doi:10.1006/jmbi.2001.5080 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **313**, 903–919

JMB

Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure

Julian Gough[1]*, Kevin Karplus[2], Richard Hughey[2] and Cyrus Chothia[1]

[1]*MRC, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

[2]*Department of Computer Engineering, Jack Baskin School of Engineering, University of California, Santa Cruz CA 95064, USA*

**Corresponding author*

Of the sequence comparison methods, profile-based methods perform with greater selectively than those that use pairwise comparisons. Of the profile methods, hidden Markov models (HMMs) are apparently the best. The first part of this paper describes calculations that (i) improve the performance of HMMs and (ii) determine a good procedure for creating HMMs for sequences of proteins of known structure. For a family of related proteins, more homologues are detected using multiple models built from diverse single seed sequences than from one model built from a good alignment of those sequences. A new procedure is described for detecting and correcting those errors that arise at the model-building stage of the procedure. These two improvements greatly increase selectivity and coverage.

The second part of the paper describes the construction of a library of HMMs, called SUPERFAMILY, that represent essentially all proteins of known structure. The sequences of the domains in proteins of known structure, that have identities less than 95%, are used as seeds to build the models. Using the current data, this gives a library with 4894 models.

The third part of the paper describes the use of the SUPERFAMILY model library to annotate the sequences of over 50 genomes. The models match twice as many target sequences as are matched by pairwise sequence comparison methods. For each genome, close to half of the sequences are matched in all or in part and, overall, the matches cover 35% of eukaryotic genomes and 45% of bacterial genomes. On average roughly 15% of genome sequences are labelled as being hypothetical yet homologous to proteins of known structure. The annotations derived from these matches are available from a public web server at: http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY. This server also enables users to match their own sequences against the SUPERFAMILY model library.

Keywords: genome; superfamily; hidden Markov model; structure; homology

Introduction

Protein structure prediction, to discover the fold and hence information about the probable function of the sequence of a gene about which nothing is known, is possible *via* homology to a sequence of known structure. Protein homology searching methods have been the central tool in sequence analysis for many years, and with the growth in the experimental determination of new sequences following Moore's law (largely due to the genome projects), they are of increasing value. There are more than 50 completely sequenced genomes at the time of writing, including five eukaryotes. They comprise approximately a quarter of a million sequences. Improvements in the speed of homology methods enables larger-scale studies, and improvements in their selectivity leads to discovery of novel relationships. The system presented here probably offers the best selectivity currently available for genomic-scale studies.

Abbreviations used: HMM, hidden Markov model; FIM, free insertion module; PDB, Protein Data Bank.

E-mail address of the corresponding author: jgough@mrc-lmb.cam.ac.uk

0022-2836/01/040903–17 $35.00/0

Of the sequence comparison methods available, pairwise searches perform much less selectively than profile-based, of which hidden Markov models[1–4] (HMMs) are apparently the best. The work by Park *et al*.[5] which is the basis of this assertion, is supported by our more recent calculations.[6] The SAM HMM package (http://www.cse.ucsc.edu/research/compbio/sam.html) includes an iterative model building procedure called T99,[7] which improves remote homology detection. Of the HMM procedures available, SAM T99 is the most effective.

The implementation of HMMs is not entirely straightforward. In the first part of this paper we describe calculations that (i) improve the performance of SAM HMMs and (ii) determine a good procedure for creating SAM HMM for sequences of proteins of known structure.

In the second part of the paper we describe the construction of a library of HMMs called SUPERFAMILY, that represent essentially all proteins of known structure. This library of models extends both aspects of performance: speed and selectivity. A sequence can be searched against it at over an order of magnitude faster than building a model from a query sequence and searching an equivalent database of target sequences. Expert creation and selection of the models, coupled with consensus information, has greatly improved the selectivity of the library.

In the third part of the paper we describe the use of the SUPERFAMILY library to annotate the sequences of over 50 genomes. For each genome, matches are made to sequences that form roughly half the cytoplasmic proteins. These annotations are available from a public web server at: http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY/cgi-bin/gen_list.cgi.

This server also enables users to match their own sequences again the SUPERFAMILY model library.

Sequences, Domains and Homologies of the Proteins of Known Structure

Many small proteins contain single domains, whereas nearly all large proteins contain two or more that have linked by recombination and which occur in other proteins in isolation, in combination with different partners, or in both these states.[8–10] This means that to search for homologies in large sets of sequences, it is more effective to use HMMs that represent protein domains (whole small proteins or the evolutionary units of large proteins). Thus to build HMMs representing all proteins of known structure we must first have the sequences of representative sets of proteins, the definition of their domain structure and a description of their homologies. These data are available from the SCOP and ASTRAL databases.

SCOP database

This database[10] contains a structural and evolutionary classification of the proteins in the PDB[11] and usually keeps up to date to within two to six months. In SCOP, a multi-domain protein is split up into its constituent domains, which are then considered separately. These protein domains are evolutionary units in that for a protein to be divided into domains they must be observed in isolation or in different combinations in nature. The fundamental units used in the work described here are the protein domains as classified by SCOP. Note that a domain which has another domain inserted within it will comprise multiple chains or regions of sequence.

SCOP is a hierarchical classification, and the level of classification relevant to this work is the ''superfamily''. A superfamily is defined as a group of domains which have structural and functional evidence for their descent from a common evolutionary ancestor. The level below superfamily is the ''family'' level which groups together those domains that have clear sequence similarities. The level above superfamily is the ''fold'' which groups domains that have the same major secondary structure with the same chain topology. Superfamilies clustered at this level have either no evidence to suggest an evolutionary relationship or only very weak evidence that requires conformation.

The superfamily level contains the most distantly related domains and so is the highest level for useful remote homology detection. Proteins in the same superfamily often have the same function, and usually but not always have related functions. Since SCOP classifies protein domains separately, a multi-domain protein may have contributions to it's overall function from the different domains.

ASTRAL database

Files in this database[12] provide sequences corresponding to the SCOP domain definitions and are derived from the SEQRES entries in PDB files.

Domains which are non-contiguous in sequence, i.e. parts of the domain separated by the insertion of another domain, are treated as a whole. Their sequences are marked with separators between the fragments representing regions belonging to other domains. All PDB entries with sequences shorter than 20 residues, or with poor quality or no sequence information are omitted.

ASTRAL also has available sequence files filtered to different levels of residue identity. The ASTRAL entries are generated entirely automatically and hence have a small number of documented errors.

SUPERFAMILY database

The sequences which are used for the work presented here are available from the SUPERFAMILY server, described in a later section, and are generated from the ASTRAL sequences yet differ in the

following ways: (1) SUPERFAMILY sequence files have any sequence shorter than 30 residues removed rather than 20 in ASTRAL. Domains which are split across more than one chain had separate entries in ASTRAL which had to be joined to make a single entry in SUPERFAMILY. The ordering of the chains was obtained from NRDB and NRDB90.[13] Subsequent releases of ASTRAL however now include these joined domains. (2) A small number of documented ASTRAL errors which are significant are corrected by hand. (3) Some errors in domain definitions in the SCOP classification were detected and corrected in the SUPERFAMILY sequence files. These were mostly due to typographical mistakes and have been corrected in a subsequent release of SCOP. (4) Sequences which are merely redundant shorter parts of other sequences are removed when filtering on sequence identity.

The SAM-T99 HMM procedure

The SAM-T99 HMM procedure was developed by Kevin Karplus and his colleagues at Santa Cruz. Here we give an outline of the procedure; which is described in detail elsewhere.[7]

SAM-T99 starts from an initial alignment of homologous sequences or a single query sequence, this sequence or alignment is called the ''seed''. The default procedure follows these steps: (1) Using the initial sequence(s) and the WU-BLASTP procedure (http://blast.wustl.edu/blast/), a search of a large non-redundant protein database is carried out to find two sets of sequences. The first set consists of close homologues of the query sequence: those that match it with *E*-values of 0.00003 or less, and these are used to create the initial HMM. The second set of sequences consists of those that match the query with *E*-values of 500 or less and, therefore, will probably include more distant homologues of the query sequence. (2) An initial HMM is built from the first set of close homologues from step (1). This is then used to search the second set from step (1) for more homologues, which are added to the first set and realigned to create a new and larger alignment from which a new and better model can be built. (3) The initial HMM is extended with additional homologues by repeating step (2) for four iterations. In step (1), the use of low *E*-values by WU-BLAST and a strict HMM scoring threshold, ensures that only close homologues are used. In the four iterations of step (2), thresholds for the HMM score are gradually decreased. (4) From the final alignment produced by the iterations in step (3), a model is built using one of the scripts provided by the SAM package. These scripts filter and weight the sequences in the alignment before building the model.

The UCSC SAM package used to create the HMMs and to score sequences with the HMMs is available from http://www.cse.ucsc.edu/research/compbio/sam.

Performance of a model from an alignment of multiple homologous sequences *versus* multiple models from single homologous sequences

Previous practice by this group[14] and others[15] has been to build one HMM model for each protein family or superfamily using an accurate alignment of the sequences of diverse family members. However, there are two problems with this approach: one practical and one theoretical. The practical problem is that producing accurate sequence alignments is a not a trivial problem and for very diverse sequences requires expert human intervention. The theoretical problem is that it has not been demonstrated that using one model built from a good alignment of selected diverse sequences produces better results than using multiple models built from different single seed sequences and their homologues (as described above). To investigate this second problem various methods of modeling five chosen superfamilies were compared.

The five superfamilies were selected because detailed structural and sequence analyses were available, along with the expert knowledge acquired from these analyses (Bashford *et al.*[16] and our unpublished work). These provide both accurate structural-based hand-built sequence alignments, and the means with which to verify the results of homology searches.

The different models were built for each superfamily and searched against NRDB90, and then checked and compared. All of the models were built using the SAM T99 iterative procedure described above.

The following four variations on the method were investigated: (1) The accurate structure-based alignments of the superfamily members were used as the seed alignment for the default T99 procedure which generates the final models. (2) The accurate hand alignments were used as the seed, but an additional constraint was applied; the structurally conserved core regions of the alignment were fixed throughout the iterative procedure which would usually re-align at every step. (3) Completely automatic alignments of the same sequences used in (1) and (2) were created with ClustalW.[17] These automatic alignments (with many observed errors) were used as the T99 seed. (4) The individual sequence members of the above alignments were each used separately as seeds for a set of models. The results from all models were concatenated to give one result as in the other procedures.

The number of homologous sequences found by these four procedures is given in Table 1. The homology criteria were as follows: any hit with a ''reverse'' score lower (better) than −15 was taken as a homologue, hence the comparisons presented above depend on the scores produced by the different methods being roughly equivalent. This value (−15) was the score found to produce a 1%

Table 1. The numbers of hits found for different SCOP families using procedures 1-4 described in the text

	Total number of hits				
Superfamily	Cupredoxins	Cytochrome *c*	Flavodoxins	Globins	Iset
Procedure (1)	94	22	121	492	8440
Procedure (2)	63	21	121	489	8431
Procedure (3)	82	22	119	492	8298
Procedure (4)	106	22	130	505	9687

These data were obtained with SAM-T98 on SCOP release 1.39. The nrdb90 database from 1998 was used including sequences available at the time.

rate of errors per query when using T98 (a slightly older version of T99) to score all of PDB *versus* PDB.[5] The reverse score is offset against the score of the sequence in reverse, this filters out low-complexity matches.

The target sequences were checked by hand for false homologues using a combination of annotation, alignments, structural knowledge and further sequence searching. A small number of unannotated potential false homologues were found to be in the search results, and only a couple of certain false homologues. The immunoglobulins were not checked because the homologues were too many, and in the case of the flavodoxins the nitric oxide reductases were not counted as false because they are a well known case of sequence similarity between different proteins.[18]

The results indicate that the use of multiple models where each starts with a single sequence produces the best results. When sequence alignments are used the addition of the constraints described in (2) have little effect. The Cupredoxin hand alignment included only members of one of the three families within the superfamily and in this case procedure (2) did badly. The data also show that there is some loss in performance using automatic alignments for seeding the T99 procedure, but not a great deal.

Further analysis of these data shows that not only did the multiple models procedure (4) provide more hits but it found everything found by the others. The individual models in this procedure mostly found the same hits. Some models were completely redundant with respect to others and some found outliers which the others did not find (see below for a more detailed analysis of model redundancy).

These results solve the theoretical problem of whether one model or multiple models are most effective, and hence remove the need for solving the practical problem of accurately aligning distantly related sequences for the purpose of generating good hidden Markov models.

The SUPERFAMILY set of HMM Models

Seed sequences for the HMM library

Given that multiple models are to be built for each SCOP superfamily, the question remains; how should the models be generated? The model building procedure comparisons described show that it is best to create models for a superfamily starting with a set of single seed sequences. Here we use as seeds for the models sets of what we call ''SUPERFAMILY'' sequences: these are based on the sequences found for each SCOP superfamily in ASTRAL filtered to remove any that have identities greater than 95 %. Thus in the SUPERFAMILY model library each superfamily is represented by one or more models depending on how many structures there are with less than 95 % sequence identity in the given superfamily. Using the current data this produces a model library of 4849 models which is computationally viable both on a genomic scale and on the scale of a fast, single query (see below).

Studies on the effect of using models built from seeds filtered to different percentages of sequence identity showed that there is a strong fall off in the coverage achieved, beginning when using seeds filtered to 40 % identity, and falling steadily from 30 % and below (Figure 1). Since reducing the library from 95 % to 40 % only gives a reduction of one-third in the computational cost, and some performance is lost (e.g. ~50 assignments for an average-sized bacterial genome), the seeds filtered to 95 % are still used. In fact, the main advantage in using the larger library is an improvement in the quality of the assignments, rather than an increase in coverage. This is discussed below in the section on the assignment procedure.

Model building parameters

There are many parameters which it is possible to vary during the model building process. The effects of a basic set of five variations was chosen as a guide to the best method for building the final set of models. For each set of parameters a model library was built and scored against the sequences as described below. The parameters which were varied were: (1) the number of iterations in the T99 procedure. The default is 4 and up to 6 iterations were tried. (2) The cut-off *E*-values of the iterations in T99. At each iteration there is a cut off *E*-value used to choose new homologues to be included in the next model. (3) The limit on the score threshold, and the maximum number of sequences included in the large set of culled sequences used in the T99 procedure. (4) The final model-building

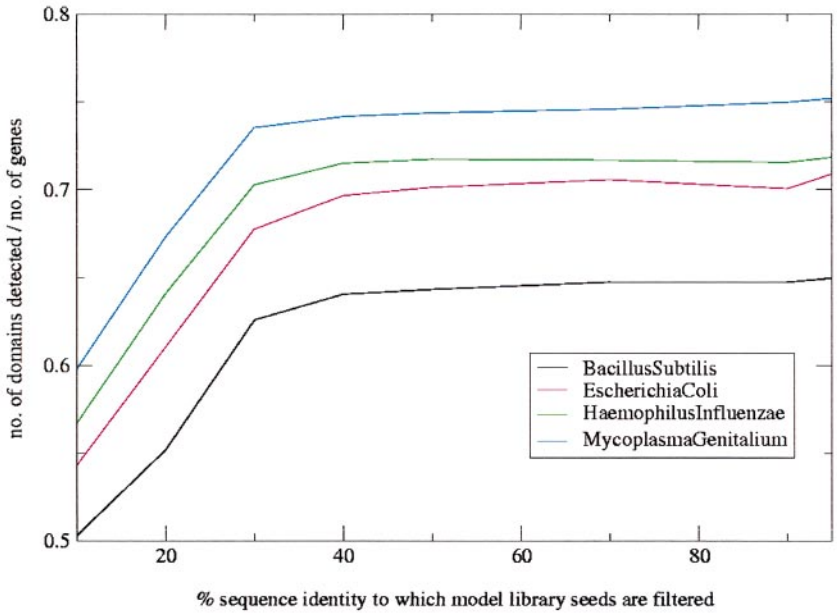

Figure 1. The effect of filtering the seed sequences on percentage sequence identity, is shown by counting the number of domains found in four different genomes. As the percentage sequence identity allowed between the seed sequences used to build the models increases, so more seeds, and hence models are allowed. There is a general trend that using more models finds more domains in the genomes.

script (mostly affecting the filtering of sequences in the final alignment). Again, the SAM package includes a selection of scripts to use. (5) The Dirichlet mixtures used in the regularizer for model-building[19] in T99. These encapsulate information about the nature of the residue distributions which are expected to be found in match states. The SAM package includes several different prior libraries of Dirichlet mixtures which can be used.

It was generally found that variations in (1), (2), and (3) produced mostly linear changes. It was possible to alter the parameters in a more ''loose'' direction improving the ratio of true to false homologues (Figure 2(a)), but also increasing the absolute error rate (Figure 2(b)). ''Tightening'' the parameters made it possible to reduce the number of badly built models, but also the coverage (Figure 2(c)). This can be attributed to the fact that looser parameters include more information in the models but increase the risk of including incorrect information, which leads to bad models. It was not only found that the absolute coverage is increased for looser models, which is dominated by a few large superfamilies, but that it increased the average coverage per superfamily (excluding singletons) in a similar way. To optimise the model library in its final form the loosest parameters were chosen to improve the coverage, and the models were analysed by hand to rebuild the bad models bringing the high error rate down. In fact, doing this to the library removes the bad models, thus producing entirely statistical (scoring) errors which match the theoretical *E*-value very closely, which is much lower than even the much tightest libraries (see below and Figure 3(b)). This procedure cannot be automated because of the complexity of the decision-making process involved in classifying inter-superfamily relationships as truly false. Some very distant inter-superfamily relationships are more acceptable than others depending on structural and functional similarities. Once the curation has been carried out on the model library the benefits are inherent within it, and carried forward to all future scoring.

Looking in more detail at the various parameters there are a few points worthy of note. The more iterations which are used, and the larger the culled set, the more computationally expensive the model building becomes. Once again; an expensive procedure is affordable if it is only to be carried out once, because the models can be used again and again at no further cost. The changes of (4) and (5) are of no extra cost. Using model-building script ''W0.5'' and prior library ''recode3.20.comp'' were found to be improvements on the defaults from work carried out at UCSC, which this work confirms. Tightening the threshold in the final iteration (also suggested by UCSC) reduces the number of bad models but does little or nothing to improve the ratio of true to false hits. For an automatic procedure this is very desirable, but in the case of the SUPERFAMILY model library which has bad models re-built by hand (see below), doing this reduces the ratio of true to false hits.

Free insertion modules

A Free Insertion Module (FIM) in a model allows the free insertion of any number of residues at that point in the sequence without penalty to the score. If you have a non-contiguous domain, it is possible to replace the inserted domain with a FIM in the model, thus giving the same score to a non-contiguous sequence with a domain inserted as a contiguous sequence. This amounts to removing any gap penalties at one point which might penalise an inserted domain where one is expected. The FIMs

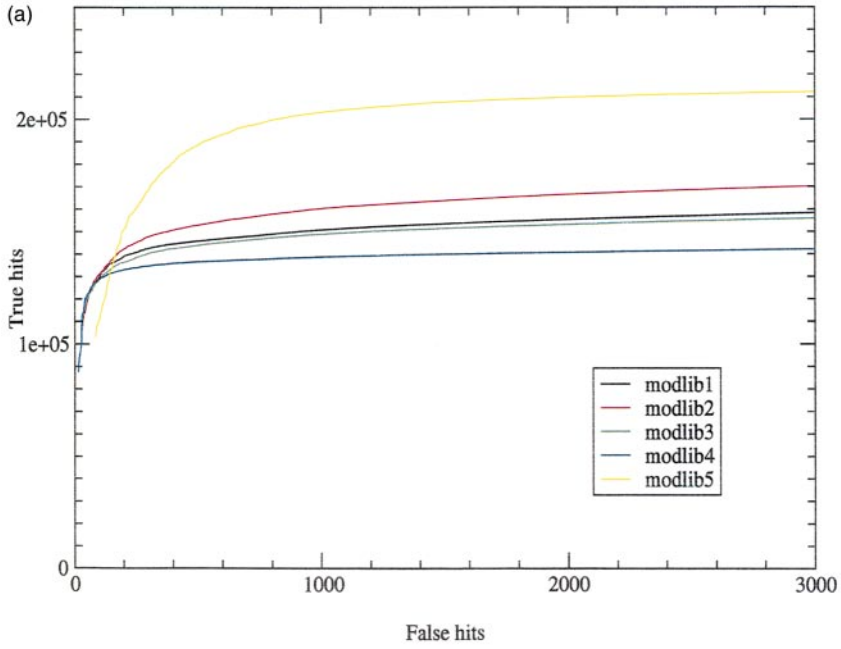

Figure 2 (*legend opposite*)

may be (i) inserted before the T99 process, or (ii) inserted in the final model after the T99 process, or (iii) not used at all.

Models were built for all non-contiguous domains in the three possible ways, scored against the PDB as before and compared. It was found that FIMs did not make much difference, but that using them ((i) and (ii) above) was slightly detrimental. The models without FIMs ((iii) above) were found to have an HMM segment with very high insert transition probabilities at the point of the inserted domain. In fact the T99 procedure successfully detects the point of insertion and allows large insertions at very low penalty without the need for intervention.

Curation of the model library

Once a library has been created it is tested. Its models are scored against every SUPERFAMILY sequence. The hits which the models make are then classified as true or false depending on whether or not they are classified by SCOP in the same superfamily.

It was discovered from this that false positives can arise for two reasons: (1) Errors in scoring. These are random scoring errors which occur because a sequence from another superfamily is assigned a score with a higher significance than it biologically deserves because of a chance similarity in sequence. Statistically this is expected to happen. The *E*-value is a prediction of how often this will happen by chance per query. (2) Errors in model-building. These are errors arising at the model-building stage. Such a problem will cause a model to consistently score certain false homologues with a significant score, because it is inherently built in to the model to do so. For example if members of other superfamilies somehow get into the alignment during the T99 procedure, then the model will also be built from these false sequences, causing it to match them with a significant score.

Models which have errors at the building stage ((2) above) can be recognised because their false homologues will all be related to each other, occur frequently and are assigned very significant scores. Scoring errors ((1) above) will inevitably produce a very few false homologues which are unrelated and will tend to have borderline scores. It is possible to plot the proportion of the models which make errors against different potential *E*-value thresholds. A change in behaviour (characterised by a jump discontinuity of the derivative) is observed at the point where the errors cease to be dominated by the model building errors, of which there will be many more with lower *E*-values than scoring errors.

Examination of the numbers of models involved in errors of both kinds, shows that a small percentage of models which have model-building errors produce most of the observed errors. By excluding less than 4 % of the models over 90 % of the errors are removed. Figure 3 compares a library with and without the 4 % of badly built models against a full library. This theory of two types of errors is confirmed by the errors of a library which has had badly built models removed. The error rate is then very much closer to the theoretical curve for statistical errors than that given by the full library (Figure 3).

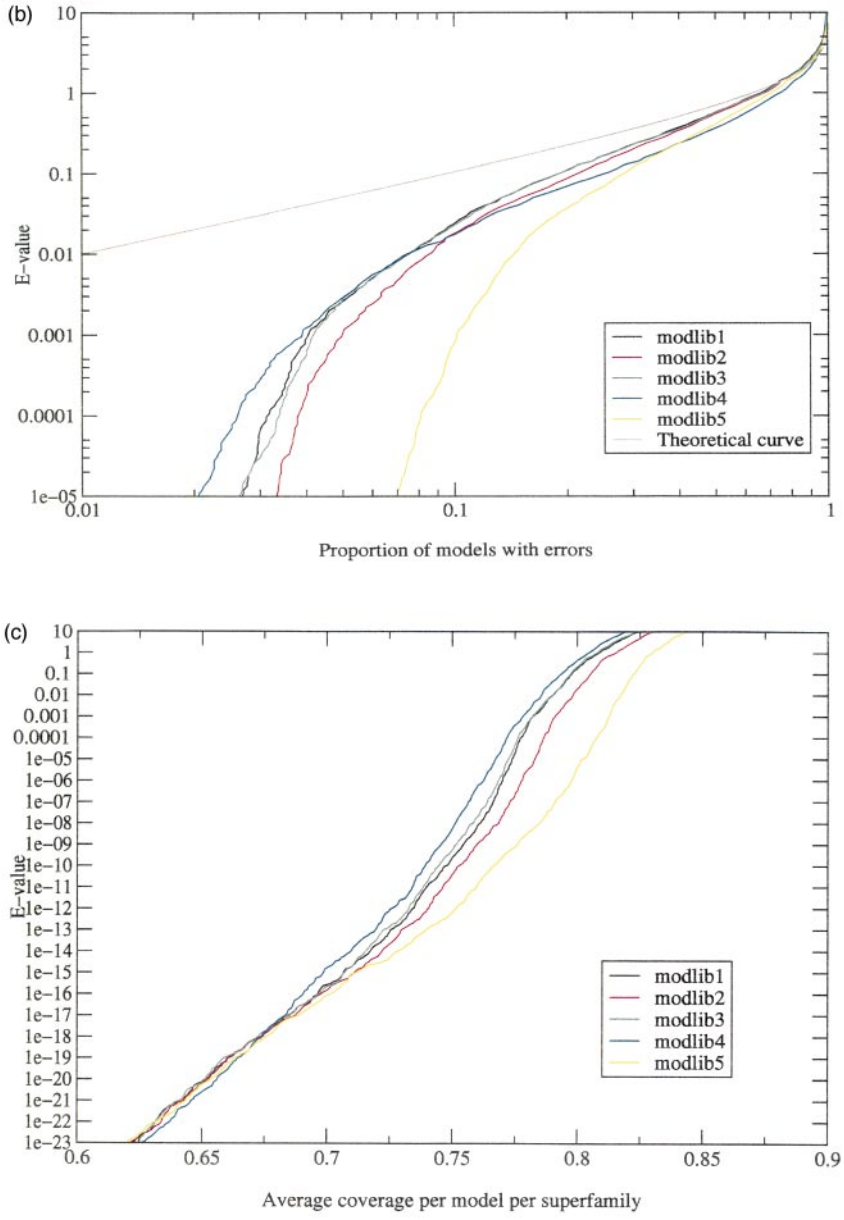

Figure 2. (a) The ability for models to discriminate between true and false positives is shown for five different model-building parameters in an all-against-all validation test of SCOP 1.53 sequences filtered to 95 % sequence identity. (b) The same validation test shows the proportion of models which produce errors at a given *E*-value threshold. The theoretical curve shows what would be unavoidably expected by chance. (c) The same validation test shows the average coverage per model per superfamily excluding singleton superfamilies at any given *E*-value. Singletons are excluded because they will have an average coverage of 100 % by merely finding themselves, and thus skew the distribution. In these Figures five different model building parameters were compared. Modlib1 refers to the default parameters with the exception of using an alternative Dirichlet mixture, and a very slightly lower *E*-value for the last iteration, and the W0.5 script. Modlib2 refers to using five iterations, and a slightly larger culled set. Modlib3 refers to the default SAM-T99 parameters with the W0.5 script. Modlib4 refers to the default parameters with no model-building script. Modlib5 refers to using six iterations, a much larger culled set, and higher *E*-value thresholds.

These excluded models are analysed by hand, comparing the structures to find the cause of their persistent errors. About half of these problem models appeared to be genuinely badly built, these are re-run with more restrictive parameters (see next section) and re-checked until they are behaving properly. In this way all of the superfamilies are properly represented. By doing this the final model library has 90 % of its false homologues removed and the error rate becomes very close to the theoretical value.

The other half of the problem models turned out to actually be behaving well and to involve either technical limitations in SCOP or the current lack of

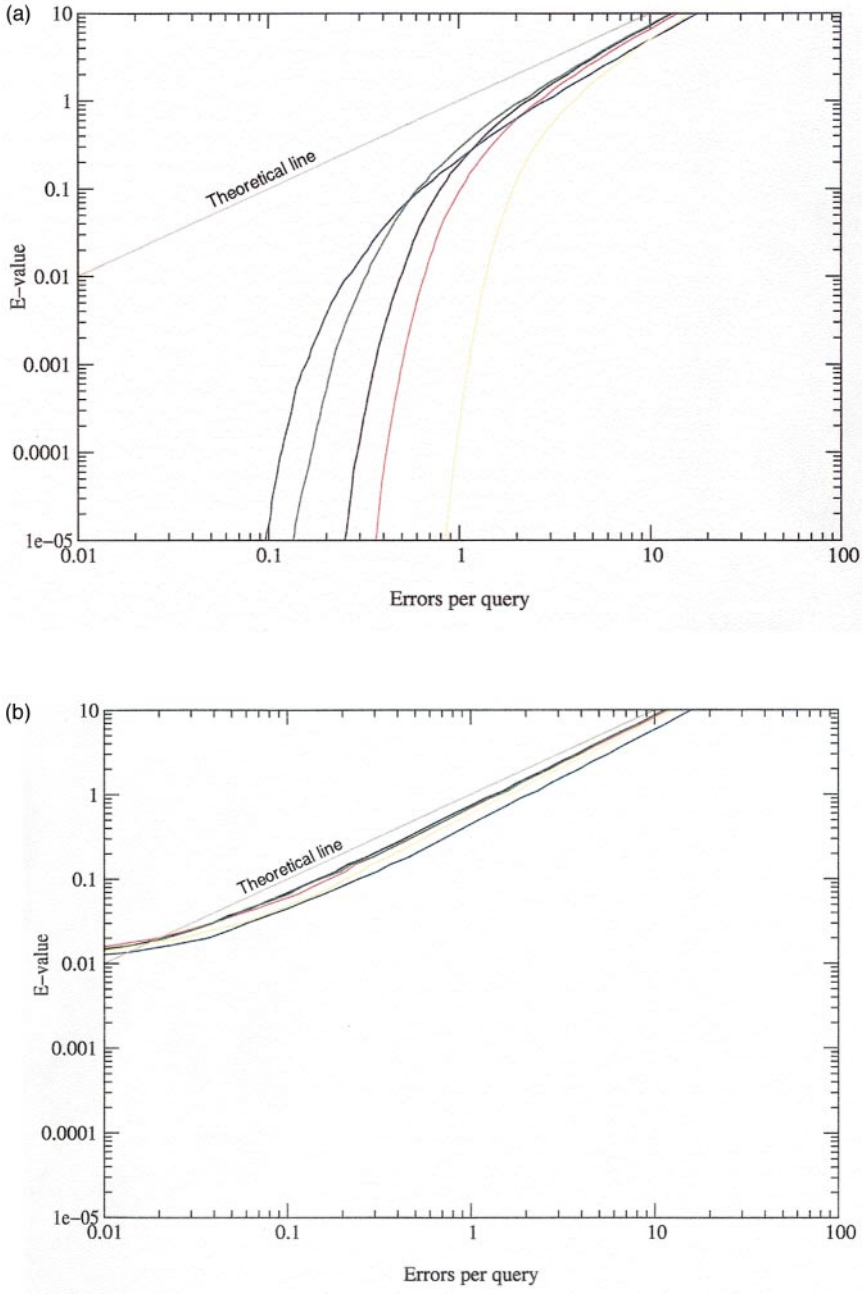

Figure 3. (a) The error rate produced by models built from five different model-building parameters in an all-against-all validation test of SCOP 1.53 sequences filtered to 95% sequence identity. The theoretical line shows the unavoidable error rate expected by chance alone. (b) The error rates of the five model-building parameters are shown here once all models with false hits below an *E*-value of 0.01 are removed. This demonstrates that a small number (~2%) of bad models are causing most of the errors.

structural data for plausible superfamily relationships: (1) The majority of these models detect inter-superfamily relationships between the few superfamilies which are in fact related, but classified separately for technical reasons. Most notable of these are the NAD(P) Rossmann domains and the FAD/NAD(P), and nucleotide binding Rossmann-like domains (see SCOP annotation). There were many relationships detected between these three superfamilies and a few others such as the N-terminal of MurD superfamily. (2) Members of SCOP superfamilies are classified according to whether, in the light of the currently known structural sequence and functional evidence, they possess a common evolutionary ancestor. Use of this evidence is conservative in that proteins, for which evidence of an evolutionary relationship is not strong, are placed in the same fold category but as separate superfamilies. This means that some models may well collect sequences that allow them to detect relationships between different superfamilies that go beyond that available from the current structural,

sequence, and functional evidence. This is particularly true for TIM barrel superfamilies.[5,20] Also 1c20 and 1bmy (ARID-like domains) have sequence homology but were classified separately in SCOP (Table 2). They have subsequently been re-classified.

(3) There are domains with local structural similarities which are detected by the models but the overall domain structure is different. 1b0p and 1fum (Figure 4) have an interesting identical discontinuous structural motif, but the rest of the domain does not even share the same secondary structure, one being all alpha and the other being mostly beta. 1psd (residues 108-295 on chain ''a'' form a Rossmann domain) and 1b6r (residues 1-78 chain a form a Biotin carboxylase N-terminal domain-like domain) superpose 35 residues to within 1.14 angstroms, but the rest of the structure does not superpose at all (Figure 5).

(4) Another common cause of these problems is due to multi-domain proteins which have similar (e.g. a long alpha-helical linker) or very variable sequence around the domain boundaries. An example is the glyceraldehyde-3-phosphate dehydrogenase-like two-domain proteins. As a result of this the models blur the domain boundaries and may detect part of the other domain. A related problem occurs when the domain boundary definitions vary slightly from one protein to another , in these cases they can be made consistent in the sequence files. 1qtm (residues 423-447) and 1kfs (residues 519-544) show an example of this; there is a six-turn alpha-helix which belongs to one domain in one definition, and the other domain in the other definition. These domain boundaries have subsequently been re-classified.

(5) A further cause is when two proteins have different domain boundaries which cause the parts to be classified differently. In the case of 1hfe (chain ''s''), 1feh (residues 520-574 chain a) which are Fe-only hydrogenases, a fragment is classified differently when it is a separate chain and when it is part of the main domain. In the case of 1qax (residues 4-108 chain a) and 1dqa (residues 462-586 chain a) a section which forms part of a domain is classified as part of another domain when the rest of its domain is not present. This has subsequently been re-classified.

(6) The last case is where there is a simple disagreement between the SCOP classification and the homology suggested by the scores. Mostly in these cases the homology suggested by the scores is simply wrong. The errors are scoring errors (natural statistical fluctuations) and are not caused by badly built models. For example 1sig (sigma70 subunit fragment from RNA polymerase) and 2fb4 (Immunoglobulin), have 16 residues which align almost identically yet the structures are unrelated. The section in question is β-sheet in one structure and helical in the other (Table 3). A rare example is of 1zrn and 1ek1 (HAD-like proteins) which match each other and superpose over 100 residues to within 1.655 angstroms r.m.s. deviation, indicating that they are clearly related structures (Figure 6), and have subsequently been re-classified in SCOP.

A further check was carried out by examination of the lengths of hits to unknown sequences. The library was scored against the *Escherichia coli* genome and the lengths of the hits were compared to the lengths of the models. It was found that most discrepancies in length were due to genuine insertions (sometimes of entire domains). Others were caused by matches to repeat domains, where there is a strong sequence similarity across several domains and the model matches parts from different domains. The EF hands were a common source of differences in length, as well as some circularly permuted sequences. As these cases were relatively few, and as there is no way to automatically separate genuine inserted domains from mis-matches, these models were left as they were. This does not greatly impair the homology detection of the model, merely its ability to distinguish the domain boundaries correctly.

Redundancy of models

Once the model library has been constructed it may be examined and filtered for redundancy. Since the redundancy of the seeds of the models is

Table 2. An example of an alignment of two unrelated sequences of known structure which produced a false hit

PDB	Alignment
1bcy	FLVALYKYMKERKTPIERIPYLGFKQINLWTMFQAAQKLGGYETITARRQWKHI
	FL L+ +M++R TPI R+P + ++L+ ++ GG + ++ W+ I
1c20	FLDDLFSFMQKRGTPINRLPIMAKSVLDLYELYNLVIARGGLVDVINKKLWQEI
1bcy	YDELGGNPGSTSAATCTRRHYERLILPYE
	L TSAA R Y + + PYE
1c20	IKGLHLPSSITSAAFTLRTQYMKYLYPYE

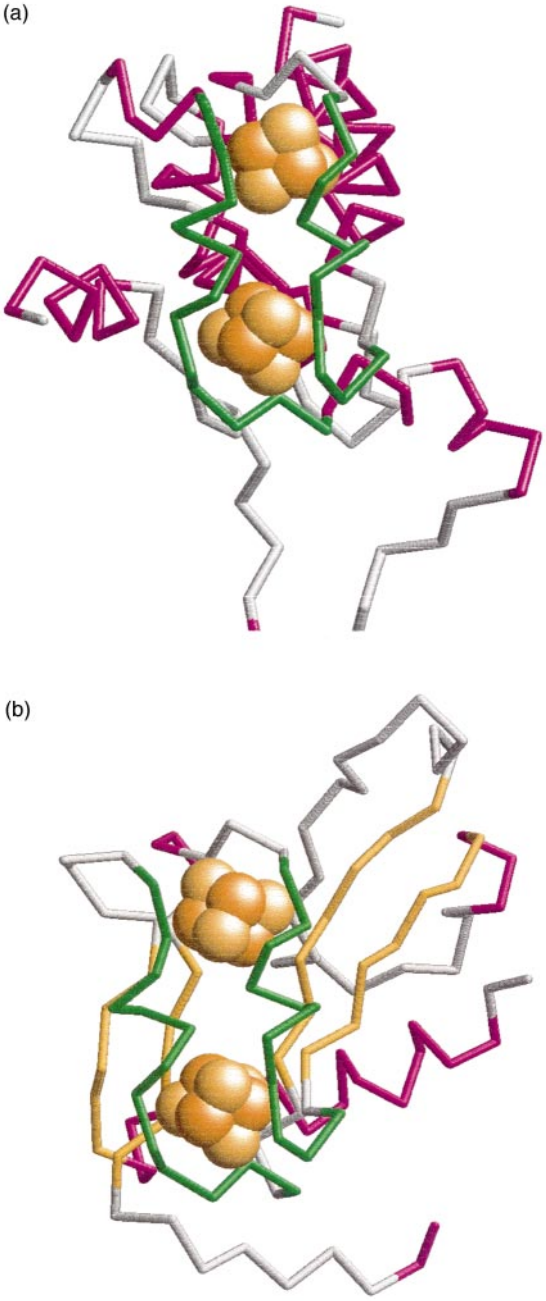

Figure 4. (a) The structure of a single domain of the multi-domain protein 1fum, which is classified in SCOP in the globin-like fold, and the superfamily is alpha-helical ferredoxin. This domain covers residues 106-243 of chain b. There is a structural motif (green) which consists of two helices which are sequentially joined by part of the supporting helical structure. (b) The structure of a single domain of the multi-domain protein 1b0p, which is classified in SCOP in the ferredoxin-like fold, and the 4Fe-4S ferredoxins superfamily. This domain covers residues 669-785 of chain a. In this case the same structural motif(green) as in (a) has the two helices sequentially joined by part of the supporting beta-sheet structure.

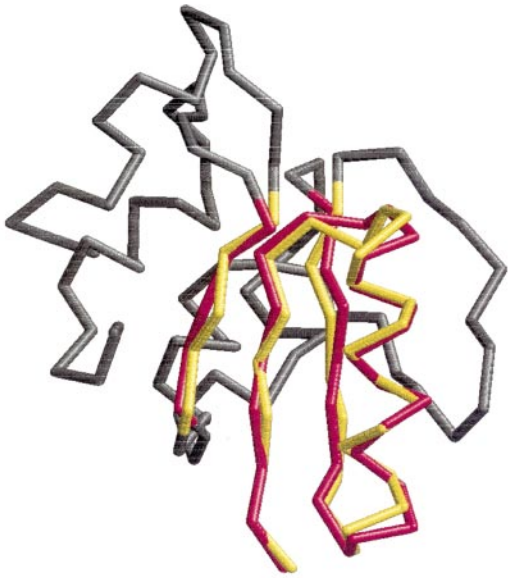

Figure 5. Shown here is the superposition of a part of two domains from the structures 1psd (residues 108-295 on chain a) and 1b6r (residues 1-78). The coloured region covers 35 residues which superpose with an r.m.s. deviation of 1.14 angstroms. These domains belong to two different superfamilies: Biotin carboxylase N-terminal domain-like, and NAD(P)-binding Rossmann-fold domains.

only indirectly linked to the percentage sequence identity of the models other measures of similarity must be introduced. Two measures of redundancy can be defined. These are based on the common percentage of: (1) the sequences of the multiple alignment from which the final models are built, (2) the sequences hit by the models from 50 complete genomes (see below).

For each superfamily, every model was compared to every other model in the same superfamily using the criteria defined above. Although no two models in the library have exactly the same state and transition probabilities, under the two criteria defined above 30 % of the models were found to be 100 % redundant on both counts. The covariance between the percentage sequence identity of the seed, model-building sequences, and genome sequence hits shows that there is no significant correlation of the sequence identity between model seeds and the sequence identity between genome hits made by the models ($r = 0.135$). What this means is that sequence identity of the seeds of models is a very poor measure of similarity. Only 58 % of models score their seed sequence higher than any other seed sequence, and only 33 % score their seed higher than any sequence in the completed genomes. This means that a model really does represent its superfamily first and foremost, and not its seed sequence as might be thought.

Table 3. An example of an excellent alignment of 16 residues, which can occur by chance in unrelated proteins

PDB	Alignment
1sig	GFIFSSYAMYWVRQA
	G+ FS+YA +W+RQA
2fb4	GYKFSTYATWWIRQA

Superfamily Assignment Procedure

Once the model library has been built it should be implemented in the best way possible. The aim is to identify the superfamilies of domains in sequences. Since most domains are part of multi-domain proteins, the domain boundaries must often be identified as well as their assignment to a superfamily. This can be particularly difficult in the case of non-contiguous domains.

For a given query sequence for which the domains and their superfamilies are not known, every model is scored (using local Viterbi scoring) across the whole sequence detecting any occurrences of a domain belonging to the superfamily which the model represents. For each region (domain) which is hit, the model which scores the highest has its superfamily assigned to that region. This reduces scoring (inevitable statistical) errors by nearly a factor of 2 since incorrect matches are only kept if they are not better matched by another superfamily (unpublished calculations). If the incorrect match is a member of a superfamily for which there is a structure, then the model(s) belonging to the correct superfamily will score higher thus knocking out the incorrect assignment. Using a model library built from seed sequences with less than 95 % identity rather than 40 % does not give a large improvement in the coverage, but does, *via* the assignment procedure, increase the quality of the assignments.

The greatest problem arises in the definition of the region which is matched. Often the hits will overlap, and when there is a domain inserted in another, the assignment will be completely within the other. Unfortunately using the domain definitions from the regions hit by the models alone did not prove adequate. The assignments to 20 genomes were analysed and a complicated procedure was developed to cope with such things as overlapping and inserted domains. This procedure was rejected because it allowed some false assignments to remain.

The solution is that every sequence region which is hit by a model is aligned to that model. For any given query sequence all alignments are compared to each other and the number of residues aligned to the same position is calculated. This number is declared as the overlap. This means that the number of residues overlapping between alignments is the sum of the residues with a match state in both models. For a full sequence the regions are assigned one by one, beginning with the highest scoring and adding each subsequent non-conflicting lower score in turn. A conflict is defined as an overlap of 20 % or more. A detailed study of non-contiguous assignments looking at all inserted domains, long sequence matches and overlapping assignments found only four errors out of 3041 assignments. The value of 20 % was suggested by studies on the initial assignment procedure which was ultimately rejected. This procedure adds a significant time to the overall procedure, but it is worth the cost and is not sufficient to threaten the viability of the whole procedure on a genomic scale.

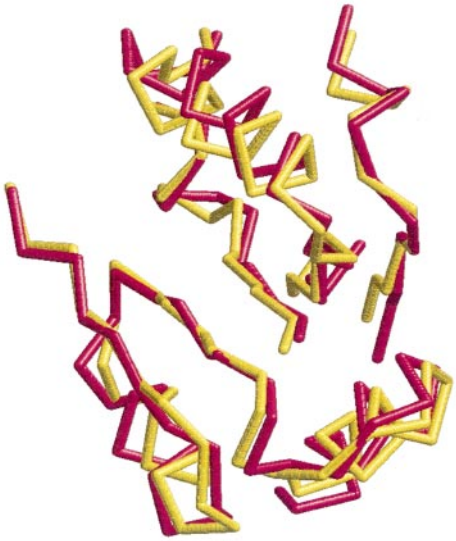

Figure 6. The superposition of PDB structures 1zrn and 1ek1 (residues 4-225 chain a), which are in different superfamilies in SCOP, yield 91 residues (shown) with 1.49 angstroms r.m.s. deviation from each other.

Assignment of superfamilies to genome sequences

A model library based on the 1.53 release of SCOP was used to carry out structural superfamily assignments of 56 complete genomes. The domains within sequences which hit with a significant score were uniquely assigned to a superfamily. The creation of the SUPERFAMILY structural assignment procedure and model library has allowed comprehensive descriptions of the domain genome sequences on a scale not previously possible (Table 4). The assignments cover roughly 45 % of prokaryotic and 35 % of eukaryotic genome peptide sequences.

No significant difference in the distribution of the scores given to each assignment was found between genomes. Differences were found how-

Table 4. Genome assignments using the SUPERFAMILY model library and procedure

A. *The extent of the SUPERFAMILY assignments to the sequences from 56 genomes*

Genome	ab	Number of genes	Matched sequences Number	%	Genome Coverage %	Number of Domains	Number of superfamilies
Eukaryotes							
Homo sapiens	hs	23867	11661	49	37	21201	595
Drosophila melanogaster	dm	14331	6851	48	34	11479	554
Caenorhabditis elegans	ce	19705	7851	40	29	12628	537
Arabidopsis thaliana	at	25470	13320	52	38	17957	564
Saccharomyces cerevisiae	sc	6297	2770	44	33	3760	461
Eubacteria							
Mesorhizobium loti	mk	6752	3552	53	44	4631	433
Pseudomonas aeruginosa	pa	5570	3079	55	45	4261	439
Escherichia coli o157	eo	5283	2502	47	41	3346	454
Escherichia coli	ec	4289	2292	53	45	3097	453
Mycobacterium tuberculosis CDC1551	mu	4187	1911	46	41	2594	391
Bacillus subtilis	bs	4100	2027	49	44	2754	417
Bacillus halodurans	bh	4066	2000	49	43	2688	415
Mycobacterium tuberculosis	mb	3918	1959	50	41	2650	392
Vibrio cholerae	vc	3835	1852	48	42	2527	424
Caulobacter crescentus	cc	3737	1997	53	46	2663	404
Clostridium acetobutylicum	ca	3672	1819	50	41	2382	401
Cyanobacterium synechocystis	cs	3169	1589	50	42	2164	379
Deinococcus radiodurans	dr	3102	1561	50	42	2007	379
Xylella fastidiosa	xf	2766	1097	40	41	1477	359
Stapylococcus aureus	sa	2594	1313	51	43	1728	368
Lactococcus lactis	ll	2266	1170	52	43	1514	334
Streptococcus pneumoniae	sr	2094	1044	50	43	1351	330
Neisseria meningitidis A	nn	2065	958	46	42	1266	342
Neisseria meningitidis	nm	2025	941	46	43	1264	342
Pasteurella multocida	pm	2014	1112	55	46	1467	359
Thermotoga maritima	tm	1846	1003	54	46	1335	343
Haemophilus influenzae	hi	1709	943	55	48	1243	341
Streptococcus pyogenes	sq	1696	887	52	44	1189	328
Campylobacter jejuni	cj	1634	845	52	43	1095	329
Mycobacterium leprae	ml	1605	844	53	48	1215	327
Helicobacter pylori	hp	1553	670	43	38	882	295
Aquifex aeolicus	aa	1522	902	59	49	1203	334
Helicobacter pylori J99	hq	1491	681	46	38	896	287
Chlamydophila pneumoniae AR39	cq	1110	443	40	36	625	243
Chlamydophila pneumoniae J138	cp	1070	446	42	36	628	243
Chlamydophila pneumoniae	cr	1052	443	42	36	625	242
Treponema pallidum	tp	1031	467	45	38	655	235
Chlamydia muridarum	cm	909	423	47	39	604	234
Chlamydia trachomatis	ct	894	419	47	40	597	235
Borrelia burgdorferi	bb	850	415	49	42	574	225
Rickettsia prowazekii	rp	834	437	52	44	605	248
Mycoplasma pulmonis	mq	782	363	46	34	485	186
Mycoplasma pneumoniae	mp	677	308	45	35	414	179
Ureaplasma urealyticum	uu	611	267	44	33	367	170
Buchnera sp.	bn	564	380	67	56	560	248
Mycoplasma genitalium	mg	480	261	54	41	362	172
Archaebacteria							
Sulfolobus solfataricus	ss	2977	1412	47	40	1790	323
Aeropyrum pernix	ap	2694	836	31	33	1067	289
Archaeoglobus fulgidus	af	2407	1238	51	45	1664	320
Pyrococcus horikoshii	ph	2064	904	44	40	1175	294
Halobacterium	hb	2058	1023	50	42	1351	306
Methanobacterium thermoautotrophicum	mt	1869	971	52	44	1297	307
Pyrococcus abyssi	pb	1765	957	54	45	1231	298
Methanococcus jannaschii	mj	1715	872	51	45	1132	288
Thermoplasma volcanium	tv	1499	795	53	45	1034	284
Thermoplasma acidophilum	ta	1478	795	54	45	1051	286

B. *The extent of the SUPERFAMILY assignments from 11 miscellaneous sequence sets including five alternative human gene sets and some incomplete genomes*

Genome	ab	Number of of genes	Matched sequences Number	%	Genome coverage %	Number of domains	Number of superfamilies
Softberry human gene predictions	hv	38170	15235	40	31	28223	613
Ensembl 0.8 human gene predictions	hx	29303	14437	49	39	25558	597
Ensembl 1.0 human gene predictions	hs	27615	13210	48	37	23402	595
Affymetrix human gene predictions	hu	21111	10339	49	37	19876	581
Known human genes	ht	8243	4995	61	41	9769	531
Mus musculus cDNAs	mm	21076	6223	30	29	8047	496
Mus musculus incomplete genome	mn	6978	3463	50	39	5599	391
Viridiplantae sequences from GenPept	sp	46369	31232	67	58	64711	546
Oryza sativa incomplete genome	os	2425	759	31	28	987	177
Guillardia theta nucleomorph genome	gt	485	203	42	33	261	92
Rhizobium plasmid	pn	417	202	48	40	250	77

A, The genome assignments for 56 genomes using the model library and assignment procedure. For each genome the table shows in order: the name of the species of the genome; a two-letter code (ab); the number of genes comprising the genome; the number of genes which have at least one SCOP domain assigned; the percentage of genes with at least one domain assigned; the percentage of the actual sequence covered by SCOP domains because multi-domain genes may have some domains assigned but not others; the total number of domains assigned; the total number (out of the possible 859) of superfamilies represented by at least one domain in the genome.

B, The assignments for 11 miscellaneous sequence sets including amongst other things five alternative human gene sets and some incomplete genomes. In A the current ensembl (version 1.1) is used for *Homo sapiens*.

ever between the distributions for each genome of the sequence identities between the matched genome sequences and the sequences used to seed the HMMs (Figure 7). For a few genomes, namely *Haemophilus influenzae* and *Buchnera*, a very high coverage of the genome was observed (the high number of *E. coli* genes with 100% identity is due to its over-representation in the PDB). This was accompanied by a markedly different sequence identity distribution across the genome. They have proportionally far more sequences with high identity to sequences of known structure than other genomes, which accounts for the high number of assignments.

As well as giving assignments that can be used in new investigations into genomes, they provide potential new annotations. An analysis of sequences previously unassigned showed that in

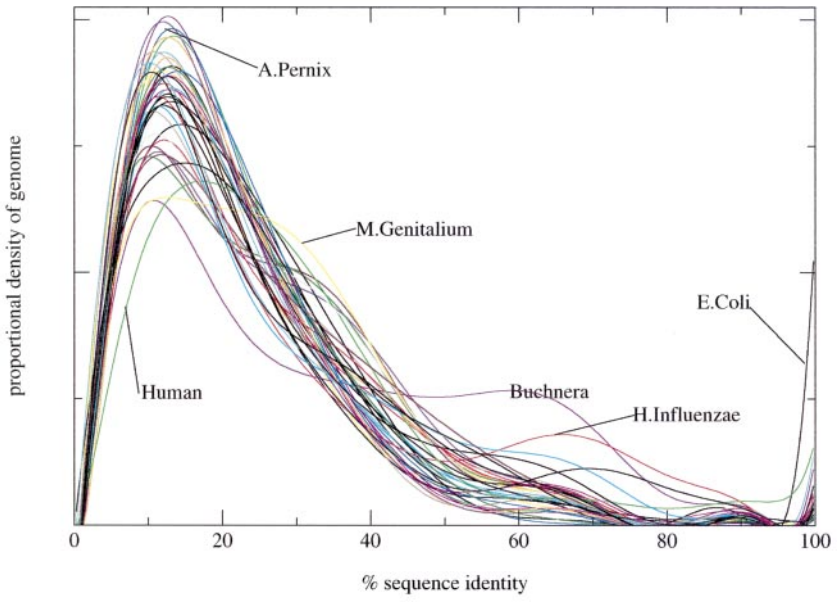

Figure 7. This Figure shows high order polynomial curves fitted to the number of domains found with a given percentage sequence identity to a PDB sequence for 38 genomes. The curves are normalised on the total number of domains found.

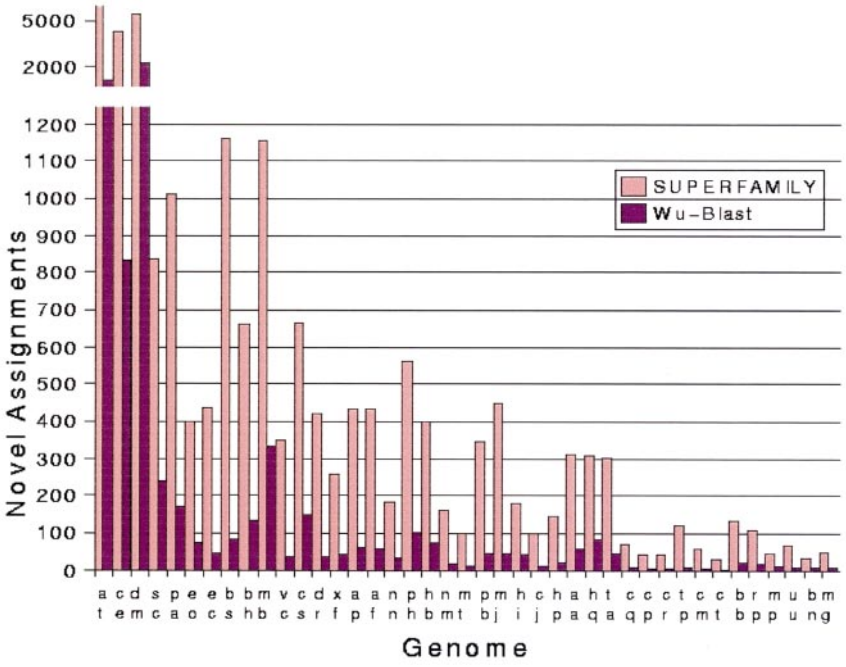

Figure 8. The genomes shown here were searched for entries with no annotation using keywords specific to each genome such as ''unknown'' and ''hypothetical protein''. Please see Table 4 for a key to the genome names. The number of SCOP domains found by SUPERFAMILY in sequences with no previous annotation is shown. Also shown are the number of previously unannotated sequences which have a hit to a SCOP domain with a *P*-value < 0.001 using Wu-BLAST. All genomes were provided with some novel annotation, and an average of 15% of the genes in a genome were provided with potential annotation where previously there was none. Please refer to Table 4 for a key to the genome names.

most genomes the SUPERFAMILY HMMs detect large numbers of novel assignments. The extent of this varies greatly with the quality of annotation carried out on the genomes by the sequencing projects (Figure 8). It is also worth noting that the novel assignments do not in general have marginal scores as one might expect; 63% of the novel assignments to eukaryotes scored with an *E*-value lower than 10^{-8} which is an excellent score.

A comparison with the homologies found for genome sequences by pairwise comparison and other methods

To assess the improvement in the detection of homology made by the SAM HMMs described here, we compared their performance with that of a pairwise sequence comparison method. For this comparison we used WU-BLAST because previous work has shown that this is one of the most effective of the pairwise methods.[21]

The 4849 SCOP sequences that were used to seed the SUPERFAMILY HMM models were directly matched by WU-BLAST to the sequences of the genomes in Table 4. For this calculation we used WU-BLAST version 2.0a19 with default parameters and matrix (BLOSSUM62). All those WU-BLAST matches with *E*-values of less than 0.001 were taken as significant. This *E*-value is expected to select matches whose significance is the same as the SAM HMM scores (see Brenner *et al*.[21]). The ratio of the number of sequences matched by the two methods is very similar for the different genomes: the WU-BLAST procedure makes matches to half the number of sequences that are matched by SAM HMMs.

On a more detailed level we compared the number of ''hypothetical'' sequences matched by the WU-BLAST and SAM HMM procedures. At this level WU-BLAST matched between 4% and 36% of those matched by the SAM HMMs (see Figure 8).

The coverage presented in Table 4 shows that 54% of the genes of *Mycoplasma genitalium* have a structural assignment. The current SUPERFAMILY HMM library which has been recently updated to release 1.55 of SCOP has structural assignments to 61% of the genes. It is possible to compare coverage for this genome with many other methods because it is very small (480 sequences) and so most methods have a comparable analysis. GenThreader[22] covers 53% (Jones, personal communication) of the genes, but is computationally costly and as a consequence has not been applied to many genomes. PSI-BLAST[23] has been used by several groups yielding coverages of 37% (Huynen *et al*.[24]), 39% (Wolf *et al*.[25]), and 41% (Teichmann *et al*.[26]). However, these figures were obtained using the fewer PDB sequences available at the time. Much more extensive use has been made recently by Gene3D (http://www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D) which covers 41% of *Mycoplasma genitalium* proteins, and has been applied to over 30 genomes including two of the smaller eukaryotes. None of these methods have been applied as extensively as the work presented here however, and most notably there is a lack of analysis of larger eukaryote genomes.

The SUPERFAMILY public server

As well as being used to produce genome assignments the library can be used to carry out

structural assignments for other sequences of interest. A public web server has been set up to serve the model library at http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY.

The library may be searched with many (amino acid or nucleotide) sequences at a time by entering them in FASTA format.[27] It includes a BLAST prefilter to remove the most obvious assignments without running the more costly HMM searches. The server will return all of the structural domains present in the sequences with the regions and links to SCOP. The service has already been extensively used, in particular by structural genomics groups to aid the selection of targets.

The server also makes the assignments to the genomes described here publicly available along with some analysis which has been carried out. The assignments can be browsed starting with either a superfamily or a genome, selecting the second from a list.

Also available on the server is an alignment procedure which allows multiple sequences to be aligned to the models in the library. Query sequences used in searches of the library may be aligned to the models, or sequences can be pasted in or uploaded in FASTA format. Genome sequences and PDB sequences are available on the server to include in the multiple alignments without the need to enter them.

It is the intention that the SUPERFAMILY models themselves will be made available to download. The library will be kept updated with every release of SCOP and is being used to provide a feedback loop for the testing and improvement of SCOP.

As the structural coverage of sequence space increases, especially from the work of structural genomics projects, the coverage of the library will increase as will the quality of the assignments.

Validation

The model library performs extremely well in validation tests against SCOP but, as it has been heavily trained on these validation tests themselves, the rigorous test is one of blind prediction such as the CASP[28] test. SUPERFAMILY was submitted to the LiveBench(http://bioinfo.pl/LiveBench) continuous bench marking of prediction servers, which is based on the CASP concept and offers difficult targets of recently solved structures. All targets have a BLAST[29] *E*-value higher than 0.1 to all other members of the PDB. Out of the 203 targets in the LiveBench-2 project (collected between 13.4.00 and 29.12.00), 45 were assigned to the correct SCOP superfamily by SUPERFAMILY and one falsely assigned. The one incorrect assignment involved a very short cysteine-rich protein and sequences of this kind are known to produce false matches with good scores.[5]

Discussion

The SUPERFAMILY HMM library

Here, we have described and assessed new procedures for the determination of homology by hidden Markov models; the construction of a library of models that represents almost all proteins of known structure, and the demonstration that sequences that comprise some 45% of bacterial genomes and 35% of eukaryotic genomes match these models

It has been established that by using multiple models each starting from a single seed sequence at least as many, if not more, homologues can be found than by using an accurate structurally based hand alignment to build a single model. This removes the need for accurate structural alignments for building a good model library. The principle does not extend to sequences of unknown structure because the SCOP classification, which is obtained by detailed structural analysis, is essential for determining the set of overlapping models which should represent a superfamily.

A model library has been made for all proteins of known structure. A superfamily is represented by many separate models each of which is attempting to model the whole superfamily. The model building procedure is very sensitive which is why a different seed sequence will produce a different model even though they are attempting to model the same superfamily. This is also the reason that the models are built from a 95% non-redundant set rather than a less redundant one. Thus, though the models will mostly hit the same sequences they do so with different scores for the common matches and some models uniquely match different outliers.

The two major factors which contribute to the performance of this library are (i) each superfamily is represented by many models rather than a single model and (ii) the ability to use SCOP/PDB sequence analysis to improve the models' performance. This second factor leads to the rejection of some models, the re-building of others and guidance in homology decision. Further to this, sequences queried against the library will be tested against every model so that the results are cross-compared, i.e. a sequence hitting one model may not be considered true if it hits another model with a higher significance. This is an extremely important contribution in genomic assignments, where regions will frequently hit many models especially within a single superfamily. The library leans heavily on the extensive and accurate information contained in the SCOP classification, ensuring that the assignments are as good as they can be given what is currently known.

Possible inter-superfamily relationships for which there is not yet structural evidence have been detected, most of which are structurally plausible.

Since the models for the library are created once and then re-used on a large scale, more effort can be spent on making the models as good as possible. The two areas in which the models can be improved at greater expense than would be practical for a single use are as follows: computationally expensive model-building parameters can be used on large computing resources over a long period of time, the models can be assessed and tuned using expert knowledge both in individual cases and across the whole library.

The models described here have been used for high-throughput genomic studies. The results of this work and the models themselves provide the basis of a public web service (http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY).

The genome sequences matched by the HMMs

The library has been used to assign structures to sequences from over 50 genomes with a coverage of roughly 35% of eukaryotic sequences and 45% of prokaryotic sequences. Membrane proteins are believed to form 20-30% of all sequences. There are only a few models for membrane sequences in SUPERFAMILY. This means that the SUPERFAMILIES HMMS match somewhat more that half of cytoplasmic proteins in bacteria and half, or a little less, of those in eukaryotes.

Since the library is based on known structures, as more novel folds are discovered through structural genomics projects, the number of matched genome sequences is expected to quickly rise.

The applications of this model library are several and it is already being used by many other projects within this laboratory and outside. The annotation provided by the genome assignments is already being used by many genome projects either as an aid to annotation or as annotation in its own right: for example those for mouse (http://www.gsc.riken.go.jp/e/FANTOM) and *Arabidopsis thaliana* (http://www.arabidopsis.org). The procedure has recently been added to the ENSEMBL (http://www.ensembl.org) pipeline for annotation of the human genome. The library is also being used by structural genomics projects to make predictions of the structure of their targets, and to suggest potential targets, e.g. SPiNE(http://spine.mbb.yale.edu/spine). The SUPERFAMILY web server is heavily accessed for genome assignments and sequence alignments. The server has already processed nearly 10,000 requests for sequence queries against the model library this year.

The genome projects themselves usually annotate the genes, but the annotation is just free text entries which are understandably very minimal and incomplete due to the volume of data involved (between 500 and 30,000 genes per genome), and the methods used. The free text is of little use for global studies of a genome because it requires human interpretation not possible on this scale. There are no standards of annotation between genomes, so inter-genome comparisons can be difficult, if not impossible. Using the model library to assign structural domains to all genome sequences provides both the necessary information about the genes, and the framework of classification (SCOP) for new comparative studies consistent across all genomes. Thus the SUPERFAMILY genome assignments have formed the basis of several comparative studies, for example on domain recombination[30] and on the evolution and formation of small molecule metabolic pathways.[31]

Acknowledgments

We thank Alexey Murzin for helpful discussions.

References

1. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.
2. Eddy, S. R. (1995). Multiple alignment using hidden markov models. *Proc. Third Int. Conf. Intelligent Systems for Molecular Biology* (Rawlings, C. *et al*., eds), pp. 114-120, AAAI Press, Menlo Park.
3. Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361-365.
4. Hughey, R. & Krogh, A. (1996). Hidden Markov models for sequence analysis: extension and analysis of the basic method. *CABIOS,* **12**, 95-107.
5. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210.
6. Madera, M. & Gough, J. (2001). A comparison of Hidden Markov Model software for remote homology detection. *Bioinformatics,* **in the press**.
7. Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics,* **14**, 846-856.
8. Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature,* **250**, 194-199.
9. Patthy, L. (1991). Introns and exons. *Curr. Opin. Struct. Biol.* **4**, 383-392.
10. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
11. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242.
12. Brenner, S. E., Koehl, P. & Levitt, M. (2000). The ASTRAL compendium for sequence and structure analysis. *Nucl. Acids Res.* **28**, 254-256.
13. Holm, L. & Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics,* **14**, 423-429.
14. Teichmann, S. A. & Chothia, C. (2000). Immunoglobulin superfamily proteins in *C. elegans*. *J. Mol. Biol.* **296**, 1371-1387.
15. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The pfam

protein families database. *Nucl. Acids Res.* **28**, 263-266.
16. Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold: unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199-216.
17. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.
18. Bredt, D. S., Hwang, P. M., Glatt, C. E., Lowenstein, C., Reed, R. R. & Snyder, S. H. (1991). Cloned and expressed nitric oxide synthase structurally resembles cytochrome P-450 reductase. *Nature,* **351**, 714-718.
19. Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, S. & Haussler, D. (1996). Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. *CABIOS,* **12**, 327-345.
20. Copley, R. R. & Bork, P. (2000). Homology among (βα) barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**, 627-640.
21. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA,* **95**, 6073-6078.
22. Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797-815.
23. Altschul, S. F., Madden, A. A., Schaffer, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
24. Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S., Yuan, Y. & Bork, P. (1998). Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* **280**, 323-326.
25. Wolf, Y. I., Brenner, S. E., Bash, P. A. & Koonin, E. V. (1999). Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9**, 17-26.
26. Teichmann, S. A., Park, J. & Chothia, C. (1998). Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA,* **95**, 14658-14663.
27. Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA,* **85**, 2444-2448.
28. Moult, J., Pederson, J. T., Judson, R. & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins:Struct. Funct. Genet.* **23**, ii-iv.
29. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
30. Apic, G., Gough, J. & Teichmann, S. A. (2001). Domain combinations in archael, eubacterial and eukaryotic proteins. *J. Mol. Biol.* **310**, 311-25.
31. Teichmann, S. A., Rison, S. C. G., Thornton, J. M., Riley, M., Gough, J. & Chothia, C. (2001). The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J. Mol. Biol.* **311**, 693-708.

Edited by G. Von Heijne

(Received 18 May 2001; received in revised form 12 September 2001; accepted 12 September 2001)