Assessing strategies for improved superfamily recognition

IAN SILLITOE, MARK DIBLEY, JAMES BRAY, SARAH ADDOU,
AND CHRISTINE ORENGO

Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology,
University College London, London WC1E 6BT, United Kingdom

(RECEIVED August 10, 2004; FINAL REVISION February 24, 2005; ACCEPTED February 25, 2005)

Abstract

There are more than 200 completed genomes and over 1 million nonredundant sequences in public repositories. Although the structural data are more sparse (~13,000 nonredundant structures solved to date), several powerful sequence-based methodologies now allow these structures to be mapped onto related regions in a significant proportion of genome sequences. We review a number of publicly available strategies for providing structural annotations for genome sequences, and we describe the protocol adopted to provide CATH structural annotations for completed genomes. In particular, we assess the performance of several sequence-based protocols employing Hidden Markov model (HMM) technologies for superfamily recognition, including a new approach (SAMOSA [sequence augmented models of structure alignments]) that exploits multiple structural alignments from the CATH domain structure database when building the models. Using a data set of remote homologs detected by structure comparison and manually validated in CATH, a single-seed HMM library was able to recognize 76% of the data set. Including the SAMOSA models in the HMM library showed little gain in homolog recognition, although a slight improvement in alignment quality was observed for very remote homologs. However, using an expanded 1D-HMM library, CATH-ISL increased the coverage to 86%. The single-seed HMM library has been used to annotate the protein sequences of 120 genomes from all three major kingdoms, allowing up to 70% of the genes or partial genes to be assigned to CATH superfamilies. It has also been used to recruit sequences from Swiss-Prot and TrEMBL into CATH domain superfamilies, expanding the CATH database eightfold.

Keywords: CATH; HMM; sequence profile benchmarking; intermediate sequence library; sequence alignments

Since the late 1980s numerous profile-based approaches have been designed to improve recognition of distant homologs (Barton and Sternberg 1987; Taylor 1987; Rice and Eisenberg 1997; Park et al. 1998; Kelley et al. 2000) (for reviews, see Eddy 1996; Madera and Gough 2002; Lee et al. 2003). In particular, it was observed that

Hidden Markov models (HMMs) are often more sensitive at recognizing very distant homologs (Madera and Gough 2002), probably because they employ more powerful statistical frameworks for handling the extensive insertions/deletions which can arise between distant homologs. The SUPERFAMILY database set up by Gough and coworkers (Madera et al. 2004) provides a library of sensitive HMMs built using the SAM-T technology of Karplus and coworkers (1998) for superfamilies in the SCOP structure database.

The relative performance of pairwise versus various profile-based sequence comparisons has been assessed by several groups (Park et al. 1998; Pearl et al. 2000).

Reprint requests to: Mark Dibley, Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, United Kingdom; e-mail: dibley@biochem.ucl.ac.uk; fax: +020-7679-7193.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.041056105>.

Results showed that for a given error rate, profiles built from the SCOP database recognized nearly twice the number of relationships than could be detected by simple pairwise methods searching the PDB alone. When considering more remote relationships, (<40% sequence identity) the performance of the profile-based methods was three times the performance of pairwise methods (Park et al. 1998). Similar results were obtained from benchmark studies exploiting profiles built from the CATH database (Salamov et al. 1999; Pearl et al. 2001; Buchan et al. 2002).

Of the profile-based methods, Park et al. (1998) showed that the SAM-T98 approach was able to identify the highest percentage of homologous relationships. More recent analysis by Gough and coworkers (Madera et al. 2004), again using the SCOP database, showed further increases in the performance of this HMM technology obtained by using a more recent version (SAM-T99) and by optimizing the parameters and manually adjusting bad models within the library. However, it is reasonable to suppose that the performance of profile-based approaches would also be dependent on the quality of the multiple sequence alignment used to generate the profile.

Ensuring the accuracy of an alignment becomes increasingly difficult as the sequences become more distant. One solution exploited in the past is to incorporate structural information. For example, Sternberg and coworkers (Kelley et al. 2000) exploited structural data in a novel fold recognition method. In 3D-PSSM, PSI-BLAST was used to provide sequence relatives for the structural sequences in a SCOP superfamily. The resulting sequence alignments were then combined in a pairwise manner using the SSAP algorithm (Taylor and Orengo 1989) to provide a reliable structural alignment between the parent structural sequences. The final multiple sequence alignment was converted to a position-specific score matrix (1D-PSSM) which encoded the sequence variation at each position of the alignment. Using this approach, the profiles generated by 3D-PSSM showed a 13% increase in sensitivity in homolog recognition over those generated with no explicit structural data (Kelley et al. 2000).

Conflicting results were obtained by Bateman and coworkers (2004), who also used structural alignments to generate improved HMMs for homolog detection (Griffiths-Jones and Bateman 2002). Their results suggested that incorporating structural data does not significantly increase the sensitivity of these probabilistic models. However, they were not explicitly attempting to recognize very remote homologs (<30% sequence identity) as in the work of Muller and coworkers (1999), and their benchmark data set was built from Pfam rather than SCOP, and so would be expected to contain a smaller proportion of these very remote homologs. Furthermore, at least two-thirds of the pairwise sequence relationships in the multiple

structure alignments they exploited shared >30% sequence identity.

However, it is clear that in some families, there can be considerable structural variation between distant relatives (<30% identity) (Orengo et al. 2001), and so improving the quality of the structural alignments in these families may help to increase performance of the HMMs. In order to get the best performance from structure-based HMMs, we developed a new strategy, SAMOSA (sequence augmented models of structure alignments), which aims to improve the sensitivity of HMMs by carefully combining structural data from the CATH database with sequence relatives. SAMOSA models have been built using the iterated model building protocol of SAM-T99 developed by the Karplus group (Karplus and Hu 2001).

An alternative strategy to using profile-based or HMM technologies is to use intermediate sequence search strategies (Park et al. 1997). These perform pairwise scans of sequences against protein family libraries, or intermediate sequence libraries (ISLs), rather than scanning databases of unclassified sequences. Many structure and sequence databases cluster sequences into families according to sequence, structural, and/or functional similarity, e.g., Pfam (Bateman et al. 2004), PRINTS (Attwood et al. 1997), CATH (Pearl et al. 2001), and SCOP (Lo Conte et al. 2000). An ISL can be constructed from any of these resources by extracting representative sequences from each family against which a query sequence can be scanned. Because pairwise sequence search methods are being employed, these techniques can sometimes be faster than the profile-based strategies, depending on the size of the library and the search method used.

Here we review several sequence-based strategies for recognizing distant homologs, which all exploit the SAM-T99 technology to build HMM models for representative sequences in the CATH domain structure database. Performance of the various protocols is compared with that of an intermediate sequence library (CATH-ISL) protocol that applies the pairwise BLAST algorithm (Altschul et al. 1990).

The performance of the models, for a reasonable error rate (<0.1%), ranges from 67% coverage when recognizing very remote homologs to 76% using a data set of less remote homologs. This contrasts with a coverage of 30% for the pairwise BLAST of CATH-ISL. No significant increase in performance was observed using an HMM model library that included the structure-based SAMOSA models. However, expanding the library of HMM models eightfold to include additional HMM models of CATH relatives in the sequence databases gave a 10% increase in coverage to 86%. It is prohibitively slow to scan large databases or hundreds

of completed genomes against this expanded library. However, we scanned selected genomes from each kingdom of life and observe increased coverage in the range of 4%–7% depending on the genome.

Results

Performance of the different homolog detection strategies was assessed by measuring the proportion of remote homologs from the CATH database that could be recognized, for a reasonable error rate ($<0.1\%$). These relationships had been validated by structure comparison and manual inspection. The most stringent data set, CATHremote, contained extremely remote homologs. The other sets, CATHfull and CATHsingle, contained less remote homologs, identified in various ways (see Materials and Methods, section II). The CATHremote data set was used to assess the benefits of incorporating structural data in the SAMOSA models for recognizing very distant homologs. None of the query sequences in the data sets had more than 35% sequence identity to any of the sequences used to generate the HMMs.

For each data set tested, query sequences in the data set were scanned against the HMMs, and the resulting matches were classified as true positives (same homologous superfamily) or false positives (different homologous superfamily or fold group) to generate coverage-versus-error plots as described in Materials and Methods. The error rate was measured as the percentage of false positives recorded for increasing E-values and therefore increasing coverage (see also Materials and Methods, section III). The motivation for assessing performance using a variety of benchmark data sets was to explore the dependence of the coverage obtained, for a given error rate, on the composition of the data set.

In the Results section, the five different sequence search protocols assessed are as follows: (1) pairwise scan using BLAST against a data set of nonredundant sequences from an ISL based on CATH (CATH-ISL), (2) scans against an HMM model library built from

representatives (S35 reps) of CATH structural families clustered at 35% sequence identity (1D-HMM-S35), (3) scans against an HMM model library built from representatives (S95 reps) of CATH structural families clustered at 95% (1D-HMM-S95), (4) scans against a combined HMM model library that included the 1D-HMM-S35 model library and also HMM models built from multiple structure alignments of structural subgroups in CATH superfamilies (3D-HMM), and (5) scans against an HMM model library built from CATH-ISL families that included related sequences from GenBank (1D-HMM-ISL). See Materials and Methods, section I for a detailed description of the protocols and data resources used.

Table 1 lists all of the search protocols tested together with the benchmarks data sets employed and the observed performances. Below we describe in more detail the results obtained for the various homolog recognition strategies tested.

1. Assessing the performance of pairwise vs. HMM-based homolog detection methods

The performance of pairwise and profile-based strategies was first assessed using the CATHfull benchmark data set. It can be seen from Figure 1 that scanning the query data set against the CATH 1D-HMM-S35 library gives a coverage of 76%. This compares with a coverage of 30% obtained by BLASTing the benchmark data set against the CATH intermediate sequence library, CATH-ISL.

The effect of sequence redundancy in the 1D-HMM-S35 model library was also investigated. It can be seen from Figure 1 that only a small increase in coverage (2%) was obtained by expanding the 1D-HMM-S35 model library to include models for all nonidentical representatives in CATH—the 1D-HMM-S95 model library. Since the 1D-HMM-S35 model library contains 44% fewer models than the 1D-HMM-S95 model library, this will give much faster database scans for equivalent performance.

Table 1. The three benchmarking sequence data sets, which version of CATH they are derived from, what sequences the data set contains, and the percentage of coverage obtained when used to benchmark the different search methods

Scan set	Version	No. of models	BENCHMARKING SET		
			CATHremote v1.7 < 25% seq. id 303 sequences	CATHfull v2.5 S35 reps 4036 sequences	CATHsingle v2.5 Superfamily reps 1467 sequences
1D-BLAST-ISL	v2.4	27,161	—	30%	—
1D-HMM-S35	v2.4	3,285	44.8%	76%	62%
1D-HMM-S95	v2.4	5,858	45.5%	78%	—
3D-HMM +					
1D-HMM-S35	v2.4	3,974	53.8%	77%	—
1D-HMM-S35	v2.5.1	4,023	—	83%	—
1D-HMM-ISL	v2.4	27,161	—	86%	—

Figure 1. Coverage-vs.-error plot for assessing the performance of the 1D-HMM model library compared to pairwise intermediate sequence search. The CATHfull data set was used for benchmarking. Scanning query sequences against the CATH-ISL using BLAST, solid line; 1D-HMM-S35 library built from CATH v2.4 S35 reps, dashed line; 1D-HMM-S95 model library built from CATH v2.4 CATH S95 reps, dotted line.

2. Assessing the effect of superfamily bias in the benchmark data set

To investigate the effect of superfamily bias in the benchmark data set on performance, a second query data set was tested. The CATHsingle data set contains only a single representative from each CATH superfamily. It can be seen from Figure 2 that the 1D-HMM-S35 library shows a lower performance in recognizing homologs when tested with the CATHsingle data set.

Some superfamilies in CATH are very highly represented. For example, the 76 most highly populated superfamilies in CATH (out of 1467 superfamilies) contain 50% of the sequence families despite representing only 8% of the superfamilies. The CATHfull data set contains a high proportion (83%) of query sequences from these highly populated superfamilies. Therefore the increased performance observed using this data set suggests that models generated for these more highly populated superfamilies are more sensitive in recognizing distant homologs within the benchmark data set. This is presumably because distant relatives in these superfamilies are more extensively represented in the GenBank NRDB100, giving rise to more informative models that better capture essential sequence features conserved throughout evolution.

3. Including HMMs built from multiple structure alignments in the model library

Although Griffiths-Jones and coworkers (2002) demonstrated that HMMs built from multiple structure alignments do not perform significantly better than HMMs

built from sequence-based alignments, they restricted their analysis to only those 348 superfamilies present in both HOMSTRAD (Mizuguchi et al. 1998) and SUPERFAMILY (Madera et al. 2004). This is only 62% of the total number of structural superfamilies for which HMMs could be built by exploiting multiple structure alignments from the CATH database. Furthermore, their benchmark data set was based on Pfam representatives and would therefore contain a smaller proportion of remote homologs (<30% sequence identity) than benchmark data sets selected using structural classifications (e.g., SCOP, CATH). Therefore, we decided to re-examine the impact of including structure-based HMMs taking these factors into account.

To assess the performance of the 1D-HMM-S35 library compared to the 3D-HMM library for detection of very remote homologs, we used the CATHremote data set, which contains a higher proportion of very remote homologs than the CATHfull or CATHsingle data sets (see Materials and Methods, section III).

Preliminary trials using models built on an earlier release of CATH (v1.7) had previously demonstrated that an ~10% increase in performance could be obtained by combining the 1D-HMM-S35 and 3D-HMM model libraries (Fig. 3), although the performance of the 3D-HMM library was below that of the 1D-HMM-S35 library (38.8% compared to 44.8%). The increase in performance obtained by combining the libraries (1D-HMM-S35 + 3D-HMM) was similar to the increased performance observed by Sternberg and coworkers (Kelley et al. 2000) and was due to the fact

Figure 2. Coverage- vs.-error plots contrasting the performance measured for the 1D-HMM model library built from CATH v2.4 using two different benchmark data sets. Scanning query sequences against the CATH-ISL by BLAST (CATHfull benchmark data set), solid line; performance of the 1D-HMM-S35 models (CATHsingle benchmark data set), dashed line; performance of the 1D-HMM-S35 models (CATHfull benchmark data set), dotted line.

Figure 3. Coverage- vs.-error plot comparing the performance of the combined 1D-HMM-S35 and 3D-HMM library with the 1D-HMM-S35 library. Performance of the 1D-HMM-S35 model library, solid line; performance of the 1D-HMM-S95 model library, dashed line; performance of the combined 1D-HMM-S35 and 3D-HMM model libraries, dotted line. All HMM model libraries were built using CATH v1.7. Performance was assessed using the CATHremote data set.

that the 3D-HMMs recognized more distant homologs at a cost of missing some of the closer homologs.

However, on reassessing the performance of the 1D-HMM-S35 + 3D-HMM model library constructed from a more recent version of CATH, it can be seen in Figure 4 that the increase in performance of the combined model library is reduced to 1%. This suggests that within three years the radius of “sequence space” recognized by the 1D-HMM models had expanded due to both the increase in the number of structural relatives in CATH and the expansion of sequence relatives in GenBank, on which the 1D-HMM-S35 library is built. Thus, there is now little gain in remote homolog recognition obtained by adding the 3D-HMMs to the library. However, scanning sequences from the 120 completed genomes against the SAMOSA 3D-HMM library does result in a slight increase in coverage (up to 3%; see below for details).

This observation of lack of any significant increase in coverage agrees with the previous findings reported by Griffiths-Jones and Bateman (2002). Although their approach did not attempt to expand the HMM models with sequence relatives as for the SAMOSA models (see Materials and Methods, section I), it is clear from our data that this expansion does not increase the performance significantly, either.

However, it can be seen from Figure 5 that there may be a very slight improvement in the accuracy of sequence alignments using the 3D-HMMs, compared to the 1D-HMM-S35s, for very remote homologs (<30% sequence identity) (see Materials and Methods, section III, for description of methods used to assess alignment

accuracy). For low sequence identity (10%–30%), the average increase in performance is 2.5 standard deviations from the mean accuracy achieved when not using structural alignments. In this range, the SAMOSA models provide twice as many improved alignments over 1D models. However, the *P*-value for a paired t-test is 0.14, which suggests that these results are not significant by this test. At lower levels of sequence identity (<10%), alignment accuracy is very poor (<20%) for both types of models.

4. Increasing the performance of homolog detection by HMMs by expanding the CATH HMM model library

The increase in performance of the 1D-HMM-S35 model library over the last two years, with increases in the CATH structure and GenBank sequence databases, is also demonstrated in Figure 6. This compares the coverage obtained by scanning the CATHfull data set against 1D-HMM-S35 models built from the most recent version of CATH (v2.5.1) and shows an increase in performance of 7% from 76%–83%. This suggests that particularly for those large, highly represented superfamilies in CATH discussed above, the 1D-HMMs are capturing the majority of distant homologs.

The performance of a very significantly expanded HMM model library built from the CATH intermediate sequence library (1D-HMM-ISL) was also assessed (see Materials and Methods, section I, for details on the generation of the 1D-HMM-ISL). In contrast to the CATH 1D-HMM-S35 model library, which contains 3285 models, the 1D-HMM-ISL model library contains

Figure 4. Exploring the cumulative effectiveness of scanning against a combined library of 1D-HMM-S35 and 3D-HMMs. All HMMs were built using CATH v2.4. Performance of all model libraries was assessed using the CATHremote data set. Performance of the 3D-HMM library, dashed line; that of the 1D-HMM-S35 library, solid line; that of the combined 1D-HMM-S35 and 3D-HMM libraries, dotted line.

Figure 5. Contrasting the accuracy of sequence alignments generated by aligning query sequences against 1D-HMM models or 3D-HMMs. Plot shows the average percentage alignment quality for 1D- and 3D-HMMs.

an additional 23,876 models built from additional sequence relatives for each CATH superfamily in GenBank. It can be seen from Figure 6 that there is a 10% increase in coverage for the expanded 1D-HMM-ISL library (86% coverage) compared to the corresponding 1D-HMM-S35 library (76% coverage).

However, the expanded library (1D-HMM-ISL) is approximately eightfold larger than the 1D-HMM-S35 model library, and this significantly increases the time required for scanning. Therefore, although the coverage is better the protocol can only realistically be implemented for small genomes or sequence data sets, unless significant computing resources are available.

5. Annotation of representative genomes using the combined 1D-HMM and 3D-HMM model libraries

The 1D-HMM-S35 library has been used to provide CATH domain annotations for 120 completed genomes, for the latest release of Gene3D (Buchan et al. 2003). Conservative thresholds on the E-values used in recognizing homologs were employed to give a low error rate (<0.1%). The coverage obtained for nine representative genomes, three for each kingdom, can be seen in Figure 7. The coverage for all 120 completed genomes ranges from 30% to 70% of genes annotated, depending on the organism. Including the 3D-HMM model library results in a 1%–3% increase in coverage, while including the expanded 1D-HMM-ISL model library provided a 4%–7% increase in coverage. All of the annotations can be viewed on the Gene3D Web site (http://www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D), and there are links to the CATH superfamily into which the gene region has been assigned. The coverage obtained for genome annotation using the

CATH-based HMM model libraries is comparable to that reported for other family-based annotation resources (e.g., SUPERFAMILY [Madera et al. 2004; <http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/>]; Pfam [Bateman et al. 2004; <http://www.sanger.ac.uk/Software/Pfam/>]).

Discussion

Scanning a query data set of structurally and manually validated remote homologs from the CATH database against a 1D-HMM-S35 model library containing representatives from each sequence family in CATH (v2.4) allowed recognition of 76% of the homologs. This performance increased to 83% using a more recent version of CATH (v2.5.1). These performances compare with a performance of 45% observed using an earlier version of CATH (v1.7) and suggest that the rapid expansions in the structure and sequence databases mean that the CATH HMM model library is now able to recognize the majority of distant homologs from the most highly represented families in CATH. Many of these superfamilies are also recurring frequently in the genomes (Ranea et al. 2004) and are thus highly populated in the sequence databases, too.

Interestingly, there was little difference in performance between the 1D-HMM library built from the 5858 S95 representative sequences and the library built from 3285 S35 representatives. This suggests that since there is no advantage in using the larger library, the smaller, and therefore faster, 1D-HMM-S35 library would be more appropriate to use for the rapid identification of homology in the classification procedure and

Figure 6. Increase in performance obtained by scanning an updated CATH 1D-HMM-S35 model library and an expanded CATH 1D-HMM-ISL model library. Performance of CATH 1D-HMM-S35 model library built from CATH v2.4, solid line; that of CATH 1D-HMM-S35 model library built from CATH v2.5.1, dashed line; that of expanded 1D-HMM-ISL model library built from CATH v2.4, dotted line.

Figure 7. The proportion of sequences from nine selected complete genomes (three from each kingdom) that can be assigned to CATH domain families by scanning the sequences against the 1D-HMM-S35 library are shown in gray. Additional matches recognized by scanning against the 3D-HMM library are shown in black; additional matches recognized by scanning against the 1D-HMM-ISL library, white. All HMM libraries were built using CATH v2.4.

when scanning large data sets of genome sequences against the models.

The combined 1D-HMM-S35 and 3D-HMM model libraries were expected to provide complimentary strengths in that the high coverage of the 1D-HMM-S35 model library would complement the sensitivity of the 3D-HMM model library. From the results, the combined model library provided only a 1.1% increase of coverage at an equivalent error rate. Although the 3D-HMM models were not able to enhance the recognition of additional remote homologs in the combined 1D-HMM-S35 + 3D-HMM model library, identification of coherent structural subgroups used to construct the SAMOSAs has allowed us to generate more accurate multiple structural alignments for highly populated superfamilies in the CATH database.

The multiple structural alignments have also been used to build consensus structural templates with improved sensitivity in recognizing structural homologs and fold similarities. These 3D templates will enable the CATH database to keep pace with the increasing numbers of protein structures being determined by the structure genomics initiatives. In turn, this expansion of the CATH structural library will further increase the performance of the CATH model library for genome annotation.

The 1D-HMM-S35 library was used to provide CATH domain structural annotations for 120 genomes in the Gene3D database (Buchan et al. 2003). These data are all publicly available via the Gene3D FTP site (Buchan et al. 2003). Since assignment of sequences to structural families and superfamilies allows more distant evolutionary relationships to be recognized, these data will be important in analyzing the evolution of protein

families and their patterns of recurrence in the genomes. The results can also be used to identify gene regions for which no structural representatives can currently be identified. These regions are suitable targets for the structure genome initiatives.

Materials and methods

GenBank NRDB

The GenBank NRDB100 is a translated, nonredundant FASTA version of the NIH genetic sequence database available from the NCBI Web site (<ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz>) (Benson et al. 2004). A new version of the GenBank NRDB100 was used for each CATH version HMM build.

CATH

CATH is a hierarchical classification of domain structures clustered into fold groups, evolutionary superfamilies, and sequence families (Orengo et al. 1997). Domains are recognized using several independent automated methods and validated by manual inspection (Orengo et al. 2003). Homologs are recognized using both sequence-based and structure-based comparisons. Pairwise sequence comparison, using a standard Needleman and Wunsch algorithm (HOMOL; Orengo et al. 1997), is applied to recognize close homologs. One-dimensional-profiles and HMMs are used to recognize more distant homologs (Pearl et al. 2002). Very remote evolutionary relationships and analogous structures are identified using several structure comparison methods (SSAP [Taylor and Orengo 1989], GRATH [Pearl et al. 2001; Harrison et al. 2003], CORA [Orengo 1999]). Manual validation is used to distinguish analogous and homologous relationships. CATH v2.4 contains 775 fold groups and 1386 homologous superfamilies. Within each superfamily, sequences are clustered into families of relatives at different levels of sequence identity (35%, 95%, 100%) from which representative sequences can be selected for building and benchmarking HMMs. In the present study, representatives from sequences clustered at 35% identity (S35 reps) and 95% identity (S95 reps) were selected. The numbers of domain sequence families (S35) in CATH v2.4 and v2.5.1, for which 1D-HMMs were built, are listed in Table 1.

CATH-ISL

The CATH-ISL was built using a library of CATH v2.4 representatives each having <35% sequence identity. These representatives were PSI-BLASTed against the GenBank NRDB100 containing 907,000 sequences that had regions of low complexity and transmembrane spans masked out using PFILT (Jones and Swindells 2002). The resulting matches were then analyzed using DomainFinder (Pearl et al. 2000) identifying 582,163 nonoverlapping hits of NRDB100 sequences to CATH domains. These matching sequences could be confidently assigned to a CATH superfamily.

For each CATH superfamily, the NRDB100 sequences, together with all of the CATH sequences for that superfamily, were then clustered into sequence families using a standard

Needleman and Wunsch algorithm (HOMOL, Orengo et al. 1997). Sequence families were generated at 35%, 95%, and 100% sequence identity. This resulted in an ISL containing 616,450 domains clustered into 1386 homologous superfamilies (as originally assigned in CATH v2.4), 26,812 S35, 80,849 S95, and 150,315 S100 families, respectively.

1. Homolog recognition methods

Below we describe five different protocols for scanning sequence databases or sequences from completed genomes, to recognize homologs of known protein superfamilies. Two HMM-based search methods were developed which exploit SAM-T99 technology to build HMMs with (3D-HMM) and without (1D-HMM) explicit structural data from the CATH database. We also investigated the increase in performance achieved for both pairwise and model-based protocols using intermediate sequence libraries.

1. Pairwise methods

The BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1990) presents a simple approach for the detection of distant but biologically sensitive relationships. This algorithm separates protein sequences into tripeptide fragments, expanded to include a series of closely related fragments, and scored using the BLOSUM62 substitution matrix. The query sequence is searched against the sequence database to identify fragments matching identically to the expanded list of tripeptide fragments and scored based on the probability that an equivalent matching fragment could have emerged by chance.

The CATH benchmark data set was BLASTed against the CATH-ISL S35 representatives using BLASTall with the *-P* option to detect multiple hits to the query sequence.

2. Hidden Markov models (1d-HMM) for CATH sequence families generated using SAM-t99

The SAM-T99 software was used to generate the 1D-HMM library using the default settings and a single structural sequence as a seed. Two versions of the 1D-HMM library were constructed using structural sequences from CATH clustered at 95% (1D-HMM-S95) and 35% (1D-HMM-S35) sequence identity (5858 and 3285 models, respectively, for CATH v2.4). An expanded 1D-HMM model library was also built based on representative sequences from each sequence family clustered at 35% identity (S35reps) from the CATH-ISL (1D-HMM-ISL, 27,161 models).

The alignments were built using target99 with an initial BLAST report of 100,000, recode3.20comp mixture, iteration E-value thresholds of 0.00001, 0.0001, 0.001, and 0.01, and the models were built using the w0.5 script.

A protocol was developed for identifying models that were performing badly. These gave low coverage for a given error rate because the model had been corrupted by the inclusion of non-homologous sequences incorrectly matched by the model. To detect when this was occurring, we ensured that the GenBank NRDB100 file used to build the models contained the sequences of known structures classified in CATH. This meant that the CATH superfamily identifier could be used to identify any non-homologs matching an HMM for a given CATH superfamily. When this occurred the model was rebuilt using stricter thresholds on homolog detection in the build phase, ensuring that performance increased as the thresholds were adjusted. Similarly, if the models failed at some stage to recognize the sequences used to seed the model, the model was rebuilt.

3. Hidden Markov models for structurally coherent families in CATH (3D-HMM) generated from multiple structure alignments (using CORA) and sequence alignments (using SAM-T99)

a. Identification of structurally coherent clusters within each CATH superfamily. In order to generate more accurate structural alignments in CATH superfamilies, we first investigated methods for identifying clusters of structurally similar relatives within the superfamily. Multiple alignments generated within these clusters, which we refer to as structural subgroups (SSGs) were then obtained using an established method (Orengo 1999).

The multiple structure alignment method, CORA, can be used to generate an alignment for each superfamily in the CATH database and to subsequently derive 3D templates from these alignments. The CORA templates have been shown to give increased sensitivity and selectivity in recognizing remote structural homologs (Orengo 1999). There are nearly 35,000 domain structures in CATH v2.4 and 1386 superfamilies. Although 90% of these superfamilies are relatively small, containing fewer than five sequence-diverse families (representatives clustered at 35% or more sequence identity), 10% of the remaining superfamilies contain 10 or more diverse families and account for 50% of nonidentical domains in CATH. In these superfamilies, the quality of the multiple structure alignments generated will be highly sensitive to the inclusion of very divergent structures with extensive structural embellishments.

Therefore, in order to improve the quality of multiple structure alignments, particularly for the most structurally variable superfamilies, we first investigated the effect of replacing a single multiple alignment for the superfamily with a set of alignments from SSGs within the superfamily. Multiple structure alignments were built for nonredundant representatives from each SSG, and 3D templates were then derived using CORA. The structural coherence of the cluster was assessed by measuring the sensitivity and selectivity of the 3D template in recognizing structural homologs from the superfamily.

A multiple-linkage algorithm was used to group the structures within each of the superfamilies into SSGs based on pairwise SSAP structure comparison scores. SSAP is a pairwise method of structural alignment that uses double dynamic programming to compare structural environments for pairs of relatives between two proteins (Taylor and Orengo 1989). The clustering algorithm selects the highest-resolution structure for each sequence family clustered at 35% identity (S35 family). Only one representative from each S35 family is selected to prevent the resulting multiple alignment and 3D template from being biased towards a particular highly populated S35 family within the SSG. Starting with the highest SSAP score, these representative structures were then clustered on the basis that a structure can only join a cluster if it has a structural similarity above a given threshold (T) to all the existing members of that cluster.

The value for this threshold score was optimized in order to select the most descriptive set of representative structures for inclusion in the structural templates. To do this, multiple structural templates were generated for four test superfamilies, from each structural class within CATH, using four different SSAP score threshold values (T = 70, 75, 80, 85). The structural templates at a given threshold were then scanned against all nonredundant structures (<35% sequence identity) in the CATH database, and the results were analyzed according to whether matches were to structures from the same homologous superfamily, fold group, or nonrelatives. The optimal

clustering threshold value provided the most discrimination between the structural similarity scores for homologous structures and nonrelated structures. An example of the results for the cytokine superfamily (CATH code 1.20.160.30) is illustrated in Figure 8. For all four test superfamilies, the most discriminatory templates were obtained using a SSAP score threshold of 80.

b. Building the 3D-HMM (SAMOSA) models. For each of the 689 SSGs, 3D-HMMs (SAMOSAs) were generated using the CORAXplode protocol. SAMOSA is an acronym for “sequence augmented models of structure alignments”, which refers to the fact that the alignment used to build the HMM is guided by a multiple structure alignment.

The CORAXplode protocol first takes a set of similar but nonredundant structures from a given SSG in a CATH superfamily, and a multiple structure alignment of these seed proteins is generated using the CORA algorithm (Orengo 1999). Sequence alignments for relatives of these seed structures are then generated by searching the translated GenBank-NRDB using the SAM-T99 protocol (Karplus and Hu 2001). The resulting sequence alignments for each seed protein are then condensed by ignoring any alignment positions that correspond to a gap in the seed structure. This step avoids the complication when combining the sequence alignments of attempting to align genomic sequences that could not be referenced back to positions in the original CORA structural alignment. The truncated sequence alignments are then combined by inserting gaps throughout the sequence alignment where gaps occurred in the CORA structural alignment. The

resulting alignment is converted to an HMM (3D-HMM, SAMOSA) using the SAM technology (Karplus and Hu 2001).

II. Benchmarking protocol for assessing the performance of the homolog recognition methods

Searching sequences against the HMM libraries

Matches obtained scanning the HMM library with query sequences were classed as true positives (TPs) if the homologous superfamily classification in CATH of the query sequences agreed with the CATH classification of the matched HMM. Matches were classed as false positives (FPs) if the classifications did not match. Homologous sequences that did not match the models were classified as false negatives (FNs), and all nonhomologous sequences that were not identified were seen as true negatives (TNs). If a query sequence matches more than one HMM from a given superfamily in the HMM library, then only the best scoring match (lowest E-value) is used. This “one-to-many” (Muller et al. 1999) approach of assigning a homologous relationship avoids artificially exaggerating the number of recognized homologs for each query sequence.

Matches between query sequences and HMMs that had the same fold classification in CATH were ignored and not counted as false positives. This is because structures are classified in CATH as homologs depending on whether they have significant structural similarity combined with significant sequence similarity or evidence of functional similarity.

Figure 8. Comparing the coverage-vs.-contact score plots for homologs (dotted lines), fold-relatives (dashed lines) and nonrelatives (solid lines) using multiple SSG templates from the cytokine superfamily (CATH 1.20.160.30), generated from four SSAP cluster cutoffs (70, 75, 80, 85).

Sometimes, for very distant homologs with similar 3D structures, no sequence or functional similarity is detected at the time of classification. These proteins are therefore classified in the same fold group in CATH but different superfamilies. However, since the HMMs used include further diverse sequences from all superfamilies present in GenBank NRDB100, these distant relationships may be detected by the models and a match is returned. Since these would need to be manually validated for classification in CATH, the simplest approach was to ignore these matches. This is also the approach adopted by other groups benchmarking the performance of similar sequence-based protocols (Muller et al. 1999).

a. Standard data sets for assessing the performance of 1D-HMMs vs. pairwise methods (CATHfull, CATHsingle). In order to assess the performance of various protocols for homolog recognition, it was necessary to use a data set of sequences that had no easily detectable sequence similarity to sequences used to generate the models or ISLs. Representatives were first selected from each S35 family in CATH (v2.5), giving a set of nonredundant representatives. This is referred to as the CATHfull data set. To test the effect of bias in the data set due to some CATH superfamilies being more sequence-diverse than others, another data set was generated which contained only a single representative from each superfamily. This is referred to as the CATHsingle data set.

When using these data sets to assess performance in detecting homologs, any matches between close homologs were ignored. In this instance, a close homolog is defined as a query sequence which shares more than 35% sequence identity to any of the sequences used in seeding the HMM or used in building the ISL library. This follows the protocol established by Chothia and coworkers (Park et al. 1998) for selecting benchmark data sets, except that their approach used a threshold of 40% to exclude close homologs.

b. Stringent data set for comparing the performance of the 1D and 3D-HMMs (cathremote). In order to compare the performance of the 1D and 3D HMMs, a further data set was generated containing the S35 reps from only those superfamilies that were diverse enough to be subclustered into SSGs. This was obtained by selecting S35 reps from the CATHfull data set for each CATH superfamily containing one or more SSGs. Those superfamilies containing only one SSG were only included in the data set provided there were additional S35 reps not clustered into the SSG that could be used to test the performance of the model built on the SSG.

In comparing the performance of the 3D-HMMs versus 1D-HMMs, any matches of query sequences to 1D-HMMs or 3D-HMMs from the same SSG as the query sequence were discarded. This is a more stringent criteria for excluding matches than that used for the CATHfull and CATHsingle data sets, as relatives from different SSGs typically have < 25% sequence identity between them.

III. Benchmarking protocol for assessing the quality of alignments to 1D and 3D-HMMs

In order to assess the quality of alignments of query sequences to the 1D- and 3D-HMMs, the sequence alignments generated were compared against structural alignments generated using the pairwise structure comparison SSAP algorithm (Taylor and Orengo 1989). Since structure is much more highly conserved than sequence during evolution (Lesk and Chothia 1980), structure

comparison gives more accurate alignments than sequence-based methods and can be used to benchmark the performance of the HMM alignments. For superfamilies with one or more SAMOSA models, sequence representatives were selected and scanned against both the SAMOSA library and the 1D-HMM libraries for their superfamily. A data set of 814 benchmarked alignments was obtained in this way, and used to assess the performance of the 1D-HMM and 3D-HMM libraries in generating sequence alignments, for a range of sequence identities.

The accuracy of the alignment between the query sequence and the target sequences contained in the HMM was assessed by simply calculating the number of aligned residues matching the aligned residues identified from a structural alignment of the query and target domains. This is expressed as a percentage of the total number of equivalent pairs identified by the structural alignment. The protocol used is similar to other benchmarking procedures (Elofsson 2002).

Acknowledgments

We acknowledge Frances Pearl for advice and comments on the benchmarking protocol and for managing the regular updates and curation of the CATH database. I.S. and M.D. acknowledge the Medical Research Council (MRC) for financial support as part of the e-Family project. J.B. acknowledges the NIH for funding as part of the Midwest Structural Genomics initiative. C.O. acknowledges the MRC for funding a Senior Non-Clinical Fellowship.

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.
- Attwood, T.K., Avison, H., Beck, M.E., Bewley, M., Bleasby, A.J., Brewster, F., Cooper, P., Degtyarenko, K., Geddes, A.J., Flower, D.R., et al. 1997. The PRINTS database of protein fingerprints: A novel information resource for computational molecular biology. *J. Chem. Inf. Comput. Sci.* **37**: 417–424.
- Barton, G. and Sternberg, M. 1987. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J. Mol. Biol.* **198**: 327–337.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. 2004. The Pfam protein families database. *Nucleic Acids Res.* **32**: D138–D141.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2004. GenBank: Update. *Nucleic Acids Res.* **32**: 23–26.
- Buchan, D., Shepherd, A., Lee, D., Pearl, F., Rison, S., Thornton, J., and Orengo, C. 2002. Gene3D: Structural assignment for whole genes and genomes using the CATH domain structure database. *Genome Res.* **12**: 503–514.
- Buchan, D.W., Rison, S.C., Bray, J.E., Lee, D., Pearl, F., Thornton, J.M., and Orengo, C.A. 2003. Gene3D: Structural assignments for the biologist and bioinformaticist alike. *Nucleic Acids Res.* **31**: 469–473.
- Eddy, S.R. 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**: 361–365.
- Elofsson, A. 2002. A study on protein sequence alignment quality. *Proteins* **46**: 330–339.
- Griffiths-Jones, S. and Bateman, A. 2002. The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. *Bioinformatics* **18**: 1243–1249.
- Harrison, A., Pearl, F., Sillitoe, I., Sidel, T., Mott, R., Thornton, J., and Orengo, C. 2003. Recognizing the fold of a protein structure. *Bioinformatics* **19**: 1748–1759.
- Jones, D.T. and Swindells, M.B. 2002. Getting the most from PSI-BLAST. *Trends Biochem. Sci.* **27**: 161–164.
- Karplus, K. and Hu, B. 2001. Evaluation of protein multiple alignments by

- SAM-T99 using the BALiBASE multiple alignment test set. *Bioinformatics* **17**: 713–720.
- Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**: 846–856.
- Kelley, L., MacCallum, R., and Sternberg, M. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**: 499–520.
- Lee, D., Grant, A., Buchan, D., and Orengo, C. 2003. A structural perspective on genome evolution. *Curr. Opin. Struct. Biol.* **13**: 259–369.
- Lesk, A.M. and Chothia, C. 1980. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**: 225–270.
- Lo Conte, L., Ailey, B., Hubbard, T., Brenner, S., Murzin, A., and Chothia, C. 2000. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **28**: 257–259.
- Madera, M. and Gough, J. 2002. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.* **30**: 4321–4328.
- 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.
- Mizuguchi, K., Deane, C.M., Blundell, T.L., and Overington, J.P. 1998. HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Sci.* **7**: 2469–2471.
- Muller, A., MacCallum, R., and Sternberg, M. 1999. Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.* **293**: 1257–1271.
- Orengo, C. 1999. CORA—Topological fingerprints for protein structural families. *Protein Sci.* **8**: 699–715.
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* **5**: 1093–1108.
- Orengo, C., Sillitoe, I., Reeves, G., and Pearl, F.M. 2001. Review: What can structural classification reveal about protein evolution? *J. Struct. Biol.* **134**: 145–165.
- Orengo, C.A., Pearl, F.M.G., and Thornton, J.M. 2003. The CATH domain structure database. In *Structural bioinformatics* (eds. P.E. Bourne and H. Weissig), pp. 239–248. Wiley-Liss, Inc., Hoboken, NJ.
- Park, J., Teichmann, S., Hubbard, T., and Chothia, C. 1997. Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* **273**: 349–354.
- Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**: 1201–1210.
- Pearl, F., Martin, N., Bray, J., Buchan, D., Harrison, A., Lee, D., Reeves, G., Shepherd, A., Sillitoe, I., Todd, A., et al. 2001. A rapid classification protocol for the CATH domain database to support structural genomics. *Nucleic Acids Res.* **29**: 223–227.
- Pearl, F.M.G., Lee, D., Bray, J.E., Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M., and Orengo, C.A. 2000. Assigning genomic sequences to CATH. *Nucleic Acids Res.* **28**: 277–282.
- Pearl, F., Lee, D., Bray, J., Buchan, D., Shepherd, A., and Orengo, C. 2002. The CATH extended protein family database: Providing structural annotations for genome sequences. *Protein Sci.* **11**: 233–244.
- Ranea, J.A., Buchan, D.W., Thornton, J.M., and Orengo, C.A. 2004. Evolution of protein families and bacterial genome size. *J. Mol. Biol.* **336**: 871–887.
- Rice, D.W. and Eisenberg, D. 1997. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* **267**: 1026–1038.
- Salamov, A., Suwal, M., Orengo, C.A., and Swindells, M.B. 1999. Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng.* **12**: 95–100.
- Taylor, W. 1987. Multiple sequence alignment by a pairwise algorithm. *Comput. Appl. Biosci.* **3**: 81–87.
- Taylor, W. and Orengo, C. 1989. Protein structure alignment. *J. Mol. Biol.* **208**: 1–22.