteins as belonging to one specific family and not to another family within the same superfamily. A match with low E-value provides the user not only with a family membership of the identified domain,

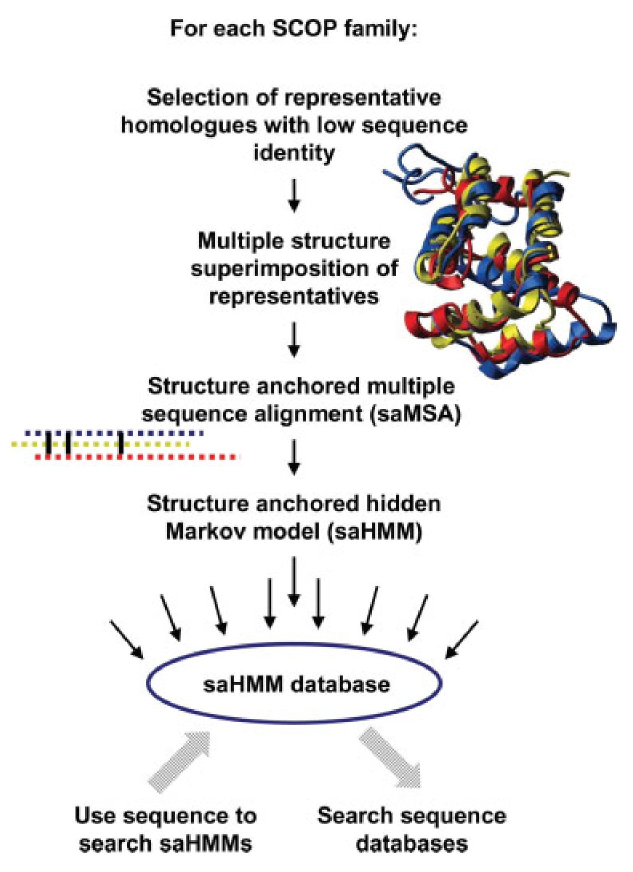
For each SCOP family:
Selection of representative homologues with low sequence identity
Multiple structure superimposition of representatives
Structure anchored multiple sequence alignment (saMSA)
Structure anchored hidden Markov model (saHMM)
saHMM database
Use sequence to search saHMMs
Search sequence databases

Figure 1

Steps involved in constructing the saHMM database. For each SCOP family we select only homologous domains with low pairwise sequence identity. For these domains we generate a multiple structure superimposition. The resulting structure-anchored multiple sequence alignment, saMSA, is used as input for building a structure-anchored hidden Markov model, saHMM, which becomes part of the saHMM database.

and hence a hint to its function, but also with an associated high resolution 3D-structure homologue.

METHODS

Structural superimposition of domains

For all structural superimpositions we use MUSTANG.[28] The program is, in principle, able to superimpose any number of structures and can produce structure-anchored sequence alignments in msf-format, which is suitable for input to HMMER.[29] MUSTANG proved to be best suited for our automated saHMM construction pipeline.

The midnight ASTRAL set

For the definition of homologous structural domains, we apply the SCOP classification on the family level.[25] The SCOP database version 1.69 contains 70,859 domains which are divided into 11 classes. We use only the seven true classes, which comprise 2845 domain families harbouring 67,220 domains. Excluded are the entries listed as the "Not a true class" such as coiled-coil proteins, peptides, low resolution structures and designed proteins. The SCOP associated ASTRAL compendium[30] provides Protein Data Bank,[31] PDB-style coordinate files for individual domains.

The PDB, and, consequently, the SCOP and ASTRAL databases are highly redundant.[32,33] To assure maximum sequence diversity within each family we include only sequences whose mutual sequence identities are equal to or less than the limiting curve $p^I(L,0)$. The function $p^I(L,n)$ is defined as

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

and depends on the alignment length L and on n, which can be interpreted as the distance, expressed in percent, from the "midnight zone curve" $p^I(L,0)$, for which n is equal to zero. The definition of $p^I(L,n)$ is similar to the HSSP-curve,[2,3,6] except that we extract the alignment length and percent sequence identity from structure superimpositions. The HSSP-curve with $n = 0$ was derived in such a way that two random sequences are in the majority of the cases homologous, if their pairwise sequence identity lies above the curve.[2,3]

A flowchart outlining the selection algorithm[27] for family representatives is depicted in Figure 2. The algorithm selects, for each 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 $p^I(L,0)$. These selection criteria will ensure a wide evolutionary spread of the homologous representatives and avoid sequence bias.

Within each family we construct pairwise structural superimpositions, in order to obtain pairwise structure-anchored sequence alignments from which we calculate percent sequence identities. If the sequence identity of a pair of superimposed domains falls above $p^I(L,0)$, we preliminarily discard the domain with the worse resolution. If the resolution values of the two structures differ from their average by less than 10%, we choose the domain with the lower mean thermal factor, B-factor. The mean B-factor is calculated as the average of the B-factors for all Cα atoms in the domain and reflects the data quality. In case of equal mean B-factors, one domain is chosen randomly.

After the first round of selection, all the preliminarily discarded domains are again compared to all remaining domains, in order to assure that only those domains with sequence identities above $p^I(L,0)$ are permanently discarded. This additional step insures that a sequence A

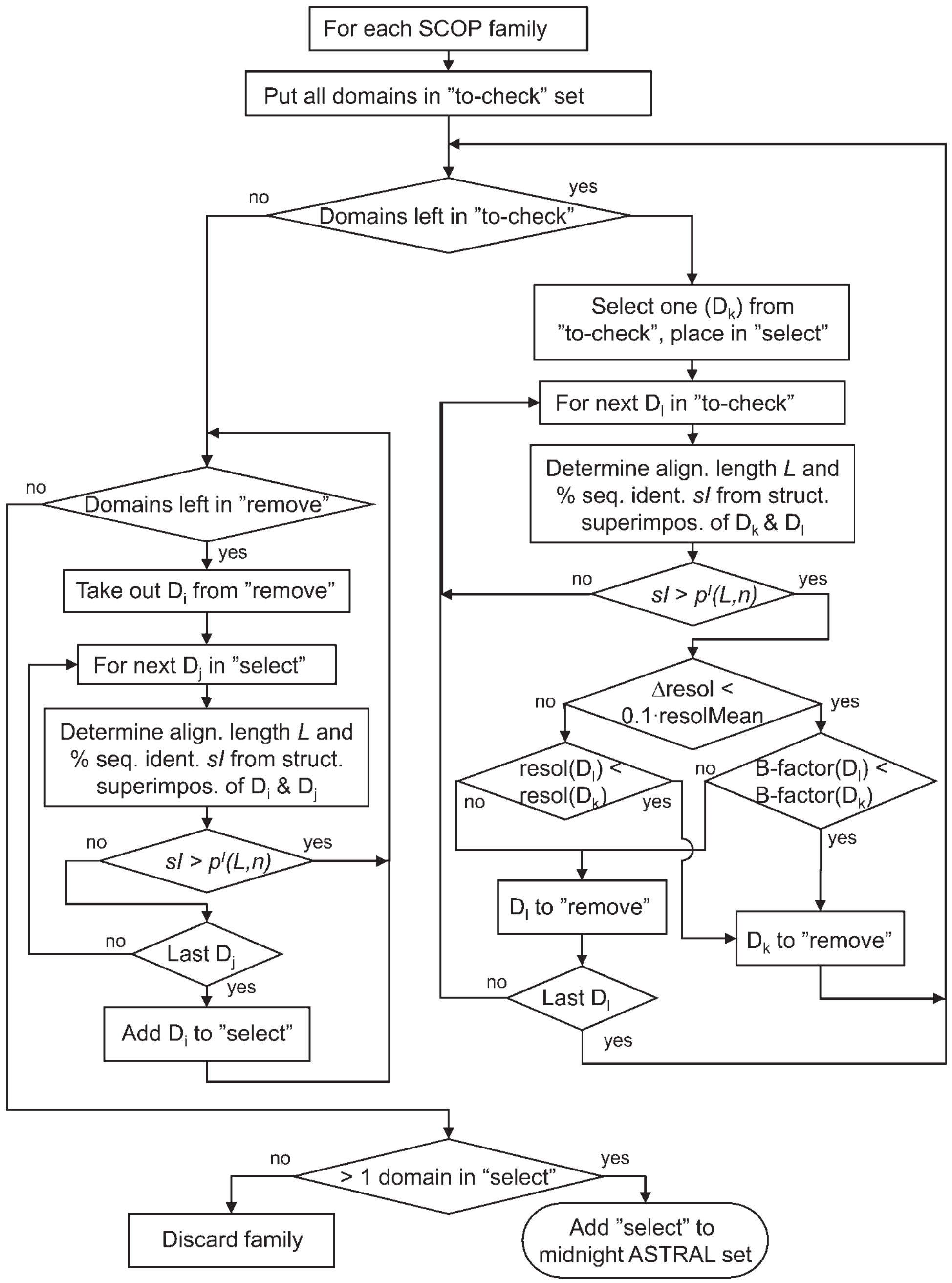
For each SCOP family
Put all domains in ”to-check” set
no
yes
Domains left in ”to-check”
Select one (D_k) from ”to-check”, place in ”select”
For next D_l in ”to-check”
Determine align. length L and % seq. ident. sI from struct. superimpos. of D_k & D_l
no
yes
$sI > p^I(L,n)$
Δresol < 0.1·resolMean
no
yes
resol(D_l) < resol(D_k)
no
yes
no
B-factor(D_l) < B-factor(D_k)
yes
D_l to ”remove”
D_k to ”remove”
no
Last D_l
yes
no
Domains left in ”remove”
yes
Take out D_i from ”remove”
For next D_j in ”select”
Determine align. length L and % seq. ident. sI from struct. superimpos. of D_i & D_j
no
yes
$sI > p^I(L,n)$
no
Last D_j
yes
Add D_i to ”select”
no
yes
> 1 domain in “select”
Discard family
Add ”select” to midnight ASTRAL set

Figure 2

Flowchart showing the construction of the midnight ASTRAL set. For each family in SCOP we structurally superimpose pairs of family members using MUSTANG. If the resulting structure-derived sequence identity of a pair falls above $p^I(L,0)$, we preliminarily place the domain with the worse resolution into the “remove” set. In case of similar resolutions, the domain with the higher mean B-factor is put into the “remove” set. After the first round of selection, all the protein domains in the “remove” set are again compared to the remaining domains. This will assure that only domains with sequence identities above $p^I(L,0)$ are permanently discarded.

which was removed due to high identity to sequence B is not unnecessarily discarded. It is possible that B will be later removed due to high identity to sequence C, and it is conceivable that A and C have low sequence identity. Hence, A must also be compared to C, and in case the identity is equal to or less than $p^I(L,0)$, both A and C are kept. The selected domain sequences, called *saHMM-members*, are taken as representatives for this particular family and are collected in the midnight ASTRAL set. As a minimum requirement for building an saHMM, the domain family must be represented by at least two structures. We therefore exclude from the midnight ASTRAL set, all families with only one representative.

Construction of saHMMs

To build saHMMs, we first construct a structure-anchored multiple sequence alignment of the saHMM-members within each family. The saMSAs are then used as input for HMMER version 2.2 g[29] with default parameters for hmmbuild. All saHMMs are calibrated using hmmcalibrate with default settings to obtain fitted *E*-values. In this way, we created 850 saHMMs, one saHMM for each SCOP protein domain family represented in the midnight ASTRAL set.

Evaluation of the performance

The performance at a given *E*-value threshold *e* is evaluated with respect to the following two criteria: the coverage, which is expressed as the percentage of all domains that are matched with the correct saHMM with an *E*-value less than or equal to *e*, and the accuracy, which stands for the percentage of all hits with an *E*-value of at most *e*, that are correct. For all our evaluations and for all methods compared, we only consider matches that cover at least 70% of the domain length.

Matches between a domain sequence and an saHMM from the same family are counted as correct hits, also called true positives, *tp*. Unless otherwise stated, all hits outside the family are considered as false positives, *fp*, even if they fall into the correct superfamily.

For all searches we use HMMER version 2.3.2[29] and, unless otherwise stated, the *E*-value cutoff is set to $e = 0.01$ for sequences searches versus saHMMs.

Determining errors per query and coverage

The number of errors per query (EPQ), is calculated as the total number of *fp* considering a certain *E*-value threshold *e*, divided by the total number of queries which is equal to the number of domains used for searching. The coverage is calculated as described in the previous section.

The advantage of the EPQ versus coverage graphs, compared to receiver operating characteristic, curves,[34,35] is that they communicate essentially the same information, while the EPQ versus coverage plots better represent the high degrees of accuracy and the vast background of nonhomologues encountered in sequence comparisons.[36]

Construction of exclude-one-saHMMs

For the 387 SCOP families with at least three saHMM-members, we construct so called exclude-one-saHMMs, exo-saHMMs, by excluding one representative sequence at a time and building new saHMMs from the superimposition of the remaining domains. In this way, we obtain a collection of *n* exo-saHMMs for a family with *n* saHMM-members. We then examine whether each of the excluded sequences can be matched with the exo-saHMM that lacks that sequence. When analysing the results on the superfamily level, we do not only count matches to exo-saHMMs as correct, but also matches to saHMMs within the proper superfamily.

Ability to find new members of a family

We use the full length protein sequences corresponding to those domains in SCOP version 1.71 (released: October 2006) that are not present in SCOP version 1.69 (released: July 2005), to search against the saHMMs, which are based on SCOP 1.69. In addition, for each family, we identify among the new domains those that have a sequence identity equal to or less than $p^I(L,0)$ compared to the saHMM-members. We then use the full length sequences harbouring these low identity domains to examine if the saHMMs are able to assign them to the correct family.

Comparing saHMMs to BLAST and PSI-BLAST

We evaluate the ability of BLAST and PSI-BLAST[37] to assign a sequence to the correct family, even at low sequence identity, and compare the results to those obtained from searches with saHMM-members versus exo-saHMMs, as described previously. Full length protein sequences corresponding to all domain sequences in the midnight ASTRAL set are used, one at a time, as queries in BLAST and PSI-BLAST searches.

For the BLAST search we use blastall (2.2.13) with default parameters to search each sequence against SCOP 1.69.

In the case of PSI-BLAST (blastpgp 2.2.13) we initially carry out five iterations versus the National Center for Biotechnology Information (NCBI) nr-database (downloaded: March 30, 2006). As threshold for including a sequence we use an *E*-value cutoff of 0.001. For all other parameters we use the default values throughout.

The resulting position specific scoring matrix, PSSM, one for each query sequence, is thereafter used to search against SCOP 1.69.

For a BLAST or PSI-BLAST search, we consider a domain as assigned to the correct family if its sequence or PSSM, respectively, obtains a match to at least one midnight ASTRAL set sequence from the same family, not counting self-hits. Matches to midnight ASTRAL set sequences outside the correct family are counted as false positives. For comparisons carried out on the superfamily level, we count matches to midnight ASTRAL set sequences within the proper superfamily as true positives.

Table I
Performance at Low Sequence Identity

	Number of sequences	Accuracy on family level (%)	Coverage on family level (%)	Accuracy on superfamily level (%)[a]	Match within correct superfamily (%)[b]	Sequences without hits (%)
E-value $\leq$ 0.01	2127	92.5	37.6	100.0	40.7	61.4
E-value $\leq$ 0.1	2127	88.0	45.6	99.5	51.5	52.4
E-value $\leq$ 10, top hits	2127	73.4	58.2	79.6	63.1	20.8

The full length sequences harbouring the domains excluded from the saHMMs are searched against the corresponding exo-saHMMs and the remainder of the saHMM database, as described in the text. The first and second row report results obtained at an *E*-value cut-off of 0.01 and 0.1, respectively. The results reported in the third row consider only the top match for each domain at an *E*-value cut-off of less than or equal to 10.

[a] Percentage of all hits that fall within the correct superfamily.

[b] Percentage of all sequences that obtain a hit within the correct superfamily.

Comparing saHMMs to SUPERFAMILY

Each SUPERFAMILY[24] HMM corresponds to one SCOP domain. Using FASTA version 34.26.5,[38] we can assign each saHMM-member to the corresponding SUPERFAMILY HMM (version 1.69) using a 95% sequence identity cutoff.

As before, we use full length protein sequences harbouring the domains of the midnight ASTRAL set and search all SUPERFAMILY HMMs using HMMER. If a domain obtains a match to at least one HMM corresponding to a midnight ASTRAL set sequence from the same family we count it as a true positive match. Self-hits are not considered. Matches to HMMs representing midnight ASTRAL set sequences outside the correct family are counted as false positives. However, when the analysis is carried out on the superfamily level, matches within the correct superfamily are counted as true positives.

Comparing saHMMs to Pfam HMMs

The classification of domains in Pfam[39] is not identical to that of SCOP. Therefore, we have mapped Pfam (version 19.0, released: November 2005) onto SCOP 1.69. The relationships between corresponding families in the two databases are established by finding the SCOP classification of PDB sequences that are part of Pfam-A seed alignments. For the comparison, we use as queries those full length sequences harbouring domains that are new in SCOP 1.71 and belong to families with both an saHMM based on SCOP 1.69 and an Pfam_ls HMM, version 19.0.

RESULTS AND DISCUSSION

The midnight ASTRAL set and corresponding saHMMs

Our midnight ASTRAL set[27] contains 3129 low identity, nonredundant domains. The domains correspond to 850, out of 2845, SCOP "true class" families. Each family is represented by at least two, and up to 38, low identity domains called the saHMM-members, from which one saHMM per family is automatically constructed.

The ability to find low sequence identity homologues

In the following, we analyze the ability of the saHMMs to identify the proper family for low identity sequences, that is to say sequences whose identity is equal to or less than $p^I(L,0)$ when compared to each one of the saHMM-members.

To carry out the "search for a specific pin in a stack of needles", we construct exclude-one-saHMMs, termed exo-saHMMs, for the domain families with at least three saHMM-members. The full length sequences corresponding to the 2127 excluded domain sequences are used, one at a time, to query the saHMM database, with one modification: for each of the query sequences we exchange the full family saHMM with the exo-saHMM that lacks that domain sequence. The search results show that 37.6% of the excluded domains can be matched to the corresponding exo-saHMM and an additional 3.1% obtain hits to saHMMs belonging to the correct superfamily (see Table I). Taken together, we obtain 865 hits, of which 92.5% are within the correct domain family.

If we relax the *E*-value cutoff to 10 and consider only the top scoring hit per domain, the coverage increases to 58.2% at the cost of reduced accuracy (see Table I). Among the top scoring matches, 79.6% are within the correct superfamily. The coverage values can be interpreted as the probability of assigning the correct family to a sequence with very low sequence identity compared to the saHMM-members.

The results show that the exo-saHMMs are able to detect very low identity homologues with high accuracy. The majority of the domains for which we obtain a hit are matched to the correct family, and the majority of the matches outside the family fall within the correct superfamily. This property demonstrates the usefulness of the saHMMs to assign the correct family to remote homologues.

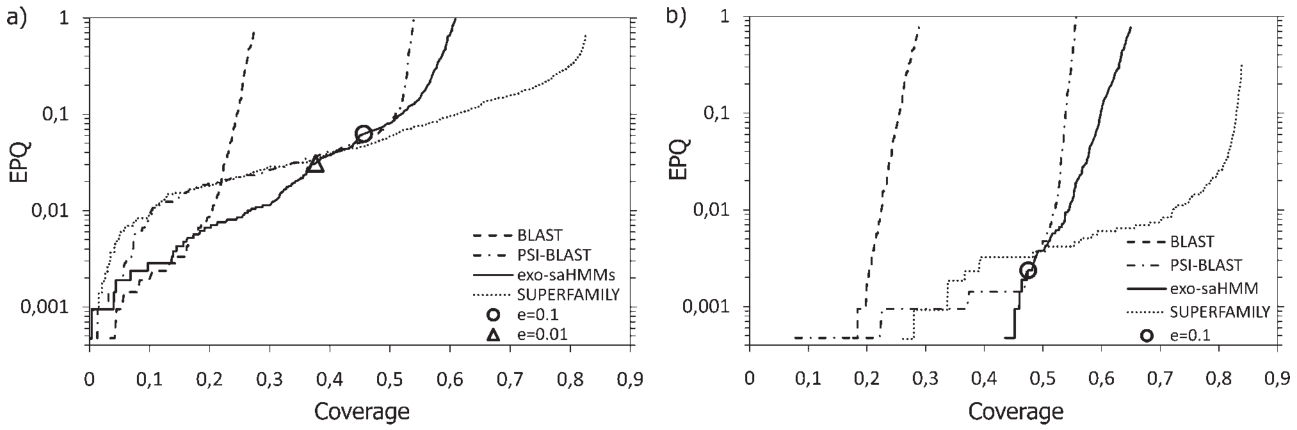

Figure 3

EPQ versus coverage plots. Single-logarithmic plots of Errors Per Query, EPQ, versus coverage for (**a**) the family level and (**b**) the superfamily level. Plotted are the results from searches with low identity sequences versus BLAST (dashes), PSI-BLAST PSSMs (dashes-dots), exo-saHMMs (solid line), and SUPERFAMILY HMMs (dots). Note that the exo-saHMM curve in (a) lies below the PSI-BLAST and SUPERFAMILY curves for a coverage below about 42% and EPQ values less than 0.04. Marked on the exo-saHMM curve are the *E*-values $e = 0.1$ (circle) and $e = 0.01$ (triangle). In (b) the mark for the *E*-value 0.01 is outside the range of the plot, as the EPQ value is zero in this case. For searches using the FISH server, the default *E*-value cutoff is set to 0.01. Note that below about 5% coverage the curves in (a) are noisy and difficult to interpret, due to the sparsity of data points.

Comparing exo-saHMMs to BLAST, PSI-BLAST PSSMs, and SUPERFAMILY HMMs

Using EPQ versus coverage plots we compare the performance of the exo-saHMMs to BLAST, PSI-BLAST PSSMs, and SUPERFAMILY HMMs (see Figure 3). The sequence identities of the query sequences, when compared to those sequences used to build the exo-saHMMs, fall below $p^I(L,0)$. It should be noted that PSI-BLAST PSSMs and SUPERFAMILY HMMs have an advantage in this comparison since they are derived from sequences obtained from iterative searches of the NCBI's nr-database. Therefore, the PSSMs and HMMs might contain so called bridging sequences, with sequence identities above $p^I(L,0)$ as compared to both the query sequence and to one or more of the saHMM-members within the same family. These bridging sequences are likely to facilitate the correct matchmaking of PSI-BLAST PSSMs and SUPERFAMILY HMMs. In order to avoid such a situation, we would have to construct a separate nr-database for each query sequence, where all sequences with sequence identities above $p^I(L,0)$ with respect to the query are removed. For practical reasons, this is not feasible.

To ensure proper sequence annotations, it is important to consider only reliable matches, that is with as low EPQ values as possible. As can be seen in Figure 3a, below about 40–45% coverage and an EPQ value of about 0.04, PSI-BLAST PSSMs and the SUPERFAMILY HMMs are outperformed by the exo-saHMMs, which obtain clearly lower EPQ values for a given coverage. For low coverage values, the BLAST curve also shows high accuracy but reaches only about 20% coverage before the number of EPQ drastically increases.

When we evaluate the results on the superfamily level, the EPQ-values, at low coverage, drop compared to the family level for all methods (Fig. 3b). In particular, below 43% coverage, the number of EPQ is zero for the exo-saHMMs. As before, the exo-saHMMs perform better than, or similar to, BLAST and PSI-BLAST over the whole range of coverage values. Up to about 50% coverage, the exo-saHMMs perform better than the SUPERFAMILY HMMs as well. However, at higher coverage the number of EPQ increases rapidly for the exo-saHMMs, while SUPERFAMILY HMMs show a lower rate of EPQ increase compared to the other methods.

Together, the graphs in Figure 3 prove our point that the exo-saHMMs, and hence the saHMMs, are highly accurate in assigning the correct family to a sequence, even at low sequence identity. The graphs also demonstrate that, for *E*-value cutoffs of 0.1 and below, false positive matches on the family level fall almost exclusively within the correct superfamily.

Ability to recognize new sequences

To assess the ability of the saHMMs to assign the correct domain families to new sequences, we use the full length sequences harbouring the 4406 domains that are present in SCOP 1.71 but not in version 1.69 to search against the saHMMs. We find that 2612 of the domains belong to families for which we have an saHMM in the database based on SCOP 1.69. Among these domains,

Table II
Performance Considering New Sequences

		Number of sequences	Accuracy on family level (%)	Coverage on family level (%)	Accuracy on superfamily level (%)[a]	Match within correct superfamily (%)[b]	Sequences without hit (%)
All sequences		4406	99.0	48.3	100	48.8	51.2
Sequence with an saHMM in the saHMM db	All sequences	2612	99.9	81.4	100	81.5	18.5
	Low identity sequences	451	100	24.4	100	24.4	75.6
Sequence without saHMM in saHMM db		1794	—	—	100	1.1	98.9

The full length sequences corresponding to domains new in SCOP version 1.71 compared to version 1.69, are searched against the saHMM database. The results in the first row refer to all sequences. Rows two and three show results for sequences that belong to families with an saHMM in the database. Whereas the second row does not consider a sequence identity cutoff, the third row reports results only for sequences with low sequence identities, below $p^I(L,0)$, compared to the saHMM-members within their families. The results in the last row refer to sequences belonging to families that are not represented by an saHMM.

[a] Percentage of all hits that are within the correct superfamily.

[b] Percentage of all sequences that obtain a hit within the correct superfamily.

81.4% obtain a top score to the correct saHMM, with an *E*-value less than or equal to $e = 0.01$. The number of domains obtaining correct top scores increases only marginally if we allow matches within the superfamily. A summary of the results is presented in Table II.

The domain sequences without a corresponding saHMM should not obtain any hits. Accordingly, our search results show that only a small fraction of these orphan sequences obtain matches at all, which are exclusively to an saHMM from the correct superfamily. Considering all domains, 99.0% of the matches are to the correct family and all hits outside the correct family are within the proper superfamily.

Next, we evaluate the ability of the saHMMs to detect low sequence identity homologues among the new sequences. By selecting, for each domain family, those domains that have a sequence identity equal to or less than $p^I(L,0)$ compared to the saHMM-members, we obtain 451 low identity domain sequences belonging to families with an saHMM. Even though the sequence identity is very low, we find that 24.4% of the sequences match the correct saHMM with perfect accuracy (Table II).

Performance of saHMMs compared to Pfam HMMs

In the following, we compare the performance of the saHMMs based on SCOP 1.69 to the performance of the corresponding Pfam_ls HMMs, version 19.0. From all the domain sequences new to SCOP 1.71 we first select the 2454 domains that belong to families with both an saHMM and a Pfam HMM. When we then screen the corresponding full length sequences against the respective HMMs, we are able to detect the correct family relationships for 86% of the domains using the saHMMs, and for 75% of the domains using Pfam HMMs. For this comparison, we consider matches with *E*-values less than 10 and count only the top hits. It is of interest to note that 412 of the domains with correct hits to saHMMs fail to obtain a match to the correct Pfam HMM. Of these 412 hits, 213 can be counted as matches within the midnight zone since they have a sequence identity of at most $p^I(L,0)$ compared to the saHMM-members based on SCOP 1.69.

Using saHMMs to annotate unknown human proteins

Public databases contain thousands of protein sequences that are labelled “unknown”. In order to investigate the ability of the saHMMs to annotate “unknown” sequences, we searched the NCBI for human proteins labelled “unknown” and found 1986 such sequences (as of November 2007). Of these, 232 proteins can be matched to at least one of our saHMMs, resulting in 530 annotated nonoverlapping domains belonging to 102 different domain families (See Additional Supporting Information File 1 for a list of all matches). As before, the *E*-values were restricted to 0.01 and below. The classic Zinc-finger domain family (SCOP family g.37.1.1, sunid 57,668) receives with 83 hits by far the most matches, which were distributed over 20 individual proteins. Each of these 20 proteins received between one to 15 hits to the classic Zn-finger domain, and, in some cases additional hits to other domain families. For 17 of the Zn-finger proteins, the NCBI annotation is incomplete in the sense that none (13 proteins, e.g., AAY14760) or not all of the Zn-finger domains (e.g., AAX93276, where five out of seven were previously not assigned) are identified in the NCBI sequence entry.

Included in the list over families receiving many hits are such common domains as the EGF-type module with 68 hits and the LDL receptor-like module with 63 hits. Many domains identified by the saHMMs are involved in signalling, for example the protein kinase catalytic subunit with 17 hits, the SH3 domain obtaining 16 hits, the PDZ domain, nine hits, and the SH2 domain with five hits. (A summary of all matches is provided as Additional Supporting Information File 2). These protein domains are often involved in cancer and other common human diseases. Correctly identifying such domains is of considerable interest for medical, structural and pharmaceutical applications. Through the structural information associated with the match, it might be possible to use the saHMM-members as template structures to build comparative models, thus providing a starting point for further computational and experimental analysis such as mutagenesis studies, identification of active sites and interaction surfaces, and possibly for drug design.

CONCLUSIONS

In a fully automated approach we constructed a collection of 850 saHMMs. The main strength of the saHMMs lies in the fact that they are built from 3D-structure alignments of low identity homologous protein domains. The structure comparisons provide structure-anchored sequence alignments even in the case of very low mutual sequence identities. Since the proper multiple structure alignment method is crucial for the success of the saHMMs, we included, after careful evaluation, the program MUSTANG[28] into our automatic pipeline.

We would like to stress the fact that the saHMM method focuses on the family level. The task of placing a sequence into the correct family might look simple, however, considering a sequence which is distantly related to the rest of the sequences within the family, it is much harder to associate the unknown sequence with its correct family, than to place it into its correct superfamily. By restricting the mutual identities of the sequences selected as representatives to values equal to or below $p^I(L,0)$, we guarantee a high sequence diversity among the saHMM-members at the same time as we preserve the sequence characteristics that define the entire family. We are able to demonstrate that the saHMMs can, with high accuracy, identify sequences as belonging to the family they represent. The saHMMs are in fact so highly family specific that they are clearly able to distinguish between members of their own family and members from other families, even within the same superfamily and at low sequence identity. Furthermore, using the family level has the advantage that the structural and functional information is more specific.

In further evaluating the ability to recognize remote homologues and by comparison with other methods, we find that BLAST, which is the usual tool for family level detection, and the exo-saHMMs show similar performance up to about 20% coverage, after which the exo-saHMMs perform significantly better with respect to both EPQ and coverage. Compared to PSI-BLAST PSSMs and SUPERFAMILY HMMs, the exo-saHMMs achieve higher coverage at low EPQ values. We assume that the full saHMMs, harbouring the complete set of saHMM-members, will perform at least as well as the exo-saHMMs.

Extending the analysis onto the superfamily level leads to a drastic drop of the EPQ values. In particular, the EPQ value is zero for $e = 0.01$, which is the default *E*-value cutoff used for the FISH server.[40] With regard to the other methods, the exo-saHMMs perform remarkably well in this coverage interval, although we did not initially intend to use the saHMMs database for searches on the superfamily level.

Comparing saHMMs to the corresponding Pfam HMMs, shows that the structure-anchored HMMs outperform Pfam in assigning the correct family membership to new sequences. In addition, the saHMMs are able to identify family relationships that are not recognized by Pfam. In a search with human sequences labelled "unknown" against the saHMM database, we demonstrate that the saHMMs are able to identify domains for which there was no previous annotation. The examples demonstrate the strength of the saHMMs and their potential to complement existing annotation methods.

In summary, we are able to construct 850 saHMMs with which we cover 65% of the domain sequences and about 30% of the seven true class families in SCOP. Without doubt, this coverage is bound to improve due to the exponential increase of deposited structures in the PDB. As we add new domains to the midnight ASTRAL set, we will be able to increase the number of saHMM-members in existing families and include new families into the saHMMs database. The database is the foundation of a publicly available server called FISH, which stands for Family Identification with Structure-anchored HMMs,[40] and is accessible at http://babel.ucmp.umu.se/fish/.

ACKNOWLEDGMENTS

The authors thank Lixiao Wang and Marek Wilczynski for support with the FISH server and Åke Sandgren for assistance with parallel computing. Part of this research was conducted using the resources of the High Performance Computing Center North (HPC2N).

REFERENCES

1. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 2000;297:233–249.
2. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
3. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
4. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.

5. Doolittle RF. Of URFs and ORFs, a primer on how to analyze derived amino acid sequences. Sausalito, CA: University Science Books, December 1986. (ISBN-10: 0935702547, ISBN-13: 978-0935702545).
6. Mika S, Rost B. UniqueProt: creating representative protein sequence sets. Nucleic Acids Res 2003;31:3789–3791.
7. Watson HC, Kendrew JC. The amino-acid sequence of sperm whale myoglobin. Comparison between the amino-acid sequences of sperm whale myoglobin and of human hemoglobin. Nature 1961; 190:670–672.
8. Zuckerkandl E, Pauling L. Evolving genes and proteins. J Theor Biol 1965;8:357.
9. Rossmann MG, Argos P. A comparison of the heme binding pocket in globins and cytochrome b5. J Biol Chem 1975;250:7525–7532.
10. Flaherty KM, McKay DB, Kabsch W, Holmes KC. Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein. Proc Natl Acad Sci USA 1991;88:5041–5045.
11. Brenner SE. A tour of structural genomics. Nat Rev Genet 2001;2: 801–809.
12. Hargbo J, Elofsson A. Hidden Markov models that use predicted secondary structures for fold recognition. Proteins 1999;36:68–76.
13. Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
14. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. J Mol Biol 2001;307:721–735.
15. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310: 243–257.
16. Gnanasekaran TV, Peri S, Arockiasamy A, Krishnaswamy S. Profiles from structure based sequence alignment of porins can identify beta stranded integral membrane proteins. Bioinformatics 2000;16: 839–842.
17. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
18. Al-Lazikani B, Sheinerman FB, Honig B. Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. Proc Natl Acad Sci USA 2001;98:14796–14801.
19. Griffiths-Jones S, Bateman A. The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. Bioinformatics 2002;18:1243–1249.
20. Tångrot J, Kågstrom B, Sauer UH. Structure anchored HMMs (saHMMs) for sensitive sequence searches, Report UMINF-03.18, Department of computing science, Umeå University, Umeå, 2003.
21. Sillitoe I, Dibley M, Bray J, Addou S, Orengo C. Assessing strategies for improved superfamily recognition. Protein Sci 2005;14:1800–1810.
22. Casbon JA, Saqi MA. On single and multiple models of protein families for the detection of remote sequence relationships. BMC Bioinformatics 2006;7:48.
23. Scheeff ED, Bourne PE. Application of protein structure alignments to iterated hidden Markov model protocols for structure prediction. BMC Bioinformatics 2006;7:410.
24. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 2001;313:903–919.
25. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
26. Buchan DW, Shepherd AJ, Lee D, Pearl FM, Rison SC, Thornton JM, Orengo CA. Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. Genome Res 2002;12:503–514.
27. Tångrot JE, Wang L, Kågström B, Sauer UH. Design, construction and use of the FISH server. Lecture Notes in Computer Science, LNCS 4699, pp. 647–657. New York: Springer, 2007.
28. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: a multiple structural alignment algorithm. Proteins 2006;64:559–574.
29. Eddy S. HMMER 2.3.2 (3 Oct. 2003): Biosequence analysis using profile hidden Markov models. Available at: http://hmmer.janelia.org/
30. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000;28: 254–256.
31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
32. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
33. Brenner SE, Chothia C, Hubbard TJ. Population statistics of protein structures: lessons from structural classifications. Curr Opin Struct Biol 1997;7:369–376.
34. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561–577.
35. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput Chem 1996;20:25–33.
36. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
38. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
39. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34 (Database issue):D247–D251.
40. Tångrot J, Wang L, Kågström B, Sauer UH. FISH–family identification of sequence homologues using structure anchored hidden Markov models. Nucleic Acids Res 2006;34 (Web Server issue):W10–W14.