PROTEINS: Structure, Function, and Genetics 28:405–420 (1997)

Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments

Erik L.L. Sonnhammer,[1] Sean R. Eddy,[2] and Richard Durbin[1]*
[1]*Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom*
[2]*Department of Genetics, Washington University School of Medicine, St. Louis, Missouri*

ABSTRACT **Databases of multiple sequence alignments are a valuable aid to protein sequence classification and analysis. One of the main challenges when constructing such a database is to simultaneously satisfy the conflicting demands of completeness on the one hand and quality of alignment and domain definitions on the other. The latter properties are best dealt with by manual approaches, whereas completeness in practice is only amenable to automatic methods. Herein we present a database based on hidden Markov model profiles (HMMs), which combines high quality and completeness. Our database, *Pfam*, consists of parts A and B. *Pfam-A* is curated and contains well-characterized protein domain families with high quality alignments, which are maintained by using manually checked seed alignments and HMMs to find and align all members. *Pfam-B* contains sequence families that were generated automatically by applying the Domainer algorithm to cluster and align the remaining protein sequences after removal of Pfam-A domains. By using Pfam, a large number of previously unannotated proteins from the *Caenorhabditis elegans* genome project were classified. We have also identified many novel family memberships in known proteins, including new kazal, Fibronectin type III, and response regulator receiver domains. Pfam-A families have permanent accession numbers and form a library of HMMs available for searching and automatic annotation of new protein sequences. Proteins: 28:405–420, 1997.** © 1997 Wiley-Liss, Inc.

Key words: classification; clustering; protein domains; genome annotation; hidden Markov model; *Caenorhabditis elegans*

INTRODUCTION

Protein sequence databases such as Swissprot[1] and PIR[2] are becoming increasingly large and unmanageable, primarily as a result of the growing number of genome sequencing projects. However, many of the newly added proteins are new members of existing protein families. Typically, between 40% and 65% of the proteins found by genomic sequencing show significant sequence similarity to proteins with known function[3,4] and usually a large fraction of them show similarity with each other.[4,5] For classification of newly found proteins, and the orderly management of already known sequences, it would therefore be advantageous to organize known sequences in families and use multiple alignment-based approaches. This requires a system for maintaining a comprehensive set of protein clusters with multiple sequence alignments.

The problem breaks down into two parts: defining the clusters (i.e., a list of members for each family) and building multiple alignments of the members. Previous approaches to construct comprehensive family databases have either concentrated on aligning short conserved regions,[6–8] often starting from the manually constructed clusters in Prosite,[9] or full domain alignments using either clusters that were derived manually from PIR[2] or automatically.[10] An issue here is whether to aim for conserved regions only or whole domain alignments. By using short conserved motifs either in the form of a pattern or an alignment can indicate when a protein contains a known domain. Motif matches are often useful to indicate functional sites. However, they usually do not give a clear picture of the domain boundaries in the query sequence. They may also lack sensitivity when compared with whole domain approaches, because information in less conserved regions is ignored. The whole domain approach therefore seems preferable for detailed family-based sequence analysis because it offers the potential for the most sensitive and informative domain annotation.

To cope with the large number of families, the existing family databases made heavy use of automatic methods to construct the multiple alignments. Almost without exception, a manually constructed alignment would have been preferred but maintaining a comprehensive collection of hand-built alignments is not feasible. If the clustering is done at a high level of similarity, such as 50% identity, the

Contract grant sponsor: National Institutes of Health National Center for Human Genome Research; Contract grant number: HG01363
*Correspondence to: Dr. Richard Durbin, Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
Received 4 June 1996; Accepted 14 October 1996

© 1997 WILEY-LISS, INC.

alignment can be generated relatively reliably with automatic methods, but this will fragment true families and compromise the speed and sensitivity of searching. To avoid this, high quality alignments of large superfamilies are needed, which frequently require manual approaches.

Apart from the multiple alignment construction problem, a fully automatic approach also has to provide a clustering, and to work for multidomain proteins, define domain boundaries. For instance, the Domainer algorithm,[10] which performs the clustering of domain families based on all versus all Blastp matching, is a fully automatic approach that was used for building the ProDom database. We are most familiar with the Domainer method but believe that other automated sequence clustering approaches share similar drawbacks. The clustering level of Domainer depends on the score level of accepted pairwise Blastp matches. Domain borders are inferred by analyzing the extent of the BLAST matches and from NH_2- and COOH-terminal ends. The main problem with Domainer is that it does not scale well. As the sequence database grows, this will have several manifestations: 1) the computing time increases in the order of N^2, 2) either the clustering level must go up or the risk of false family fusions will increase, 3) the domain boundaries become less reliable due to more noise in the Blastp data, and 4) the quality of the alignment drops as more members are added. Further drawbacks of Domainer are that it is sensitive to incorrect data and that it is a one-off process that does not allow incremental updates but must be completely rerun at each source database update. This is not only very costly computationally, but also means that the families are volatile, due to the heuristic character of the algorithm, and cannot be permanently referenced from other databases. It is not well suited for classification because the families lack family level annotation.

Currently available fully automatic methods are thus not suitable for a high quality family-based classification system. Could a combination of manual and automatic approaches be a solution? The question here is really how much manual work has to be done to achieve a comprehensive database. This depends on the distribution of protein family sizes. Based on sequence similarity, it is clear that the universe of proteins is dominated by a relatively small number of common families.[11] The same type of analysis on the structural level reveals that there are a few families of very frequently occurring folds,[12] and it has been estimated that a third of all proteins adopts one of nine "superfolds."[13] This led us to believe that a semimanual approach initially applied to the largest families could capture a substantial fraction of all proteins. For practical reasons, however, it is usually not possible to build correct alignments solely based on the sequence data from members sharing a common fold because often there is essentially no sequence similarity at this level. The structural information required to produce a correct alignment is available only for a fraction of proteins. It therefore makes more sense to perform the clustering at the superfamily or family level, where common ancestry and sequence similarity are reasonably clear.

A major stumbling block of manual approaches is the problem of keeping the alignments up to date with new releases of protein sequences. A robust and efficient updating scheme is required to ensure stability of the database. These requirements are met in Pfam by using two alignments: a high quality **seed** alignment, which changes only little or not at all between releases, and a **full** alignment, which is built by automatically aligning all members to a hidden Markov model-based profile (HMM) derived from the seed alignment. The method that generates the best full alignment may vary slightly for different families, so the parameters used are stored for reproducibility. This split into seed/full is the main novelty of Pfam's approach. If a seed alignment is unable to produce an HMM that can find and properly align all members, it is improved and the gathering process is iterated until a satisfactory result is achieved.

The seed and full alignments, accompanied by annotation and cross-references to other family and structure databases and to the literature and the HMMs, are what make up Pfam-A. Each family has a permanent accession number and can thus be referenced from other databases. For release 1.0, we strived to include every family with more than 50 members in Pfam-A. All sequence domains not in Pfam-A were then clustered and aligned automatically by the Domainer program into Pfam-B. Together, Pfam-A and Pfam-B provide a complete clustering of all protein sequences. The quality of the Pfam-B alignments is generally not sufficient to construct useful HMMs. The main purposes of Pfam-B are instead to function as a repository of homology information and a buffer of yet uncharacterized protein families. As these families become larger they will benefit more from being incorporated into Pfam-A. Our goal is to progressively introduce the largest Pfam-B families into Pfam-A.

This study describes how Pfam was constructed and presents results from applying the Pfam HMM library to analyze protein families in Swissprot and to classify 4874 proteins found in 30 Mb of genomic DNA from *Caenorhabditis elegans*.

METHODS

Pfam-A

HMMs

HMMs have been used extensively both for the construction of Pfam and for detecting matches to Pfam families in database sequences. Although

HMMs are a general probabilistic modeling technique, we will use HMM in this study to mean a specific form of model that describes the sequence conservation in a family. This type of HMM consists of a linear chain of match, delete, and insert states.[14,15] The match state contains probabilities for amino acids in a given column, whereas the transition probabilities to and from insert and delete states reflect the propensity to insert a residue or skip one at a given position. The HMM parameters can either be estimated directly from a multiple alignment or iteratively by an expectation-maximization procedure from unaligned sequences. A protein sequence can be aligned to an HMM by using dynamic programming to find its most probable path through the states. The logarithm of this probability over the probability of a random model gives the score of the match, usually expressed in bits (logarithm base 2).

Score matrix-based profiles[16] are similar and might also have been used throughout. However, there are reasons to believe that HMMs are a somewhat superior approach to matrix-based profiles.[14] A practical reason for choosing HMMs was the suitability to the task of the HMMER package,[17] which includes the programs Hmmls for finding multiple nonoverlapping complete domains in a target sequence, and Hmmfs for finding multiple nonoverlapping partial and/or full domains.

Seed and full alignments

The philosophy behind Pfam-A is to construct a seed alignment for each family from a nonredundant representative set of full-length domain sequences trusted to belong to the family. The quality of each seed alignment was controlled by manual checking. From the seed alignment an HMM was built, which then was used to find new members and to generate the alignment of all detected members. The process of seed alignment and member gathering was iterated as outlined in Figure 1 if the initial seed was unsatisfactory. The HMMs were not built from the all-member alignment because this may contain incomplete or incorrect sequences that may affect the HMM adversely. The full alignments were never edited; if they were unacceptable, either the seed alignment was improved or the method to generate the full alignment from the seed was changed.

Seed alignment construction

The initial members of a seed were collected from one of several sources: Swissprot, Prosite, structural alignments,[18] ProDom [10], BLAST results, repeats found by Dotter,[19] or published alignments. Families were chosen on an ad hoc basis, with a bias toward families with many members. If the source provided a complete alignment of the seed members, this was used, but usually an alignment had to be built and compared with known salient features such as active site residues or structurally important residues. Of

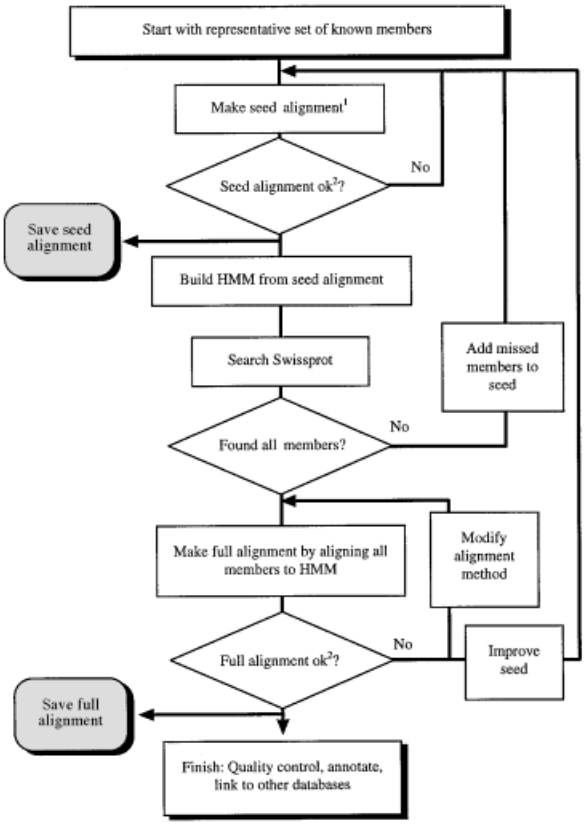


Fig. 1. The procedure to construct the alignments and HMM for a Pfam-A family. [1]Initial seed alignments are taken either from a published alignment or are made by one of the methods described in the text. [2]By 'ok' we mean that known conserved features are correctly aligned and that the overall alignment has sufficiently high information content to separate known positives from negatives.

the automated alignment methods used (Clustalw,[20] Clustalv,[21] HMM training[22]), Clustalw most often produced the best alignment. In a few cases manual editing of the seed alignment was necessary. Any sequence that was suspected to contain an error such as truncation, frameshift, or incorrect splicing was not included in the seed alignment to avoid adding noise to the HMM. This is important because up to 5% of the sequences in Swissprot may contain such errors (T. Gibson, personal communication).

HMM construction

From each seed alignment an HMM was built by using the Hmmb program. Although care was taken to ensure that the seed members did not include very similar sequences, one of two different weighting schemes[23,24] was applied to minimize any potential bias toward a subgroup.

To avoid overfitting and to make the HMM more general, amino acid frequency priors were normally derived according to an ad hoc pseudocount[25] method using the BLOSUM62 substitution matrix. How-

ever, for some families (e.g., EGF, EF-hand, globin, ig) the less specific Laplace ("plus one") priors gave better results and were therefore used.

Full alignment construction

Each HMM thus constructed was then compared with all sequences in Swissprot. This was either done directly with the search programs Hmmls or Hmmfs, or by converting the HMM to a GCG profile[26] to be able to use the very fast Bioccellerator hardware from Compugen.[27] These programs all perform variants of dynamic programming: the programs bic_profilesearch on the Bioccellerator and Hmmfs use a fully local algorithm, whereas Hmmls is local in the query sequence but matches the entire HMM. A further difference is that bic_profilesearch only reports the highest score, whereas Hmmls and Hmmfs report all scores above a threshold with coordinates. Although the Bioccellerator is ~50 times faster than a workstation, the result has to be postprocessed with Hmmfs or Hmmls to extract the coordinates of all matches. This was done by retrieving the entire sequence of all proteins that match according to bic_profilesearch with the Efetch program[28] into a minidatabase, which was then searched with Hmmfs or Hmmls.

If a list of known members of a family was available, the search result was compared with it to make sure that no known members were missed inadvertently. If the seed alignment is very small, one cannot expect to find all members at once. In such cases, selected newly found members were incorporated in a new seed alignment and the search was iterated. For the families where the initial seed alignment was derived from structural superpositions, the new HMM was constructed with a modified training algorithm that constrains the known structural alignment, allowing only the sequences of unknown structure to be realigned.

By extracting all matching sequence fragments and aligning them to the HMM with the program Hmma, a full alignment is created. Depending on the nature of the family, either Hmmfs or Hmmls will give more accurate matching segments. Hmmfs occasionally breaks a domain artificially into two or more fragments if unexpectedly large insertions or gaps are encountered. Hmmls does not do this, but may penalize partial matches (to fragments) so much that they are not found at all. Usually Hmmfs is used, but in some cases Hmmls was preferred. The method used for constructing the full alignment and the score cutoffs used were recorded for each family. The default score cutoff was 20 bits, but this was adjusted for some families as described below.

Quality control

Once the seed and full alignments of a family have been constructed, a number of quality controls were performed. False-positives and false-negatives relative to a reference clustering, usually from Prosite, were examined. Because Prosite describes motifs, the clusterings cannot always agree completely. It is ensured that neither the seed nor full alignment overlaps by even a single residue with any other family. Both the alignments and the annotation are checked for format errors.

A problem with Pfam's strategy is that there is no intrinsic protection against one protein scoring high with two HMMs if its sequence lies 'in between' the two families. This typically happens when two families are treated as separate, although they are known to be related. One case of this is the EGF domains and the related EGF-like domains found in laminins, where the laminin EGF-like modules are 20–30 residues longer than normal EGF domains and have eight instead of six conserved cysteines, possibly forming a fourth disulfide bond. When training an HMM on a cross-section of many EGF domains, this HMM will typically give a high score to laminin EGF-like domains. However, it was possible to train a tight EGF HMM where the alignment was very strict about features that are different from laminin EGF-like domains, such as the exact spacing between some conserved cysteines. This HMM would only recognize nonlaminin EGF domains.Pfam-A is checked for any overlaps between families and if this is found either the seed alignment is modified or the score cutoffs are raised slightly.

Format

The Pfam format for the alignments is for each sequence segment: name/start-end followed by the padded sequence on one line. The name is the Swissprot acronym and the start and end are the coordinates of the first and last residues of the sequence segment. In the release flat file the Swissprot accession number is added to the end of each sequence line. The annotation follows the Swissprot flatfile format closely; each family in Pfam-A has a permanent referenceable accession number (Pfxxxxx), an ID name, and a definition line. An example of annotation and alignment is shown in Figure 2. The field labels in Figure 2A follow the Swissprot syntax,[1] with the addition of AU (alignment author), SE (seed membership source), AL (seed alignment method), GA (gathering method to find all members), and AM (alignment method of all members to HMM).

Pfam-B

To cluster all protein sequences not covered by Pfam-A, the Domainer program,[10] version 1.6, was run. Domainer uses pairwise homology data reported from Blastp[29] to construct aligned families. Blastp was only run on the part of Swissprot that was not present in Pfam-A. In release 1.0 of Pfam this was 81% of Swissprot 33. These sequences were prepared by extracting all sequence sections larger

than 30 residues that were not covered in Pfam-A into separate entries. A protein with a Pfam-A domain in the center that has long flanking regions on either side will thus generate two entries. By doing this, Domainer will consider each section as an independent sequence and the boundary to the Pfam-A segment will be used as a real domain boundary. All sequences known to be fragments were omitted because these would induce false domain boundaries in Domainer.

The Domainer process was further improved by filtering the Blastp output with MSPcrunch[28] to remove biased composition matches, trim off overlapping ends of consecutive BLAST matches, and to reduce redundancy. As shown in Figure 3, the growth of homologous sequence sets (HSSs) is practically linear with the number of homologous sequence pairs (HSPs) processed, whereas running Domainer on all of Swissprot gives rise to a large plateaux in areas of large redundancy.[10] Although Pfam 1.0 is based on release 33 of Swissprot, which contains more than twice as many sequences as release 21, which ProDom 21 was based on, the number of HSPs was slightly reduced. Without reduction in redundancy by Pfam-A and MSPcrunch, a quadrupling would have been expected. The time consumption for processing the HSPs into HSSs was 26.3 hours on one workstation. Performing the Blastp all versus all comparison took a total of 184.6 hours but the elapsed time was reduced by running on a number of workstations in parallel. These timings show that it is clearly feasible to rerun the process periodically.

The Pfam-B alignments are released together with Pfam-A in one flat file. The format is essentially the same but each Pfam-B cluster is assigned a volatile accession number (PDxxxxx), which is only valid for a particular release. Information-sparse alignments that Domainer sometimes produces are avoided by excluding any alignment where more than 25% of the residues are gaps. In Pfam 1.0 this eliminated 34 of 11,963 alignments.

Incremental updating

Pfam was designed with easy updating in mind. When new sequences are released, they are compared with the existing models and if they score above the cutoff they are automatically added to the full alignment. Normally the seed alignment is not altered, except for the updating of corrected seed sequences. However, if new sequences give rise to problems, such as strong cross-reaction between families, the seeds may have to be improved to become more specific for the respective families. Once Pfam-A is brought up to date, Pfam-B is regenerated on the rest of Swissprot as described above.

RESULTS

We have constructed and made available a comprehensive library of protein domain families, as described in the Methods section. Together with the HMM technology, this can provide an advance over traditional database searching in sequence analysis for classification purposes. Figure 4A illustrates the proportions of Swissprot that are covered by Pfam-A and Pfam-B. One-third of all Swissprot proteins have one or more domains in Pfam-A and a fifth of all residues are aligned in a Pfam-A family. Pfam-B is roughly twice the size of Pfam-A, leaving only 22% of all proteins without any segment in Pfam at all. Pfam is available via anonymous FTP at ftp.sanger.ac.uk and genome.wustl.edu in /pub/databases/Pfam. There are two main data files: pfam, which contains the annotation and alignments of all Pfam families, and swissPfam, which contains the Pfam domain organization for each Swissprot entry in Pfam. There are also WorldWide Web servers on http://www.sanger.ac.uk/Pfam and http://genome.wustl.edu/Pfam, which allow browsing and HMM searching against Pfam-A with a query sequence. Table I summarizes the families currently in Pfam-A and the sizes of the seed and full alignments. On average, the full alignments have 3.5 times as many members as the seed alignments. Approximately 60% of the Pfam-A families have at least one member with a known structure. These families are cross-referenced to the protein structure database PDB,[30] which is used to link them to the structural classification database SCOP[12] from the Pfam WWW servers.

The primary use of Pfam is as a tool to identify and classify domains in protein sequences. We applied it to Wormpep 10, a database of 4874 predicted proteins from genomic sequencing of *C. elegans*.[31] The 2973 proteins for which no informative similarity has been found using the standard Blast/MSPcrunch approach[28] were searched for Pfam matches. As significance cutoffs, the previously recorded cutoffs that exclude negatives for each Pfam family were used. The 211 Pfam matches were found in 144 unannotated sequences. A number of these matches had very high scores, indicating that they would probably have been found by BLAST too but had been missed because of human error. We have found empirically that most matches found by Pfam but not by BLAST have scores below 35 bits. Table II lists the 118 matches with scores below 35 bits, representing genuinely novel classifications. Adding all of them to the already annotated *C. elegans* predicted proteins yields a classification rate of ~42%. As seen in Figure 4B, already half that amount, 21%, is covered by matches to the Pfam-A HMM library.

An interesting case of family merging that illustrates the level of clustering in Pfam is shown in Figure 5. Here two families that were previously not considered related could be merged. One family is the glycoprotein hormones (Prosite: PDOC00234) and the other is a family of connective tissue growth factor-like and COOH-terminal domains in extracel-

```
ID   response_reg
AC   PF00072
DE   Response regulator receiver domain
AU   Sonnhammer ELL
SE   Prodom
AL   Clustalw
GA   Bic_raw 25 hmmls 25
AM   hmma -qR
RA   Pao, G.M., Saier, M.H.
RL   J. Mol. Evol. 40:136-154(1995).
DR   SCOP; 3chy; fa;
CC   This domain receives the signal from the sensor partner in
CC   bacterial two-component systems.  It is usually found N-terminal
CC   to a DNA binding effector domain.
```

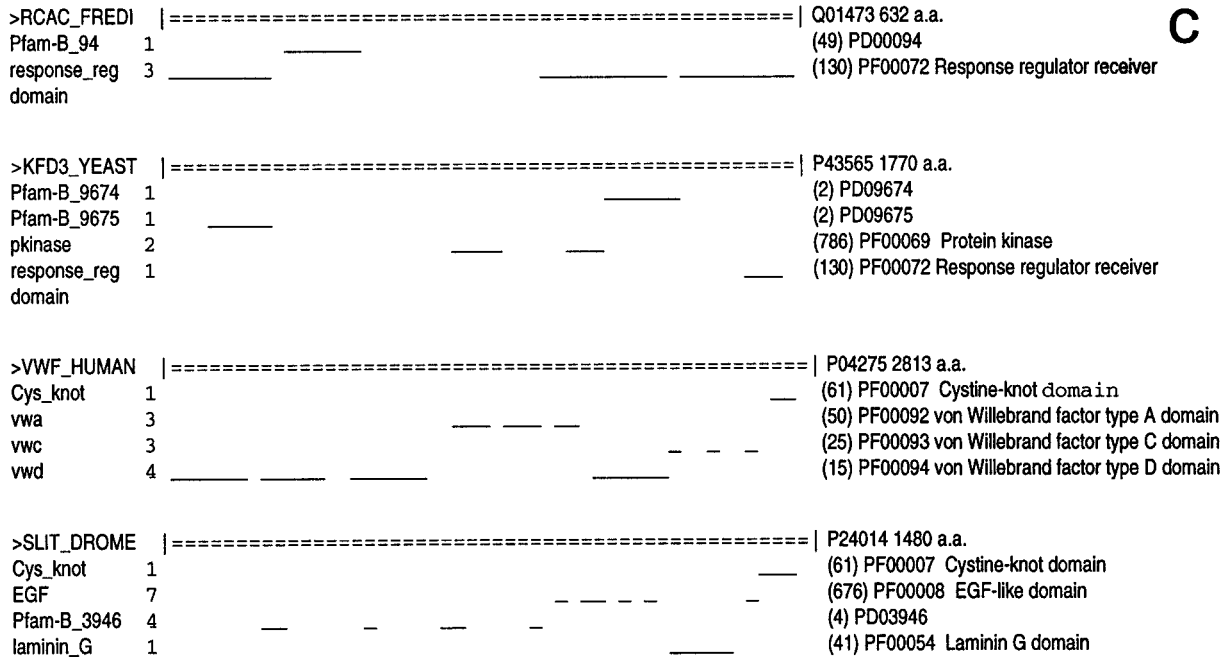
a

b

Fig. 2. Example of the Pfam-A family response_reg (PF00072) with annotation (**A**) and alignment (**B**) (only part shown). KFD3_YEAST and the middle domain of RCAC_FREDI are novel members of this family (see text). The Pfam domain (**C**) organization of these two proteins and two other examples of modular proteins. This schematic representation is provided for each protein in Pfam in the release file swissPfam. The entire sequence is represented with '=' and the Pfam domains with '-' on the lines below. The columns of the domain lines are: Pfam ID, nr. of domains, schematic, nr. of members in the family, Pfam accession nr., description (Pfam-A families only), and start and end coordinates of the segments (not shown here). Example of a Pfam-B family (**D**) produced by Domainer. This family contains the DNA binding effector domain of RCAC_FREDI.

lular proteins.[32] None of these references mention the other family. After we had noticed this family merger, which gives a good quality alignment, we learned that the structure of a glycoprotein hormone had recently been determined to be a cystine-knot fold,[33] which is the fold adopted by the growth factors TGF-β2,[34] NGF,[35] and PDGF-B.[36] The link between these and the family of extracellular COOH-terminal domains had already been made.[32] Ironically, TGF-β2, NGF, and PDGF-B share so few sequence features with the glycoprotein hormones, the connective tissue growth factors, and the extracellular COOH-terminal domains that they could not be included in the Pfam family.

During the construction of Pfam, a number of strong matches were found that despite good sequence similarity had not been classified as true members before. The alignments in Figure 2B and C contain two examples of this in the family Pfam: response_reg. Members of this family are usually found as a single NH_2-terminal domain in response regulators of two-component systems, where it receives a signal by phosphorylation by a sensor molecule. The signal is then usually transduced to a COOH-terminal DNA binding transcription factor, which turns on the expression of a set of downstream genes. Sometimes the receiver domain is not combined with any other domains on the same chain or is



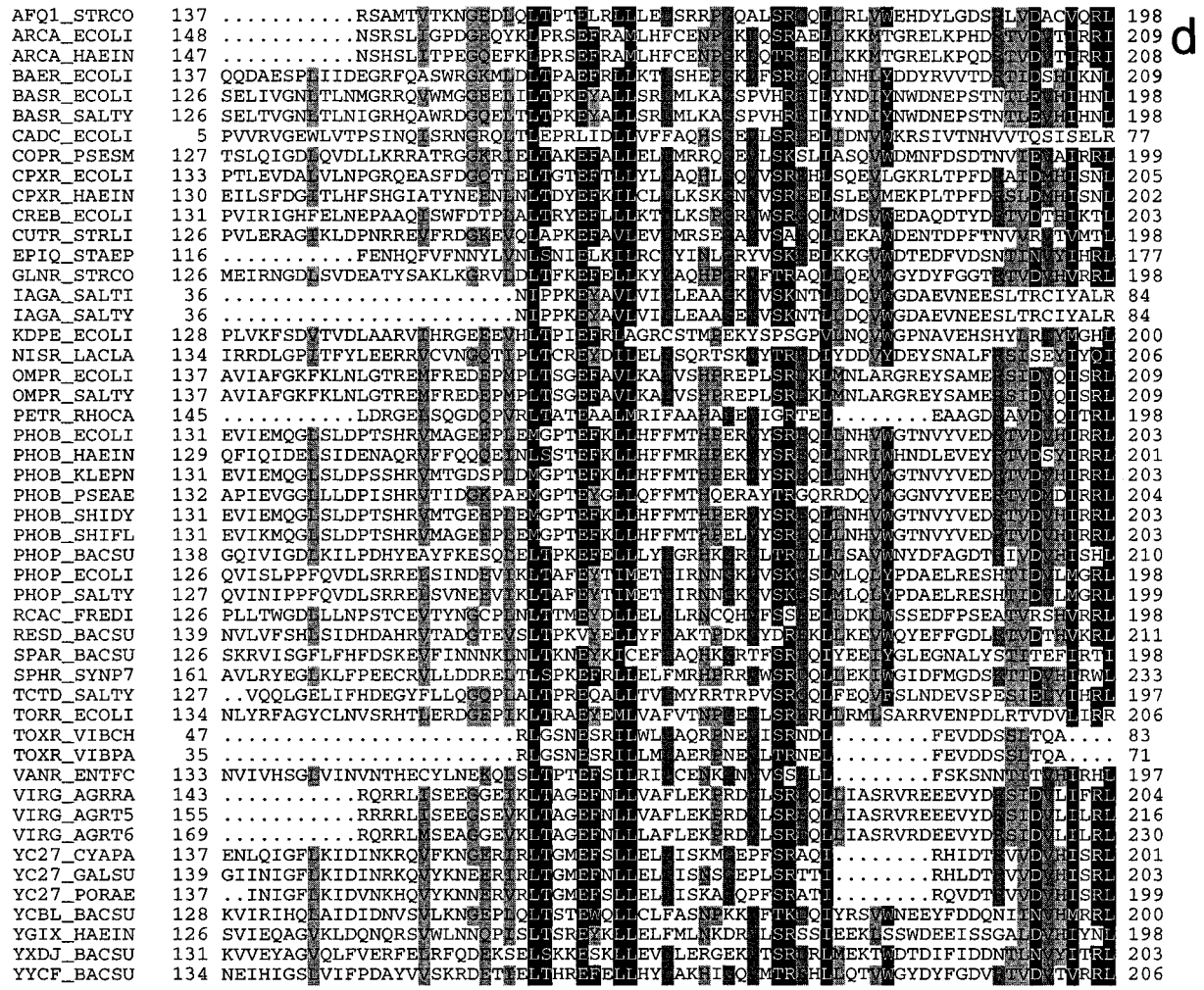


Figure 2 *(Continued).*

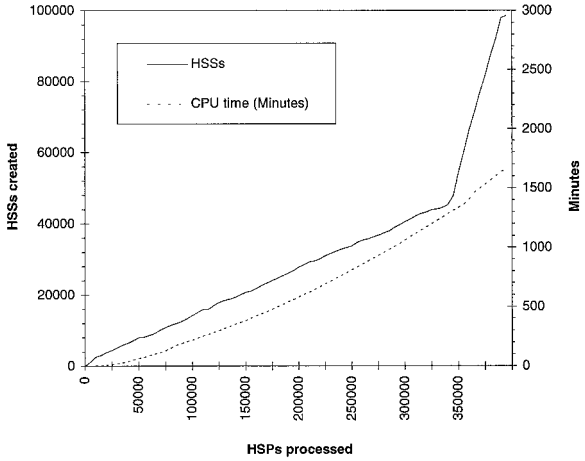


Fig. 3. Construction of Pfam-B by Domainer. Plot of Domainer run on Swissprot 33, excluding sequences in Pfam-A. Domainer groups the pairwise matches (HSPs) into stacks of matches (HSSs) if different pairs share sequence regions. The 46,293 subsequences gave rise to 392,207 HSPs, which resulted in 98,551 HSSs in 11,929 families after subsequent clustering by Domainer. When Domainer is run on the entire Swissprot, much time is spent on processing redundant pairs generated by large families, generating long horizontal plateaus in the plot (see ref. 10). In contrast, the Pfam plot is virtually linear because the most redundant families are already in Pfam and was thus removed before running Domainer. The sharp increase of the curve's slope at the end is caused by adding all full-length sequences as pseudomatches after all the heterogeneous matches.

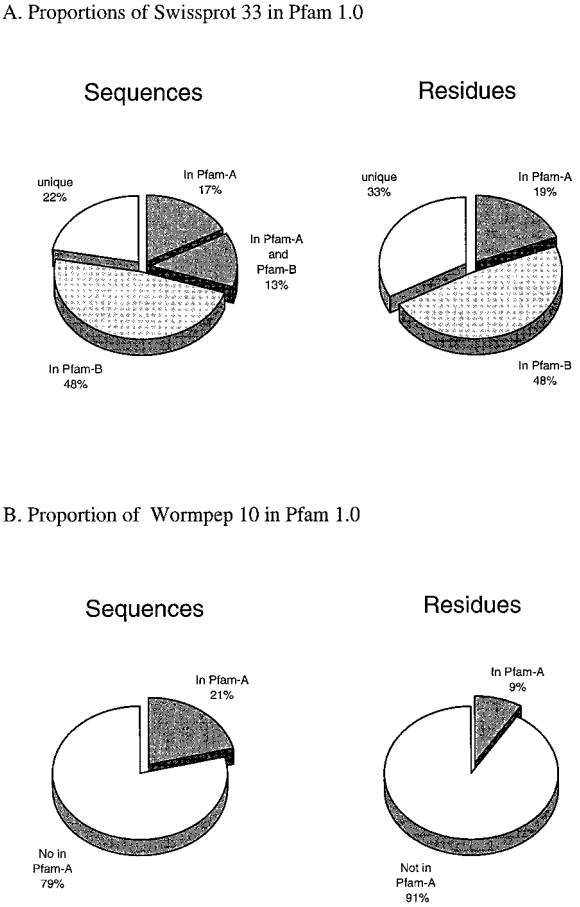


Fig. 4. Proportion of Swissprot 33 (**A**) in Pfam, based on sequences and residues. The portion of unique sequences is slightly overestimated because of the exclusion of fragments and sequences shorter than 30 residues from Pfam-B. Proportion of Wormpep 10 (**B**) comprising 4874 predicted *C. elegans* proteins that is covered by Pfam matches.

combined with other types of modules, such as kinase domains. The cyanobacterial protein *rcaC* (Swissprot: RCAC_FREDI Q01473) was previously found to have a duplicated receiver domain.[10] We now report a third receiver-like domain between the two previously described ones. Most of the conserved features are still clearly recognizable in this third domain, although it has diverged further from the other two domains. The other novel annotation in Figure 2B and C is in the yeast protein KFD3_YEAST (Swissprot P43565), which was found as ORF YFL033c by genomic sequencing of *Saccharomyces cerevisiae* chromosome VI.[37] As seen in Figure 2C, this protein has a protein kinase domain (split up in two matches) and one receiver domain. In the original analysis it was only described as "protein kinase." It further shares domains (Pfam-B_9674 and Pfam-B_9675) with *cek1* in *Schizosaccharomyces pombe* (Swissprot CEK1_SCHPO P38938), which also contains the protein kinase domain but lacks the receiver domain.

Another example is the finding of a new fibronectin type III (FN3) domain[38] in a mammalian glycohydrolase. FN3 domains have already been found in many bacterial glycohydrolases[39,40] but since this domain combination was found to be limited to the bacterial kingdom it was assumed that horizontal gene transfer had taken place from animal proteins with a completely different function. We have detected an FN3 domain in the COOH-terminal part of human, dog and mouse α-l-iduronidase (Swissprot IDUA_HUMAN P35475, IDUA_CANFA Q01634, and IDUA_MOUSE P48441) (Figure 6A). The closest homologue is β-xylosidase from the bacterium *Thermoanaerobacter saccharolyticum*, which lacks the FN3 domain. The discovery of an animal glycohydrolase linked to an FN3 domain raises questions about the conclusion that all FN3 domains in bacterial glycohydrolases have arisen by horizontal transfer of the FN3 domain from an animal source. An alternative scenario is that some ancestral glycohydrolases also possessed FN3 domains.

We have also detected previously undescribed Kazal-type protease inhibitor domains[41] in human and rat organic anion transporters (Swissprot OATP_HUMAN P46721 and OATP_RAT P46720) and in rat prostaglandin transporters (Swissprot PGT_RAT Q00910), as shown in Figure 7. As far as we know, this is the first time a Kazal domain has

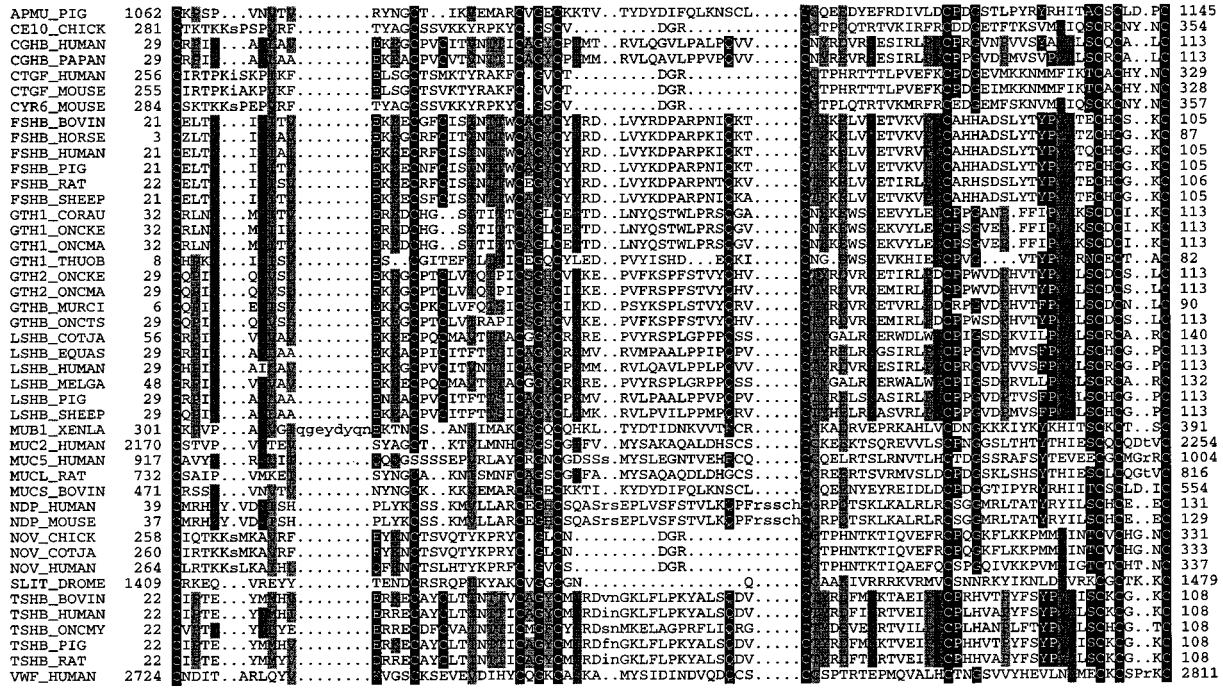


Fig. 5. Selected members from Pfam:Cys_knot (PF0007). This family clusters the two previously described subfamilies CTGF-like (connective tissue growth factor) and glycoprotein hormones in one single superfamily. The similarity has recently been structurally confirmed.

been described in transmembrane proteins. From the hydrophobicity profile of these transporters,[42] it is clear that the predicted Kazal domain lies in a region of ~90 residues between transmembrane helices 9 and 10. This region was predicted to protrude on the outside of the membrane by the program TopPred II[43] for both PGT and OATP. This supports the possibility of a disulfide-rich globular Kazal domain, which may well be important for substrate binding.

To what extent are proteins modular? With Pfam, we can address this problem with higher accuracy than before. Of the proteins in Swissprot 33 containing at least one Pfam-A domain, 17% contain two or more domains, whereas 2.5% have five or more domains. This is only a lower bound because: 1) not all domains are present in Pfam-A, 2) HMMs are not perfectly sensitive, and 3) it is based on proteins in Swissprot, which probably is biased toward single domain proteins. We have done the same analysis on Wormpep 10, which should represent a relatively unbiased set of proteins. Twenty-eight percent of the proteins that matched Pfam-A families matched in two or more domains, whereas 4% matched in five or more domains. We expect that this number is higher for the nematode *C. elegans* than it would be for single cell organisms.

DISCUSSION

We have presented a database that combines high quality alignment information with high coverage of known protein sequences. The level of clustering in Pfam-A is largely a result of the sort of alignments we aimed at: full domain alignments. If subfamilies are too diverse, aligning them together will produce a poor alignment with poor discriminative power. The clusters are thus on a level that gives maximum cluster sizes without disrupting the alignment. In many Pfam-A families the overall sequence similarity is discernible but not very strong. Clustering at a higher similarity level, like PIRALN[2] where the average family only has 6.7 members (Table III), would give alignments of very tight subfamilies where little evolutionary information is contained. This would diminish the advantages of multiple alignment-based search methods like HMM by rendering them less sensitive to recognizing distant members. In Pfam related subfamilies are generally merged into one family to achieve as diverse clusters as possible without compromising alignment quality.

We have chosen a flat structure of families for Pfam rather than a hierarchy of clusters. Maintaining a hierarchy of clearly related families would have the advantage of more fine-grained classification. The current clustering of Pfam often will not permit functional inference of a match, because proteins with a common structural origin but diverged functions may be bundled in one family. However, there were a number of reasons not to choose hierarchical clustering. Creating the hierarchy of clusters for each family remains a hard and labor-intense problem, for which no efficient and robust algorithm is

TABLE I. The Families Included in Release 1.0 of Pfam-A and the Number of Members in the Full and Seed Alignments

Description	Members in full/seed
7 transmembrane receptor (Rhodopsin family)	530/64
7 transmembrane receptor (Secretin family)	36/15
7 transmembrane receptor (metabotropic glutamate family)	12/8
ATPases Associated with various cellular Activities (AAA)	79/42
ABC transporters	330/63
ATP synthase A chain	79/30
ATP synthase subunit C	62/25
ATP synthase alpha and beta subunits	183/47
C2 domain	101/34
Cytochrome C oxidase subunit I	80/27
Cytochrome C oxidase subunit II	114/36
Carboxylesterases	62/27
Cysteine proteases	95/36
Cystine-knot domain	61/28
Phorbol esters/diacylglycerol binding domain	108/34
C-5 cytosine-specific DNA methylases	57/31
DNA polymerase family B	51/37
E1–E2 ATPases	117/24
EGF-like domain	676/75
Fibroblast growth factors	39/10
Glutamine amidotransferases class I	69/39
Elongation factor Tu family	184/63
Helix-loop-helix DNA binding domain	133/35
Heat shock hsp^{20} proteins	132/52
Heat shock hsp^{70} proteins	171/34
Bacterial regulatory helix-loop-helix proteins, lysR family	101/65
Bacterial regulatory helix-loop-helix proteins, araC family	65/42
KH domain family of RNA binding proteins	51/20
Kunitz/Bovine pancreatic trypsin inhibitor domain	79/44
Methyl-accepting chemotaxis protein (MCP) signaling domain	24/10
Class I Histocompatibility antigen, domains alpha 1 and 2	151/25
NADH dehydrogenases	61/25
Phosphoglycerate kinases	51/25
PH (Pleckstrin homology) domain	77/41
Purine/pyrimidine phosphoribosyl transferases	45/26
Ribosome inactivating proteins	37/19
Ribulose bisphosphate carboxylase, large chain	311/17
Ribulose bisphosphate carboxylase, small chain	107/49
Ribosomal protein S12	60/23
Ribosomal protein S4	54/19
Src Homology domain 2	150/58
Src Homology domain 3	161/62
Ser/Thr protein phosphatases	88/17
Transforming growth factor beta like domain	79/16
Triosephosphate isomerase	42/20

TABLE I. (Continued)

Description	Members in full/seed
TNFR/NGFR cysteine-rich region	91/51
u-PAR/Ly-6 domain	18/13
Protein-tyrosine phosphatase	122/38
Fungal Zn(2)-Cys(6) binuclear cluster domain	54/29
Actins	160/24
Alcohol/other dehydrogenases, short chain type	186/52
Zinc-binding dehydrogenases	129/45
Aldehye dehydrogenases	69/34
Alpha amylases (family glycosyl hydrolases)	114/54
Aminotransferases class I	63/29
Ank repeat	305/83
Apple domain	16/16
Arf family	43/21
Eukaryotic aspartyl proteases	72/26
Basic region plus leucine zipper transcription factors	95/22
Beta-lactamases	51/38
Cyclic nucleotide binding domain	69/32
Cadherin	168/58
Cellulases (glycosyl hydrolases)	40/30
Connexin	40/16
Copper binding proteins, plastocyanin/ azurin family	61/31
Chaperonins 10 kDa subunit	58/29
Chaperonins 60 kDa subunit	84/32
Crystallins beta and gamma	103/37
Cyclins	80/48
Cystatin domain	88/51
Cytochrome b(COOH-terminal)/b6/petD	133/10
Cytochrome b(NH_2-terminal)/b6/petB	170/9
Cytochrome c	175/58
Double-stranded RNA binding motif	22/16
EF-hand	739/86
Enolases	41/12
2Fe-25 iron-sulfur cluster binding domains	88/18
4Fe-4S ferredoxins and related iron-sulfur cluster binding domains	156/60
4Fe-4S iron sulfur cluster binding proteins, NifH/frxC family	49/16
Fibrinogen beta and gamma chains, COOH-terminal globular domain	18/17
Intermediate filament proteins	146/36
Fibronectin type I domain	49/21
Fibronectin type II domain	37/17
Fibronectin type III domain	456/109
Glutamine synthetase	78/35
Globin	683/62
Glutathione S-transferases	144/61
Glyceraldehyde 3-phosphate dehydrogenases	117/23
Heme-binding domainin cytochrome b5 and oxidoreductases	55/16
Hemopexin	37/14
Bacterial transferase hexapeptide (four repeats)	82/61
Core histones H2A, H2B, H3, and H4	178/30

TABLE I. (Continued)

Description	Members in full/seed
Homeobox domain	385/64
Protein hormones (family of somatotropin, prolactin and others)	111/17
Peptide hormones (fmaily of glucagon, GIP, secretin, VIP)	110/29
Pancreatic hormone peptides	53/15
Ligand binding domain of nuclear hormone receptors	127/32
IG superfamily	1280/65
Small cytokines (intecrine/chemokine), interleukin-8 like	67/33
Insulin/IGF-Relaxin family	132/44
Interferon alpha nad beta domains	47/17
Kazal-type serine protease inhibitor domain	155/53
Beta-ketoacyl synthases	46/11
Kringle domain	126/25
Laminin B (Domain IV)	15/9
Laminin EGF-like (Domains III and V)	134/72
Laminin G domain	41/26
Laminin N-terminal (Domain VI)	10/9
L-lactate dehydrogenases	90/30
Low-density lipoprotein receptor domain class A	98/43
Low-density lipoprotein receptor domain class B	61/23
Lectin C-type domain short and long forms	128/44
Legume lectins alpha domain	43/25
Legume lectins beta domain	40/25
Ligand-gated ionic channels	30/11
Lipases	23/16
Lipocalins	115/58
C-type lysozymes and alpha-lactabulmin	72/21
Metallothioneins	62/21
Mitochondrial carrier proteins	62/32
Myosin head (motor domain)	52/21
Neuroaminidases	55/7
Neurotransmitter-gated ion-channel	145/51
Notch	24/10
FAD/NAD-binding domain in oxidoreductases	101/56
Molybdopterin binding domain in oxidoreductases	35/15
Oxidoreductases, nitrogenase component I and other families	79/31
Cytochrome P450	204/64
Peroxidases	55/26
Phospholipase A2	122/37
Photosynthetic reaction center protein	73/27
Philins (bacterial filaments)	56/23
Protein kinase	786/67
Pou domain-NH_2-terminal to homeobox domain	47/10
peptidyl-prolyl *cis-trans* isomerases	50/28
Pyridine nucleotide-disulphide oxidoreductase class-I	43/23
Ras family	213/61
recA bacterial DNA recombination proteins	74/31
Response regulator receiver domain	130/55
Picornavirus capsid proteins	117/108
Pancreatic ribonucleases	71/30

TABLE I. (Continued)

Description	Members in full/seed
RNase H	87/31
RNA recognition motif (aka RRM, RBD, or RNP domain)	279/70
Retroviral aspartyl proteases	82/34
Reverse transcriptase (RNA-dependent DNA polymerase)	147/50
Serpins (serine protease inhibitors)	105/43
Sigma-54 transcription factors	56/41
Sigma-70 factors	61/33
Copper/zinc superoxide dismutases (SODC)	68/29
Iron/manganses superoxide dismutases (SODM)	69/28
Subtilase family of serine proteases	91/43
Sugar (and other) transporters)	107/51
Sushi domain	346/80
tRNA synthetases class I	35/19
tRNA synthetases class II	29/20
Thiolases	25/24
Thioredoxins	103/52
Thyroglobulin type I repeat	49/22
Snake toxins	172/48
Trefoil (P-type) domain	39/28
Trypsin	246/65
Thrombospondin type I domain	91/32
Tubulin	197/26
von Willebrand factor type A domain	50/37
von Willebrand factor type C domain	25/17
von Willebrand factor type D domain	15/6
WAP-type (Whey Acidic Protein) 'four-disulfide core'	19/18
wnt family of developmental signaling proteins	105/15
Zinc finger, C2H2 type	1452/165
Zinc finger, C3HC4 type	69/52
Zinc finger, C4 type (two domains)	139/27
Zinc finger, CHC class	188/122
Zinc-binding metalloprotease domain	152/45
Zona pellucida-like domain	26/11
Total	22306/6300

Because the seed alignments are smaller than the full alignments, quality control and maintenance become more feasible tasks.

known to us. Subgroups of one superfamily would often be very similar to each other, which would significantly increase the complexity of maintaining the families in a nonoverlapping manner. Furthermore, by using subgroups for similarity searching will increase the search time substantially, but preliminary experiments suggest that no significant increase in sensitivity is gained by searching against subfamilies with the current HMM implementation (data not shown).

It is interesting to compare Pfam clusters with those in Prosite. Although often very similar, they sometimes differ substantially. The reason is that Prosite clusters are usually constructed with a different goal in mind (i.e., describing very short motifs

TABLE II. Excerpt of the Weakest Pfam Matches (scores up to 35 bits) to Previously Unclassified *C. elegans* Proteins

Pfam family ID/Accession	Description	Query	Score
7tm_1/PF00001	7 transmembrane receptor (Rhodopsin family)	B0244.6	27.9
		B0244.7	24.8
		C30B5.5	24.2
		R11F4.2	24.4
		ZK418.6	27.9
		ZK418.7	33.1
		ZK1307.7	26.9
C2/PF00168	C2 domain	2 × T12A2.4	22.6–28.9
DAG_PE-bind/PF00130	Phorbol esters/diacylglycerol binding domain	F13B9.5	29.0
EGF/PF00008	EGF-like domain	F35D2.3	17.6
		K07D8.2	22.3
		5 × R13F6.4	18.2–27.1
		13 × ZK783.1	17.4–30.4
		F28E10.2	25.5
HLH/PF00010	Helix-loop-helix DNA binding domain	C17C3.7	26.4
		C17C3.8	25.5
		C17C3.10	26.4
PH/PF00169	PH (pleckstrin homology) domain	ZK1248.10	34.8
SH2/PF00017	*Src* Homology domain 2	T06C10.3	34.5
ank/PF00023	Ank repeat	3 × M60.7	28.4–34.7
		K04C2.4	33.1
cadherin/PF00028	Cadherin	B0034.3	27.7
cyclin/PF00134	R02F2.1	29.6	
fer4/PF00037	4Fe-4S ferredoxins and releated iron-sulfur cluster binding domains	C25F6.3	23.7
fn3/PF00041	Fibronectin type III domain	K09E2.4	28.6
		ZC374.2	34.3
gluts/PF00043	Glutathione S-transferases	C25H3.7	25.4
ig/PF00047	IG superfamily	F48C5.1	16.0
		3 × K09E2.4	15.9–30.2
		T02C5.3	22.8
		C18A11.7	18.1
		3 × K02E10.8	17.8–25.4
lectin_c/PF00057	Lectin C-type domain short and long forms	ZK666.7	30.5
pkinase/PF00069	Protein kinase	W07A12.4	32.1
rrm/PF00076	RNA recognition motif (aka RRM, RBD, or RNP domain)	C01F6.5	26.0
		EEED8.1	27.1
		C26E6.9A	30.9
sushi/PF00084	Sushi domain	2 × T07H6.5	29.0–34.5
thiored/PF00085	Thioredoxins	C06A6.5	27.3
		C35D10.10	23.3
tsp_1/PF00090	Thrombospondin type I domain	D1022.2	20.0
		F01F1.13	30.5
		F57C12.1	27.2
vwa/PF00092	von Willebrand factor type A domain	ZK666.3	31.2
		ZK666.7	33.9
		ZK673.9	32.8
zf-C2H2/PF00096	Zinc finger, C2H2 type	2 × C09F5.3	23.7–25.6
		D1046.2	20.6
		F21D5.9	28.1
		2 × F26F4.8	24.2–31.1
		4 × F53B3.1	22.3–32.9
		T20H4.2	26.6
		2 × ZC395.9	23.1–31.4
zf-C3HC4/PF00097	Zinc finger, C3HC4 type	C26B9.6	27.8
		EEED8.9	30.4
		F26F4.7	27.5
zf-C4/PF00105	Zinc finger, C4 type (two domains)	F21D12.1B	32.7
zf-CCHC/PF00098	Zinc finger, CCHC class	C27B7.5	24.2
zn-protease/PF00099	Zinc binding metalloprotease domain	F53A9.2	21.2
		F58A6.4	23.5
		F42A10.8	31.3
		F57C12.1	28.6
		K11G12.1	22.8

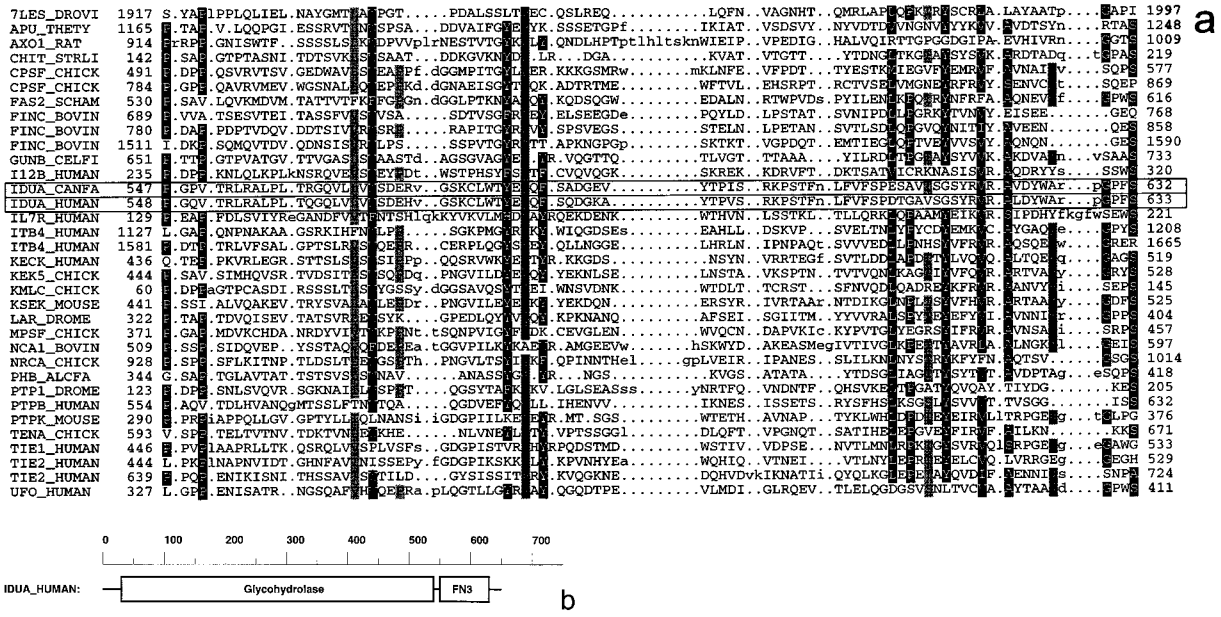


Fig. 6. Selected members (**A**) from Pfam:fn3 (PF00041). The domain (**B**) organization of iduronidase from humans and dogs (IDUA_HUMAN and IDUA_CANFA); the first examples of a mammalian glycohydrolase combined with a fibronectin type III domain.

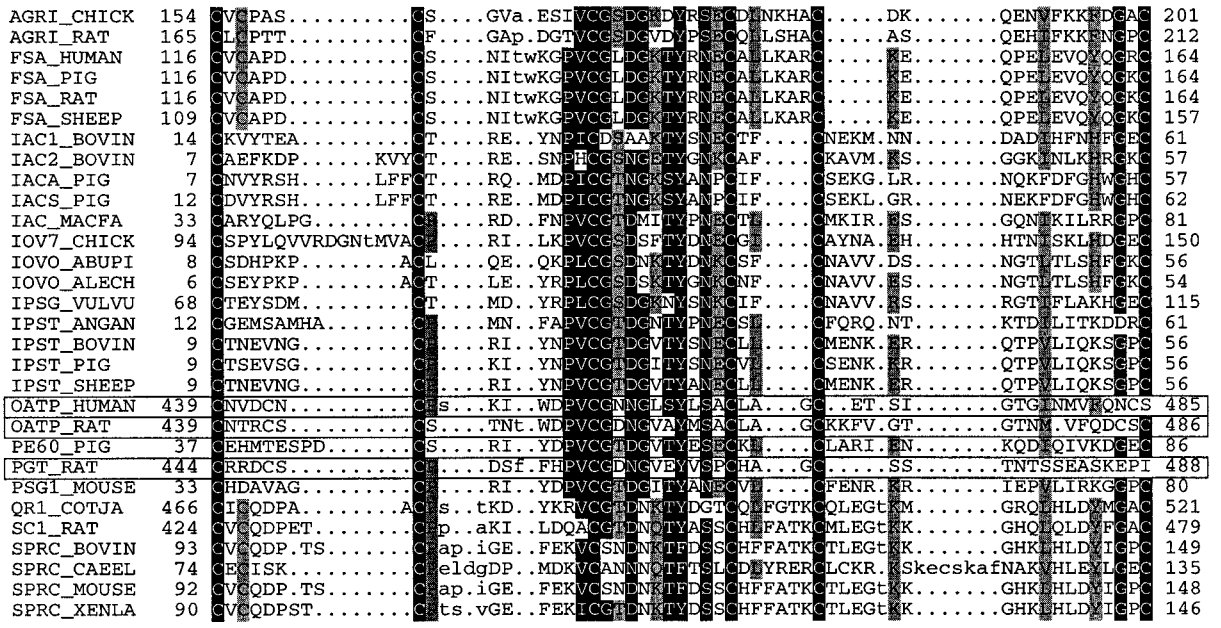


Fig. 7. Selected members from Pfam:kazal (PF00050) showing the novel members OATP_HUMAN, OATP_RAT, and PGT_RAT, which are organic anion and prostaglandin transporters.

important for function). Prosite clusters therefore tend to include as many members as possible without destroying the pattern. The level of Prosite clustering thus depends on how well a pattern can be developed, which in turn depends on the conservation characteristics throughout the family. In some cases several Prosite families are merged together into one Pfam family. For instance Pfam:lipocalin contains the members of both Prosite:PDOC00187 (lipocalin) and PDOC00188 (cytosolic fatty acid binding proteins). In other cases Pfam extends Prosite families with new members, e.g., Pfam:Cys_knot

TABLE III. Comparison of Databases That Contain Protein Family Clusters and Multiple Alignments

	Pfam-A 1.0	Pfam-B 1.0	ProDom 28.0	PIRALN 11.0	BLOCKS 13.0	PRINTS 10.0
Alignment construction	Manual, clustal, HMM	Domainer	Domainer	Pileup	Motif	SOPMA
Source database	Swissprot 33	Swissprot 33	Swissprot 28	PIR 48	Swissprot 32	OWL 26
Clusters	175	11,929	8,031	2,059	872	500
Sequences	15,604	31,931	23,048	11,367	18,593	16,231
Average alignment width (including gaps)	297	180	154	354	32	18
Average cluster size	127	5.7	3.3	6.5	19	37

contains both Prosite:PDOC00234 (glycoprotein hormones β chain) and cystine knot domains from primarily growth factors and extracellular proteins (Figure 5). Prosite families are often overlapping in the sense that one family corresponds to most members, but additional subfamilies are needed to find all members of divergent subfamilies. For example, there are four Prosite patterns for protein kinases (PDOC00100, PDOC00212, PDOC00213, and PDOC00629) but only one Pfam HMM is needed. On the other hand, families that share only a tiny motif of only a few residues, like the P-loop[44] (defined in Prosite PDOC00017 as [AG]xxxxGK[ST]), are not merged in Pfam if there is no interfamily similarity beyond the common motif. Often such patterns are in any case too short to discriminate true matches from false, as is the case for the P-loop. Pfam-A 1.0 contains some 35 families that are absent from Prosite, possibly because no discriminative pattern could be found. Some of these families are currently being added to Prosite as 'matrix' entries instead of patterns.[9]

The protein family databases Prints[45] and Blocks[46] are both based on a set of short ungapped blocks of aligned residues to describe each family. Although the Blocks alignments were generated automatically for all Prosite families, Prints was constructed using a more manual approach to define the family clusters, similar to the Pfam member gathering step (Figure 1). Hence, Prints also contains many clusters that are either absent from Prosite or have a different clustering level. The ungapped block approach has the advantage that robust and fast methods can be used both to discover conserved regions within a family and to search a database for more members.[47] By not allowing gaps, hard to align regions that could easily cause misalignments are avoided. However, gaps also occur in conserved regions and not allowing them may cause either misalignments or truncation of the domain. The principal practical difference from Pfam's approach is that PRINTS and BLOCKS contain short conserved regions, whereas Pfam alignments represent complete domains, facilitating automated annotation.

ProDom is a protein family database that was entirely generated by the Domainer program[10] purely from pairwise sequence homology data with no human knowledge to guide clustering or domain boundary definition. It is useful as a catalogue of comprehensive low quality alignments, but the quality of the alignments and clusters is generally too low to produce information-rich HMMs. Unfortunately, the quality is inversely proportional to the number of family members and very poor for short domain families. For instance, nearly all zinc finger domains were lost due to the crude 'edge trimming' of domain boundaries.

There are a number of other databases that contain valuable aspects of protein family classification but were excluded from the comparison in Table III for various reasons. For instance, Sbase[48] and the matrix entries in Prosite[9] do not provide multiple alignments for the families. The structural clustering in FSSP[49] could in theory be combined with the structure-sequence alignments in HSSP[50] to produce a protein family clustering with multiple alignments, but because this is not explicitly provided and a wide choice of different clustering levels are supplied, we have not attempted to generate this. The Conserved Regions database[51] is only indirectly accessible via the Beauty BLAST server on WWW and not as a complete aligned family database. The MBCRR[52] and Taylor's[53] databases were not included because they were based on relatively small datasets and have not been updated for many years.

The seed/full alignment strategy of Pfam was intended to make updates easy; our aim is to make a new Pfam release for each new release of Swissprot. To make Pfam an integral part of the analysis process of genomic sequencing project, tools to store and display matches to Pfam families are currently being added to ACEDB.[54] This will allow inspection of HMM matches aligned to Pfam seed alignments and significantly improve large-scale classification of proteins.

Our results suggest that Pfam is valuable for genomic sequence analysis. The improvement in protein annotation relative to a human expert annotator by using an integrated analysis workbench based on pairwise similarities is more than just an increase in percentage annotated proteins. It avoids many problems inherent to single sequence database searching, such as overreliance on the annotation of the highest-scoring match and misannotation caused

by multidomain proteins. Pfam thus significantly reduces the task of annotators and helps establish a coherent nomenclature.

ACKNOWLEDGMENTS

We thank C. Chothia and M. Gerstein for providing the structural alignment of the globin family, E. Birney for the RNA recognition motif alignment, and Peer Bork for helpful discussions on the fibronectin type III and cystine knot domains. The Sanger Centre is supported by the Wellcome Trust and the MRC. S.R.E. gratefully acknowledges support from Grant HG01363 from the National Institutes of Health National Center for Human Genome Research.

REFERENCES

1. Bairoch, A., Apweiler, R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res. 24:21–25, 1996.
2. George, D.G., Barker, W.C., Mewes, H.-W., Pfeiffer, F., Tsugita, A. The PIR-International Protein Sequence Database. Nucleic Acids Res. 24:17–21, 1996.
3. Casari, G., De Daruvar, A., Sander, C., Schneider, R. Bioinformatics and the discovery of gene function. Trends Genet. 12:244–245, 1996.
4. Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N., Hayes, W.S., Borodovsky, M., Rudd, K.E., Koonin, E.V. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. Curr. Biol. 6:279–291, 1996.
5. Brenner, S.E., Hubbard, T., Murzin, A., Chothia, C. Gene duplications in *H. influenzae*. Nature 378:140, 1995.
6. Gribskov, M., Homyak, M., Edenfield, J., Eisenberg, D. Profile scanning for three-dimensional structural patterns in protein sequences. Comput. Appl. Biosci. 4:61–66, 1988.
7. Attwood, T.K., Beck, M.E., Bleasby, A.J., Degtyarenko, K., Parry Smith, D.J. Progress with the PRINTS protein fingerprint database. Nucleic Acids Res. 24:182–189, 1996.
8. Pietrokovski, S., Henikoff, J.G., Henikoff, S. The Blocks database: A system for protein classification. Nucleic Acids Res. 24:197–201, 1996.
9. Bairoch, A., Bucher, P., Hofmann, K. The PROSITE database, its status in 1995. Nucleic Acids Res. 24:189–196, 1996.
10. Sonnhammer, E.L.L., Kahn, D. Modular arrangement of proteins as inferred from analysis of homology. Protein Sci. 3:482–492, 1994.
11. Green, P., Lipman, D.J., Hillier, L., Waterson, R., State, D., Claverie, J.-M. Ancient conserved regions in new gene sequences and the protein databases. Science 259:1711–1716, 1993.
12. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247:536–540, 1995.
13. Orengo, C.A., Jones, D.T., Thornton, J.M. Protein superfamilies and domain superfolds. Nature 372:631–634, 1994.
14. Krogh, A., Brown, M., Mian, I.S., Sjoelander, K., Haussler, D. Hidden Markov model in computational biology: Applications to protein modelling. J. Mol. Biol. 235:1501–1531, 1994.
15. Eddy, S.R. Hidden Markov models. Curr. Opin. Struct. Biol. 6:361–365, 1996.
16. Gribskov, M., McLachlan, M., Eisenberg, D. Profile analysis: Detection of distantly related proteins. Proc. Natl. Acad. Sci. USA 84:4355–4358, 1987.
17. Eddy, S.R. In: 'The HMMER package.' World Wide Web URL: http://genome.wustl.edu/eddy/hmm.html. 1995.
18. Overington, J.P. Comparison of three-dimensional structures of homologous proteins. Curr. Opin. Struct. Biol. 2:394–401, 1992.
19. Sonnhammer, E.L.L., Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1–10, 1996.
20. Thompson, J.D., Higgins, D.G., Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680, 1994.
21. Higgins, D.G., Bleasby, A.J., Fuchs, R. CLUSTAL V: Improved software for multiple sequence alignment. Comput. Appl. Biosci. 8:189–191, 1992.
22. Eddy, SR. Multiple alignment using hidden Markov models. In: 'ISMB-95; Proceedings Third International Conference on Intelligent Systems for Molecular Biology.' Menlo Park, CA: AAAI Press, 1995:114–120.
23. Gerstein, M., Sonnhammer, E.L.L., Chothia, C. Volume changes in protein evolution. J. Mol. Biol. 236:1067–1078, 1994.
24. Eddy, S.R., Mitchison, G., Durbin, R. Maximum discrimination hidden Markov models of sequence consensus. J. Comput. Biol. 2:9–23, 1995.
25. Tatusov, R.L., Altschul, S.F., Koonin, E.V. Detection of conserved segments in proteins: iterative scanning. Proc. Natl. Acad. Sci. USA 91:12091–12095, 1994.
26. Devereux, J., Haeberli, P., Smithies, O. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12:387–395, 1984.
27. Esterman, L. Bioccelerator: A currently available solution for fast profile and Smith-Waterman searches. Embnet News 2:5–6, 1995.
28. Sonnhammer, E.L.L., Durbin, R. A workbench for large-scale sequence homology analysis. Comput. Appl. Biosci. 10:301–307, 1994.
29. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman D.J. Basic local alignment search tool. J. Mol. Biol. 215:403–410, 1990.
30. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. Protein data bank. In: 'Crystallographic Databases: Data Commission of the International Union of Crystallography.' Cambridge, UK: Chester, 1987:107–132.
31. Hodgkin, J., Plasterk, R.H., Waterston, R.H. The nematode *Caenorhabditis elegans* and its genome. Science 270:410–414, 1995.
32. Bork, P. The modular architecture of a new family of growth regulators related to connective tissue growth factor. FEBS Lett. 2:125–130, 1993.
33. Lapthorn, A.J., Harris, D.C., Littlejohn, A., Lustbader, J.W., Canfield, R.E., Machin, K.J., Morgan, F.J., Isaacs, N.W. Crystal structure of human chorionic gonadotropin. Nature 369:455–461, 1994.
34. Schlunegger, M.P., Gruetter, M.G. Refined crystal structure of human transforming growth factor beta 2 at 1.95 A resolution. J. Mol. Biol. 231:445–458, 1993.
35. McDonald, N.Q., Lapatto, R., Murray-Rust, J., Gunning, J., Wlodawer, A., Blundell, T.L. New protein fold revealed by a 2.3–A resolution crystal structure of nerve growth factor. Nature 354:411–414, 1991.
36. Oefner, C., D'Arcy, A., Winkler, F.K., Eggimann, B., Hosang, M. Crystal structure of human platelet-derived growth factor BB. EMBO J. 11:3921–3926, 1992.
37. Murakami, Y., Naitou, M., Hagiwara, H., Shibata, T., Ozawa, M., Sasanuma, S.I., Sasanuma, M., Tsuchiya, Y., Soeda, E., Yokoyama, K., et al. Analysis of the nucleotide sequence of chromosome VI from *Saccharomyces cerevisiae*. Nat. Genet. 10:261–268, 1995.
38. Bazan, J.F. Structural design and molecular evolution of a cytokine receptor superfamily. Proc. Natl. Acad. Sci. USA 87:6934–6938, 1990.
39. Little, E., Bork, P., Doolittle, R.F. Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. J. Mol. Evol. 39:631–643, 1994.
40. Bork, P., Doolittle, R.F. Proposed acquisition of an animal protein domain by bacteria. Proc. Natl. Acad. Sci. USA 89:8990–8994, 1992.

41. Kazal, L.A., Spicer, D.S., Brahinsky, R.A. Isolation of a crystalline trypsin inhibitor-anticoagulant protein from pancreas. J. Am. Chem. Soc. 70:3034–3040, 1948.
42. Kanai, N., Lu, R., Satriano, J.A., Bao, Y., Wolkoff, A.W., Schuster, V.L. Identification and characterization of a prostaglandin transporter. Science 268:866–869, 1995.
43. Claros, M.G., von-Heijne, G. TopPred II: An improved software for membrane protein structure prediction. Comput. Appl. Biosci. 10:685–686, 1994.
44. Saraste, M., Sibbald, P.R., Wittinghofer, A. The P-loop: A common motif in ATP- and GTP-binding proteins. Trends. Biochem. Sci. 15:430–434, 1990.
45. Attwood, T.K., Beck, M.E. PRINTS: A protein motif fingerprint database. Protein Eng. 7:841–848, 1994.
46. Henikoff, S., Henikoff, J.G. Protein family classification based on searching a database of blocks. Genomics 19:97–107, 1994.
47. Neuwald, A.F., Green, P. Detecting patterns in protein sequences. J. Mol. Biol. 239:698–712, 1994.
48. Murvai, J., Gabrielian, A., Fabian, P., Hatsagi, Z., Degtyarenko, K., Hegyi, H., Pongor, S. The SBASE protein domain library, release 4.0: A collection of annotated protein sequence segments. Nucleic Acids Res. 24:210–214, 1996.
49. Holm, L., Sander, C. The FSSP database: Fold classification based on structure-structure alignment of proteins. Nucleic Acids Res. 24:206–210, 1996.
50. Schneider, R., Sander, C. The HSSP database of protein structure-sequence alignments. Nucleic Acids Res. 24:201–205, 1996.
51. Worley, K.C., Wiese, B.A., Smith, R.F. BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. Genome Res. 5:173–184, 1995.
52. Smith, R.F., Smith, T.S. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc. Natl. Acad. Sci. USA 87:118–122, 1990.
53. Taylor, W.R. Hierarchical method to align large numbers of biological sequences. Methods Enzymol. 183:456–474, 1990.
54. Durbin, R., Thierry-Mieg, J. ACEDB. World Wide Web URL: ftp://ftp.sanger.ac.uk/pub/acedb, 1996.