

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

Erik L.L. Sonnhammer^{*1}, Sean R. Eddy² and Richard Durbin¹

¹Sanger Centre, Hinxton Hall, Cambridge CB10 1RQ, UK and

²Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA

*To whom correspondence should be addressed.

Fax: +44 1223 494919

Tel: +44 1223 494991

Running Title:

A Database of Protein Domain Families

Keywords:

Classification, Clustering, Protein domains, Genome annotation, Hidden Markov model,

Caenorhabditis elegans

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 one hand and quality of alignment and domain definitions on the other. The latter properties are best dealt with by manual approaches, while completeness in practise is only amenable to automatic methods. Here 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 characterised 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.

Using Pfam, a large number of previously unannotated proteins from the *C. 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.

Introduction

Protein sequence databases such as Swissprot¹ and PIR² are becoming increasingly large and unmanageable, mainly 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, as well as orderly management of already known sequences, it would therefore be advantageous to organise 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 regions⁶⁻⁸, often starting from the manually constructed clusters in Prosite⁹, or full-domain alignments using either clusters that were derived manually from PIR² or automatically¹⁰. An issue here is whether to aim for conserved regions only, or whole-domain alignments. 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 to whole-domain approaches, since information in less conserved regions is ignored. The whole-domain approach therefore seems preferable for detailed family-based sequence analysis since 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 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 multi-domain proteins, define domain boundaries. For instance, the Domainer algorithm¹⁰ which performs the clustering of domain families based on all versus all BLASTP matching, is a fully automatic approach that we have used. We are most familiar with its drawbacks and 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 analysing the extent of the Blast matches and from N- and C-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. 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 can not be permanently referenced from other databases. It is not well-suited for classification, since the families lack family-level annotation.

Presently 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 families¹¹. The same type of analysis on the structural level reveals that there are a few families of very frequently occurring folds¹², and it has been estimated that a third of all proteins adopts one of nine 'superfolds'¹³. This led us to believe that a semi-manual 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, since 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 which 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. We strived to include every family with more than 50 members in Pfam-A. All sequence domains not yet 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 uncharacterised 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 paper describes how Pfam was constructed and presents results from applying the Pfam HMM library to analyse protein families in Swissprot and to classify 4874 proteins found in 30 Mb of genomic DNA from *C. elegans*.

Methods

Pfam-A

HMMs

Hidden Markov model based profiles (HMMs) have been used extensively both for the construction of Pfam, and for detecting matches to Pfam families in database sequences. Although hidden Markov models are a general probabilistic modelling technique, we will use HMM in this paper to mean a specific form of model which 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, while 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-Maximisation procedure from unaligned sequences. A protein sequence can be aligned to an HMM 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 profiles¹⁶ 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 profiles¹⁴. A practical reason for choosing HMMs was the suitability to the task of the HMMER package¹⁷, which includes the programs hmmls for finding multiple non-overlapping complete domains in a target sequence, and hmmfs for finding multiple non-overlapping 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 non-redundant 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 since this may contain incomplete or incorrect sequences which 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 alignments¹⁸, Prodom, Blast results, repeats found by Dotter¹⁹ or published alignments. Families were chosen on an *ad hoc* basis, with a bias towards 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 to known salient features such as active site residues or structurally important residues. Of the automated alignment methods used (Clustalw²⁰, Clustalv²¹, HMM training²²), 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 since 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 using the hmmb program. Although care was taken to assure that the seed members did not include very similar sequences, one of two different weighting schemes^{23 24} was applied to minimise any potential bias towards 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 method²⁵ using the BLOSUM62 substitution matrix. However, 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 to all sequences in Swissprot. This was either done directly with the search programs hmmls or hmmfs, or by converting the HMM to a GCG profile²⁶ in order to be able to use the very fast Bioccellerator hardware from Compugen²⁷.

These programs all perform variants of dynamic programming: the programs `bic_profilesearch` on the Bioccellerator and `hmmfs` use a fully local algorithm, while `hmmls` is local in the query sequence but matches the entire HMM. A further difference is that `bic_profilesearch` only reports the highest score, while `hmmls` and `hmmfs` report all scores above a threshold, with co-ordinates. Although the Bioccellerator is about 50 times faster than a workstation, the result has to be post-processed 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` program²⁸ into a mini-database, which was then searched with `hmmfs` or `hmmls`.

If a list of known members of a family was available, the search result was compared to it to make sure that no known members were missed inadvertently. If the seed alignment is very small, one can not 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 penalise 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 negatives relative to a reference clustering, usually from Prosite, were examined. Since Prosite describes motifs, the clusterings can not always agree completely. It is made sure 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 are 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 disulphide 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 recognise non-laminin 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 co-ordinates 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 centre 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 since these would induce false domain boundaries in Domainer.

The Domainer process was further improved by filtering the BLASTP output with MSPcrunch²⁸ to remove biased composition matches, trim off overlapping ends of consecutive Blast matches, and to reduce redundancy. As can be seen in figure 3, the growth of Homologous Sequence Sets (HSSs), is practically linear with the number of homologous Sequence Pairs (HSPs) processed, while running Domainer on all of Swissprot gives rise to large plateaux in areas of large redundancy¹⁰. 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 out of 11963 alignments.

Incremental updating

Pfam was designed with easy updating in mind. When new sequences are released, they are compared to 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 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. A third of all Swissprot proteins have one or more domain 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 in Pfam at all. Pfam is available via anonymous FTP at [ftp.sanger.ac.uk](ftp://ftp.sanger.ac.uk) and [genome.wustl.edu](http://genome.wustl.edu/pub/databases/Pfam) in /pub/databases/Pfam. There are two data files: pfam, which lists all the Pfam families with annotation and alignment, and swissPfam, which contains the Pfam domain organisation for each SwissProt entry in Pfam. There are also World Wide 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 summarises the families currently in Pfam-A and the sizes of the seed and full alignments. On average, the full alignments have four times as many members as the seed alignments. The structure of 60% of the Pfam-A families is known. These families are cross-referenced to PDB³⁰, which is used to link them to the structural classification database SCOP¹² from the Pfam WWW servers

The main 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 approach²⁸ were searched for Pfam matches. As significance cutoffs, the previously recorded cutoffs that exclude negatives for each Pfam family were used. 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 due to human error. We have found empirically that most matches found by Pfam but not by Blast have scores below approximately 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 about 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 which 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 C-terminal domains in extracellular proteins³². 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 fold³³, which is the fold adopted by the growth factors TGF- β 2³⁴, NGF³⁵ and PDGF-B³⁶. The link between these and the family of extracellular C-terminal domains had already been made³². Ironically, TGF- β 2, NGF and PDGF-B share so few sequence features with the glycoprotein hormones, the connective tissue growth factors and the extracellular C-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 figures 2b and 2c contain two examples of this in the family Pfam:response_reg. Members of this family are usually found as a single N-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 C-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 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 domain¹⁰. We now report a third receiver-like domain between the two previously described ones. Most of the conserved features are still clearly recognisable 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 *S. cerevisiae* chromosome VI³⁷. 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 *S. 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) domain³⁸ 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 C-terminal part of human, dog and mouse alpha-l-iduronidase (Swissprot IDUA_HUMAN P35475, IDUA_CANFA Q01634 and IDUA_MOUSE P48441) (see 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 domains⁴¹ in human and rat organic anion transporters (Swissprot OATP_HUMAN P46721 and OATP_RAT P46720), and in rat prostaglandin transporter (Swissprot PGT_RAT Q00910), as shown in figure 7. As far as we know, this is the first time a Kazal domain has been described in transmembrane proteins. From the hydrophobicity profile of these transporters⁴², it is clear that the predicted Kazal domain lies in a region of about 90 residues between transmembrane helices 9 and 10. This region was predicted to protrude on the outside of the membrane by the program TopPred II⁴³ for both PGT and OATP. This supports the possibility of a disulphide-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, while 2.5% have five or more domains. This is only a lower bound since (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 towards single-domain proteins. We have done the same analysis on Wormpep 10, which should represent a relatively unbiased set of proteins. 28% of the proteins that matched Pfam-A families, matched in two or more domains, while 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 which 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 which 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 PIRALN² where the average family only has 6.7 members (see 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 HMMs by rendering them less sensitive to recognising 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 will often not permit functional inference of a match, since 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 labour-intense problem, for which no efficient and robust algorithm is 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 non-overlapping manner. Furthermore, 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 with subfamilies (data not shown).

It is interesting to compare Pfam clusters to 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 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 contains both Prosite:PDOC00234 (Glycoprotein hormones beta chain) and cystine knot domains from mainly 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 e.g. the P-loop⁴⁴ (defined in Prosite PDOC00017 as [AG]xxxxGK[ST]), are not merged in Pfam if there is no inter-family 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 patterns⁹.

The protein family databases Prints⁴⁵ and Blocks⁴⁶ are both based on a set of short ungapped blocks of aligned residues to describe each family. While 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 (see 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 members⁴⁷. 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 program¹⁰ 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 a variety of reasons. For instance, Sbase⁴⁸ and the matrix entries in Prosite⁹ do not provide multiple alignments for the families. The structural clustering in FSSP⁴⁹ could in theory be combined with the structure-sequence alignments in HSSP⁵⁰ to produce a protein family clustering with multiple alignments, but since this is not explicitly provided, and since a wide choice of different clustering levels are supplied, we have not attempted to generate this. The Conserved Regions database⁵¹, is only indirectly accessible via the Beauty Blast server on WWW and not as a complete aligned family database. The MBCRR⁵² and Taylor's⁵³ databases were not included since they were based on relatively small datasets and have not been updated for many years.

The seed/full alignment strategy of Pfam was intended 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 ACEDB⁵⁴. 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 using an integrated analysis workbench based on pairwise similarities²⁸ is more than just an increase in percentage annotated proteins. It avoids many problems inherent to single sequence database searching, such as over-reliance 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.

Acknowledgements

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 NIH 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. Abstract.
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. Petrokovski 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. Abstract.
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. Abstract.

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.
Abstract.
13. Orengo C.A., Jones D. T., Thornton J. M. Protein superfamilies and domain superfolds. *Nature* 372:631-634, 1994.
Abstract.
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.
Abstract.
15. Eddy S.R. Hidden Markov Models. *Curr. Opinion Struct. Biol.* 6:361-365, 1996.
16. Gribskov M., McLachlan M., Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.* 84:4355-4358, 1987.
Abstract.
17. Eddy SR. 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.
Abstract.
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." AAAI Press, Menlo Park. 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. U.S.A.* 91:12091-12095, 1994.
Abstract.
26. Devereux J., Haeberli P., Smithies O. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 12:387-395, 1984.
Abstract.
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.
Abstract.
30. Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. Protein Data Bank. In: "Crystallographic Databases. Data Commission of the International Union of Crystallography." Bonn, Cambridge, 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. Laphorn 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 Å 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-Å 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. and others. 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. U.S.A.* 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. U.S.A.* 89:8990-8994, 1992.
41. Kazal L.A., Spicer D. S., Brahinsky R. A. *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.
Abstract.
46. Henikoff S., Henikoff J. G. Protein family classification based on searching a database of blocks. *Genomics* 19:97-107, 1994.
Abstract.
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 Research* 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. U.S.A.* 87:118-122, 1990.
Abstract.
53. Taylor W.R. Hierarchical method to align large numbers of biological sequences. *Meth. Enzymol.* 183:456-474, 1990.
Abstract.
54. Durbin R., Thierry-Mieg J. ACEDB. World Wide Web URL: <http://www.sanger.ac.uk/acedb>, 1996.

Figures

Figure 1. The procedure to construct the alignments and HMM for a Pfam-A family. ¹Initial seed alignments are taken either from a published alignment or are made by one of the methods described in the text. ²By ‘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.

Figure 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 and are novel members of this family (see text). C. The Pfam domain organisation 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 whole 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 co-ordinates of the segments (not shown here). D. Example of a Pfam-B family produced by Domainer. This family contains the DNA binding effector domain of RCAC_FREDI.

Figure 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. 46293 subsequences gave rise to 392207 HSPs, which resulted in 98551 HSSs in 11929 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 plateaux in the plot (See ¹⁰, figure 3). In contrast, the Pfam plot is virtually linear, since 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 pseudo-matches after all the heterogeneous matches.

Figure 4. A. Proportion of Swissprot 33 in Pfam, based on sequences and residues. The portion of unique sequences is slightly overestimated due to the exclusion of fragments and sequences

shorter than 30 residues from Pfam-B. B. Proportion of Wormpep 10, comprising 4874 predicted *C. elegans* proteins that is covered by Pfam matches.

Figure 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.

Figure 6. A. Selected members from Pfam:fn3 (PF00041). B. The domain organisation of iduronidase from human and dog (IDUA_HUMAN and IDUA_CANFA), the first examples of a mammalian glycohydrolase combined with a fibronectin type III domain.

Figure 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.

Table I. The families included in release 1.0 of Pfam-A and the number of members in the full and seed alignments. Because the seed alignments are smaller than the full alignments, quality control and maintenance become more feasible tasks.

Description	Members in full / seed		
		Cytochrome b(N-terminal)/b6/petB	170 / 9
		Cytochrome c	175 / 58
7 transmembrane receptor (Rhodopsin family)	530 / 64	Double-stranded RNA binding motif	22 / 16
7 transmembrane receptor (Secretin family)	36 / 15	EF hand	739 / 86
7 transmembrane receptor (metabotropic glutamate family)	12 / 8	Enolases	41 / 12
ATPases Associated with various cellular Activities (AAA)	79 / 42	2Fe-2S iron-sulfur cluster binding domains	88 / 18
ABC transporters	330 / 63	4Fe-4S ferredoxins and related iron-sulfur cluster binding domains.	156 / 60
ATP synthase A chain	79 / 30	4Fe-4S iron sulfur cluster binding proteins, NifH/frxC family	49 / 16
ATP synthase subunit C	62 / 25	Fibrinogen beta and gamma chains, C-terminal globular domain	18 / 17
ATP synthase alpha and beta subunits	183 / 47	Intermediate filament proteins	146 / 36
C2 domain	101 / 34	Fibronectin type I domain	49 / 21
Cytochrome C oxidase subunit I	80 / 27	Fibronectin type II domain	37 / 17
Cytochrome C oxidase subunit II	114 / 36	Fibronectin type III domain	456 / 109
Carboxylesterases	62 / 27	Glutamine synthetase	78 / 35
Cysteine proteases	95 / 36	Globin	683 / 62
Cystine-knot domain	61 / 28	Glutathione S-transferases.	144 / 61
Phorbol esters / diacylglycerol binding domain	108 / 34	glyceraldehyde 3-phosphate dehydrogenases	117 / 23
C-5 cytosine-specific DNA methylases	57 / 31	Heme-binding domain in cytochrome b5 and oxidoreductases	55 / 16
DNA polymerase family B	51 / 37	Hemopexin	37 / 14
E1-E2 ATPases	117 / 24	Bacterial transferase hexapeptide (four repeats)	82 / 61
EGF-like domain	676 / 75	Core histones H2A, H2B, H3 and H4	178 / 30
Fibroblast growth factors	39 / 10	Homeobox domain	385 / 64
Glutamine amidotransferases class-I	69 / 39	Protein hormones (family of somatotropin, prolactin and others)	111 / 17
Elongation factor Tu family	184 / 63	Peptide hormones (family of glucagon, GIP, secretin, VIP)	110 / 29
Helix-loop-helix DNA-binding domain	133 / 35	Pancreatic hormone peptides	53 / 15
Heat shock hsp20 proteins	132 / 52	Ligand-binding domain of nuclear hormone receptors	127 / 32
Heat shock hsp70 proteins	171 / 34	IG superfamily	1280 / 65
Bacterial regulatory helix-loop-helix proteins, lysR family	101 / 65	Small cytokines (intercrine/chemokine), interleukin-8 like	67 / 33
Bacterial regulatory helix-loop-helix proteins, araC family	65 / 42	Insulin/IGF/Relaxin family	132 / 44
KH domain family of RNA binding proteins	51 / 20	Interferon alpha and beta domains	47 / 17
Kunitz/Bovine pancreatic trypsin inhibitor domain	79 / 44	Kazal-type serine protease inhibitor domain	155 / 53
Methyl-accepting chemotaxis protein (MCP) signaling domain	24 / 10	Beta-ketoacyl synthases	46 / 11
Class I Histocompatibility antigen, domains alpha 1 and 2	151 / 25	Kringle domain	126 / 25
NADH dehydrogenases	61 / 25	Laminin B (Domain IV)	15 / 9
Phosphoglycerate kinases	51 / 25	Laminin EGF-like (Domains III and V)	134 / 72
PH (pleckstrin homology) domain	77 / 41	Laminin G domain	41 / 26
Purine/pyrimidine phosphoribosyl transferases	45 / 26	Laminin N-terminal (Domain VI)	10 / 9
Ribosome inactivating proteins	37 / 19	L-lactate dehydrogenases	90 / 30
Ribulose biphosphate carboxylase, large chain	311 / 17	Low-density lipoprotein receptor domain class A	98 / 43
Ribulose biphosphate carboxylase, small chain	107 / 49	Low-density lipoprotein receptor domain class B	61 / 23
Ribosomal protein S12	60 / 23	Lectin C-type domain short and long forms	128 / 44
Ribosomal protein S4	54 / 19	Legume lectins alpha domain	43 / 25
Src Homology domain 2	150 / 58	Legume lectins beta domain	40 / 25
Src Homology domain 3	161 / 62	Ligand-gated ionic channels	30 / 11
Ser/Thr protein phosphatases	88 / 17	Lipases	23 / 16
Transforming growth factor beta like domain	79 / 16	lipocalins	115 / 58
Triosephosphate isomerase	42 / 20	C-type lysozymes and alpha-lactalbumin	72 / 21
TNFR/NGFR cysteine-rich region	91 / 51	Metallothioneins	62 / 21
u-PAR/Ly-6 domain	18 / 13	Mitochondrial carrier proteins	62 / 32
Protein-tyrosine phosphatase	122 / 38	Myosin head (motor domain)	52 / 21
Fungal Zn(2)-Cys(6) binuclear cluster domain	54 / 29	Neuramidases	55 / 7
Actins	160 / 24	Neurotransmitter-gated ion-channel	145 / 51
Alcohol/other dehydrogenases, short chain type	186 / 52	Notch	24 / 10
Zinc-binding dehydrogenases	129 / 45	FAD/NAD-binding domain in oxidoreductases	101 / 56
Aldehyde dehydrogenases	69 / 34	Molybdopterin binding domain in oxidoreductases	35 / 15
Alpha amylases (family of glycosyl hydrolases)	114 / 54	Oxidoreductases, nitrogenase component 1 and other families	79 / 31
Aminotransferases class-I	63 / 29	Cytochrome P450	204 / 64
Ank repeat	305 / 83	Peroxidases	55 / 26
Apple domain	16 / 16	Phospholipase A2	122 / 37
Arf family	43 / 21	Photosynthetic reaction center protein	73 / 27
Eukaryotic aspartyl proteases	72 / 26	Pilins (bacterial filaments)	56 / 23
Basic region plus leucine zipper transcription factors	95 / 22	Protein kinase	786 / 67
Beta-lactamases	51 / 38	Pou domain - N-terminal to homeobox domain	47 / 10
Cyclic nucleotide-binding domain	69 / 32	Peptidyl-prolyl cis-trans isomerases	50 / 28
Cadherin	168 / 58	Pyridine nucleotide-disulphide oxidoreductases class-I	43 / 23
Cellulases (glycosyl hydrolases)	40 / 30	Ras family	213 / 61
Connexin	40 / 16	recA bacterial DNA recombination proteins	74 / 31
Copper binding proteins, plastocyanin/azurin family	61 / 31	Response regulator receiver domain	130 / 55
Chaperonins 10 Kd subunit	58 / 29	picornavirus capsid proteins	117 / 108
Chaperonins 60 Kd subunit	84 / 32	Pancreatic ribonucleases	71 / 30
Crystallins beta and gamma	103 / 37	RNase H	87 / 31
Cyclins	80 / 48		
Cystatin domain	88 / 51		
Cytochrome b(C-terminal)/b6/petD	133 / 10		

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/manganese 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-1 repeat	49 / 22
Snake toxins	172 / 48
Trefoil (P-type) domain	39 / 28
Trypsin	246 / 65
Thrombospondin type 1 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, CCHC class	188 / 122
Zinc-binding metalloprotease domain	152 / 45
Zona pellucida-like domain	26 / 11
Total	22306 / 6300

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 DAG_PE-bind / PF00130 EGF / PF00008	C2 domain Phorbol esters / diacylglycerol binding domain EGF-like domain	2 x T12A2.4	22.6 - 28.9
		F13B9.5	29.0
		F35D2.3	17.6
		K07D8.2	22.3
		5 x R13F6.4	18.2 - 27.1
		13 x 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
		ZK1248.10	34.8
PH / PF00169 SH2 / PF00017 ank / PF00023	PH (pleckstrin homology) domain Src Homology domain 2 Ank repeat	T06C10.3	34.5
		3 x M60.7	28.4 - 34.7
		K04C2.4	33.1
cadherin / PF00028 cyclin / PF00134 fer4 / PF00037	Cadherin Cyclins 4Fe-4S ferredoxins and related iron-sulfur cluster binding domains.	B0034.3	27.7
		R02F2.1	29.6
		C25F6.3	23.7
fn3 / PF00041	Fibronectin type III domain	K09E2.4	28.6
		ZC374.2	34.3
gluts / PF00043 ig / PF00047	Glutathione S-transferases. IG superfamily	C25H3.7	25.4
		F48C5.1	16.0
		3 x K09E2.4	15.9 - 30.2
		T02C5.3	22.8
		C18A11.7	18.1
		3 x K02E10.8	17.8 - 25.4
lectin_c / PF00059 pkinase / PF00069 rrm / PF00076	Lectin C-type domain short and long forms Protein kinase RNA recognition motif. (aka RRM, RBD, or RNP domain)	ZK666.7	30.5
		W07A12.4	32.1
		C01F6.5	26.0
		EEED8.1	27.1
		C26E6.9A	30.9
		2 x T07H6.5	29.0 - 34.5
sushi / PF00084 thioredo / PF00085	Sushi domain Thioredoxins	C06A6.5	27.3
		C35D10.10	23.3
		D1022.2	20.0
tsp_1 / PF00090	Thrombospondin type 1 domain	F01F1.13	30.5
		F57C12.1	27.2
		ZK666.3	31.2
vwa / PF00092	von Willebrand factor type A domain	ZK666.7	33.9
		ZK673.9	32.8
zf-C2H2 / PF00096	Zinc finger, C2H2 type	2 x C09F5.3	23.7 - 25.6
		D1046.2	20.6
		F21D5.9	28.1
		2 x F26F4.8	24.2 - 31.1
		4 x F53B3.1	22.3 - 32.9
		T20H4.2	26.6
		2 x 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 zf-CCHC / PF00098 zn-protease / PF00099	Zinc finger, C4 type (two domains) Zinc finger, CCHC class Zinc-binding metalloprotease domain	F21D12.1B	32.7
		C27B7.5	24.2
		F53A9.2	21.2
		F58A6.4	23.5
		F42A10.8	31.3
		F57C12.1	28.6
		K11G12.1	22.8

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, Clus- tal, HMM	Domainer	Domainer	Pileup	Motif	SOPMA
Alignment quality control	Manual	No	No	No	No	No
Source database	Swissprot 33	Swissprot 33	Swissprot 28	PIR 48	Swissprot 32	OWL 26
Clusters	175	11929	8031	2059	872	500
Sequences	15,604	31,931	23,048	11,367	18,593	16,231
Residues	3,560,959	8,957,230	6,632,274	4,376,550	1,858,812	1,634,436
Average alignment width	297	180	154	354	32	18
Average cluster size	127	5.7	3.3	6.5	19	37

Figure 1

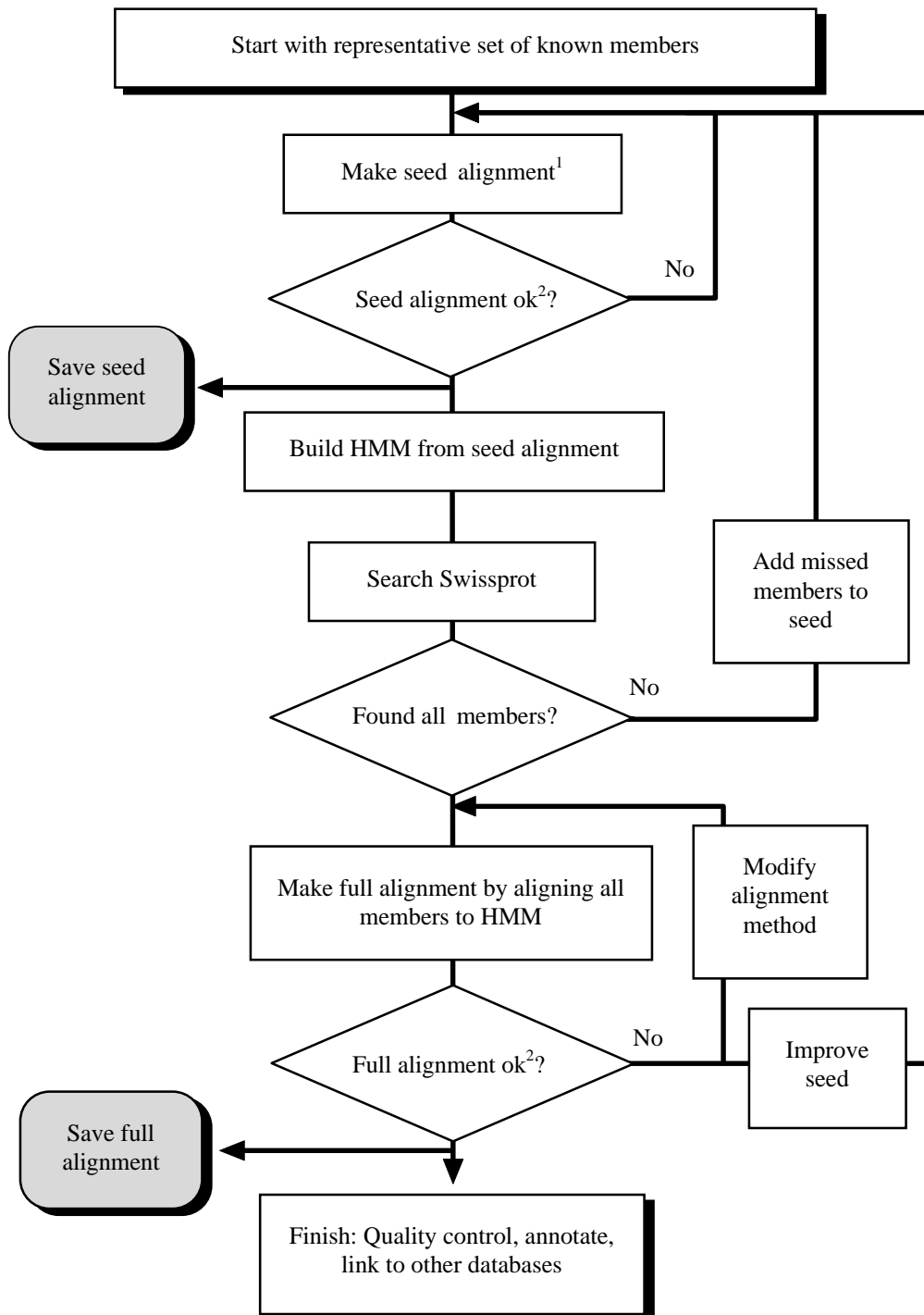


Figure 2 A

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.

Figure 2 C

```
>RCAC_FREDI |=====| Q01473 632 a.a.
Pfam-B_94 1 _____ (49) PD00094
response_reg 3 _____ (130) PF00072 Response regulator receiver domain

>KFD3_YEAST |=====| P43565 1770 a.a.
Pfam-B_9674 1 _____ (2) PD09674
Pfam-B_9675 1 _____ (2) PD09675
pkinase 2 _____ (786) PF00069 Protein kinase
response_reg 1 _____ (130) PF00072 Response regulator receiver domain

>VWF_HUMAN |=====| P04275 2813 a.a.
Cys_knot 1 _____ (61) PF00007 Cystine-knot domain
wva 3 _____ (50) PF00092 von Willebrand factor type A domain
wvc 3 _____ (25) PF00093 von Willebrand factor type C domain
wvd 4 _____ (15) PF00094 von Willebrand factor type D domain

>SLIT_DROME |=====| P24014 1480 a.a.
Cys_knot 1 _____ (61) PF00007 Cystine-knot domain
EGF 7 _____ (676) PF00008 EGF-like domain
Pfam-B_3946 4 _____ (4) PD03946
laminin_G 1 _____ (41) PF00054 Laminin G domain
```

Figure 3

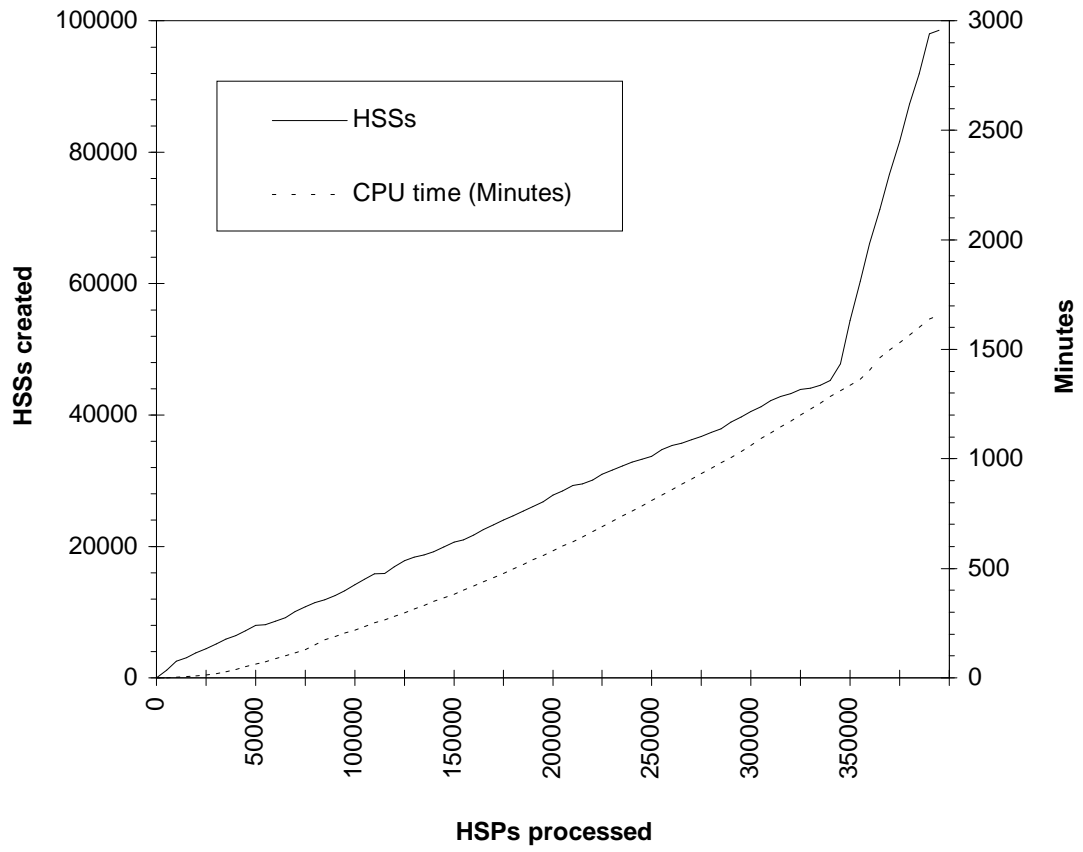
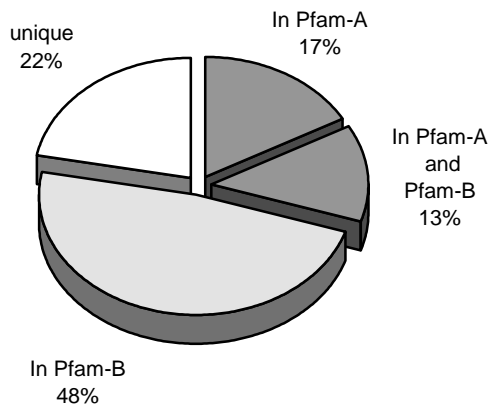


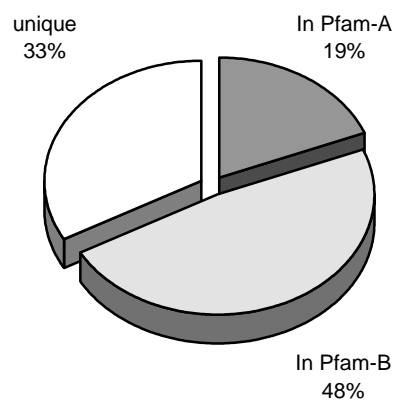
Figure 4

A. Proportions of Swissprot 33 in Pfam 1.0

Sequences

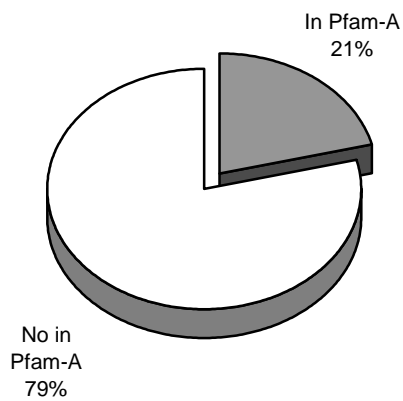


Residues

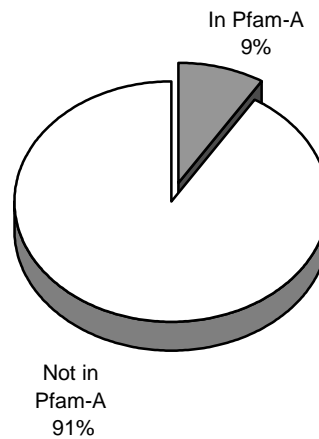


B. Proportion of Wormpep 10 in Pfam 1.0

Sequences



Residues



APQ1_STRCO	2	P	S	I	D	D	E	A	I	R	T	A	E	L	S	T	R	O	G	H	R	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	P	D	I	T	V	I	D	V	I	G	I	D	E	V	G	R	I	R	R	E	D	Q	I	P	I	L	L	R	N	D	D	I	V	V	G	E	S	G	A	M	F	I	R	K	D	P	A	B	I	R	I	112					
APRA_STAUB	1	M	K	E	F	C	E	D	P	K	O	R	E	N	N	V	T	I	K	N	T	I	M	E	E	S	P	M	I	A	L	T	P	E	V	E	L	E	D	A	K	M	N	D	I	G	X	F	L	I	O	S	T	I	N	E	C	K	S	E	R	K	H	D	V	G	N	I	F	V	A	S	H	S	E	L	T	L	F	V	K	A	M	F	I	R	K	D	P	A	B	I	R	I	120
ALGB_PSEAF	9	G	R	I	D	V	D	S	A	L	I	R	E	V	E	K	E	G	Y	S	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	O	G	A	V	I	A	S	P	O	L	I	A	120																						
ARCB_ECOLI	5	2	4	L	M	V	D	E	L	V	A	R	E	V	E	K	E	G	Y	S	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	M	V	A	M	D	V	A	M	T	K	A	L	E	M	E	R	G	E	V	L	K	O	E	Y	N	G	M	D	V	I	S	K	P	S	P	O	L	I	A	637																	
AROC_ECOLI	5	N	R	L	I	V	D	E	D	N	A	R	R	M	S	T	E	F	A	O	G	Y	S	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	E	T	H	C	A	N	N	G	R	T	L	H	F	A	D	H	P	E	781																																				
BAGA_ECOLI	6	M	T	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	B	E	L	C	D	S	H	O	Q	A	V	E	K	A	O	M	P	D	I	L	N	D	781																																			
BBGA_BORDE	3	N	K	V	L	I	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	B	E	L	C	D	S	H	O	Q	A	V	E	K	A	O	M	P	D	I	L	N	D	781																																
CHBB_ECOLI	4	I	R	V	D	S	A	L	I	R	E	V	E	K	E	G	Y	S	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	118																																						
CHBY_ECOLI	5	L	K	E	V	D	P	S	T	A	R	I	W	E	L	E	K	E	G	Y	S	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	119																																			
CMAL_BACSU	1	K	L	I	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	117																																	
COBR_PSEEM	2	S	K	E	L	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	112																																
COXR_HAETN	4	E	T	V	M	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	115																																
COXR_STRLI	1	M	R	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	115																																		
DOCT_RHIME	5	P	S	V	I	D	D	R	O	D	R	I	K	A	O	O	T	E	L	E	A	G	F	T	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	112																																
DGDU_BACSU	4	V	N	T	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	117																																	
ERR1_ARATH	6	L	K	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	117																																		
ERR1_ECOLI	1	M	N	A	I	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	112																																
FIMZ_SALTY	4	A	S	V	I	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	117																																	
FIXJ_RHIME	4	Y	T	H	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	117																																	
FLBD_CADUR	1	M	R	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	119																																		
FRZE_MYXXA	6	5	9	L	R	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	115																																
GACA_PSEFL	2	I	R	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	115																																		
GNBR_PSEAF	8	K	O	I	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	116																																	
HNR_ECOLI	6	A	T	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	117																																		
HOXA_ALCEB	6	P	A	L	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	117																																	
HOPR_RHOCA	6	I	D	H	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O	H	V	A	T	A	S	E	B	D	E	L	K	I	L	R	B	O	R	F	D	C	I	D	L	R	E	G	W	R	V	A	N	A	H	S	A	V	D	I	V	A	117																																	
HVDG_ECOLI	2	T	N	V	L	V	D	E	P	A	N	K	I	L	G	A	L	E	B	E	M	V	O																																																																															

Figure 2 D

AFQ1_STRCO 137RSAMTVTKNGEDLQLTPTTELRLILLELSRRPGQALSROQLRLVWEHDYLGDSRLVDAVQRL 198
ARCA_ECOLI 148NSRSLIGPDGEQYKLPSEFRAMLHFCENPGKIQSRAELLKKMTGRELKPHDRTVDTVIRRI 209
ARCA_HAEIN 147NSHSLITPEGQEFKLPSEFRAMLHFCENPGKIQTRSELLKKMTGRELKPDRTVDVTVIRRI 208
BAER_ECOLI 137 QQDAESPLIIDEGRFQASWRGKMLDLTPAEFRLLKTLISHEPGKVFSSREQLLNHLVDDYRVVTDRTIDSHIKNL 209
BASR_ECOLI 126 SELIVGNLTLMGRRQVWMGGEELILTPKEYALLSRLMLKAGSPVHREILYNDIYNWDNEPSTNTLEVHIHNL 198
BASR_SALTY 126 SELTVGNLTLMIGRHQAWRDGQELTLTPKEYALLSRLMLKAGSPVHREILYNDIYNWDNEPSTNTLEVHIHNL 198
CADC_ECOLI 5 PVRVGEWLVTPSINQISRNGRQLTLEPRLIDLLVFFAQHSGEVLSRDELIDNVVKRSIVTNHVVTQSISELR 77
COPR_PSESM 127 TSLQIGDLQVDLLKRRATRGGKRIELTAKEFALLELLMRRQGEVLSKSLIASQVWDMNFDSDTNVLEVAIRRL 199
CPXR_ECOLI 133 PTLEVDAVLNPNRQEQASFDGQTELELTGTEFTLLYLLAQHLGQVVSREHLSQEVLGKRLTPFDRADMHISNL 205
CPXR_HAEIN 130 EILSFDGTLHFHSHGIATYNEENLNTDYEFKILCLLLKSKGNVVSREELSLEVMEKPLTPFDRSMDHISNL 202
CREB_ECOLI 131 PVIRIGHFELNEPAAQISWFDTPALATRYEFLLLKTLKSPGRVWSRQQLMDSVWEDAQDQTYDRTVDTHIKTL 203
CUTR_STRLI 126 PVLERAGIKLDPNRREVFVRDGEVQLAPKEFAVLEVLMRSEGAVVSAEQLEKAWDENTDPTFNTVVRVTVMTL 198
EPIQ_STAEP 116FENHQVFVNNYLVNLSNIEFKILRCLYINLGRYVSKEELKKGVDWTEDFVDSNTINVIHRL 177
GLNR_STRCO 126 MEIRNGDLSVDEATYSAKLKGRVLDLTFKEFELLKYLAQHPGRVFTRAQLLQEVWGYDYFGGTRTVDVHVRRL 198
IAGA_SALTI 36NIPPKEYAVLVILLEAAGKIVSKNTLLDQVWGDAEVNEESLTRCIYALR 84
IAGA_SALTY 36NIPPKEYAVLVILLEAAGEIVSKNTLLDQVWGDAEVNEESLTRCIYALR 84
KDPE_ECOLI 128 PLVKFSDVTVDLAARVTHRGEVHVTPIEERLAGRCSTMPEKYSVSPGPVNLQVWGNVAVESHYLRITVMGHL 200
NISR_LACLA 134 IRRDLGPIIFYLEERRVCVNGQTIPLTCREYDITIELSQRTSKVYTREDITYDDVYDEYSNALFRSISEYIYQI 206
OMPR_ECOLI 137 AVIAFGKFKLNLGTREMFREDEPMPLTSGEFAVLKALVSHPREPLSRDKLMNLARGREYSAMERSIDVQISRL 209
OMPR_SALTY 137 AVIAFGKFKLNLGTREMFREDEPMPLTSGEFAVLKALVSHPREPLSRDKLMNLARGREYSAMERSIDVQISRL 209
PETR_RHOCA 145LDRGELSQGDQPVRLTATEAALMRIFAAHAGEVIGRTEL.....EAAGDRAVDVQITRL 198
PHOB_ECOLI 131 EVIEMQGLSLDPTSHRVMAGEEPEMGPTEFKLLHFFMTHPERVYSREQLLNHVWGTNVYVEDRTVDVHIRRL 203
PHOB_HAEIN 129 QFIQIDELSIDENAQRVFFQQQEIINLSSTEFKLLHFFMRHPEKVYSREQLLNRIWHNDLEVEYRTVDSYIRRL 201
PHOB_KLEPN 131 EVIEMQGLSLDPSHRVMTGDSPLDMGPTEFKLLHFFMTHPERVYSREQLLNHVWGTNVYVEDRTVDVHIRRL 203
PHOB_PSEAE 132 APIEVGGLLLDPISHRVTIDGKPAEMGPTEYGLLQFFMTHQERAYTRGQRDQVWGNVYVEERTVDMVIRRL 204
PHOB_SHIDY 131 EVIEMQGLSLDPTSHRVMAGEEPEMGPTEFKLLHFFMTHPERVYSREQLLNHVWGTNVYVEDRTVDVHIRRL 203
PHOB_SHIFL 131 EVIKMQGLSLDPTSHRVMAGEEPEMGPTEFKLLHFFMTHPELVYSREQLLNHVWGTNVYVEDRTVDVHIRRL 203
PHOP_BACSU 138 GQIVIGDLKILPDHYEAYFKESQLELTPKEFELLYLGRHKGRVLTDRLLSVAWNYDFAGDTRIVDVHISHL 210
PHOP_ECOLI 126 QVISLPPFQVDLSRRELSINDEVIKLTAFEYTIMETLIRNNGKVVSKDSLMLQLYPDAELRESHTIDVLMGRL 198
PHOP_SALTY 127 QVINIPPFQVDLSRRELSVNEEVIKLTAFEYTIMETLIRNNGKVVSKDSLMLQLYPDAELRESHTIDVLMGRL 199
RCAC_FREDI 126 PLLTWGDLNLPSTCEVYNGCPLNLTMEYDLELLLRNCQHVFSSEELDKLWSSSEDFPSEATVRSHVRL 198
RESD_BACSU 139 NVLVFSHLSIDHDAHRVTADGTEVSLTPKVYELLYFLAKTPDKVYDREKLLKEVWQYEFFGDLRTVDVHVKRL 211
SPAR_BACSU 126 SKRVISGFLFHFDSKEVFINNKNLNTKNEYKICEFLAQHKGRFTFSREQIYEEIYGLEGNALYSTITEFIRTI 198
SPHR_SYNP7 161 AVLRYEGLKLFPEECRVLDDRELTLSPKEFRLLLELFMRHPRRVWSRDQLEKIWIDFMGDSKTIDVHIRWL 233
TCTD_SALTY 127 ..VQQLGELIFHDEGYFLQGPALALTPREQALLTVLMYRRTRPVSRQQLFEQVFSLNDEVSPESITELYIHL 197
TORR_ECOLI 134 NLYRFAGYCLNVSRLTLERDGEPIKLTAEYEMLVAFVTPNPGELSRERLLRMLSARRVENPDLRTVDVILRR 206
TOXR_VIBCH 47RLGSNESRILLWLLAQRPNVIRSRNDL.....FEVDDSSLTQA.... 83
TOXR_VIBPA 35RLGSNESRILLMLAERPNEVLRNEL.....FEVDDSSLTQA... 71
VANR_ENTFC 133 NVIVHSGLVINVNTHCYLNEKQLSLTPTEFSILRILCENKGNVVSSELL.....FSKSNNTITVHTRHL 197
VIRG_AGRRA 143RQRRLISEEGGEIKLTAGEFNLLVAFLEKPRDVLSREQLLIASRVREEVYDRSIDVILIRL 204
VIRG_AGR5 155RRRRLISEEGSEVKLTAGEFNLLVAFLEKPRDVLSREQLLIASRVREEVYDRSIDVILIRL 216
VIRG_AGR6 169RQRRLMSEAGGEVKLTAGEFNLLVAFLEKPRDVLSREQLLIASRVREEVYDRSIDVILIRL 230
YC27_CYAPA 137 ENLQIGFLKIDINKRQVFKNGERIRLTGMEFSLLELLISKMGPEFSRAQI.....RHIDTRVVDVHISRL 201
YC27_GALSU 139 GIINIGFLKIDINKRQVYKNEERIRLTGMEFNLELLISNSGEPLSRRTTI.....RHLDRVVDVHISRL 203
YC27_PORAE 137 ..INIGFLKIDVNHQVYKNNERVRLTGMEFSLLELLISKAGQPFSTRATI.....RQVDTRVVDVHISRL 199
YCBL_BACSU 128 KVIRIHQLAIDIDNVSVLKNGEPLQLTSTEWQLLCLFASNPKKVFTKDQIYRSVWNEEYFDDQNIINVHMRL 200
YGIX_HAEIN 126 SVIEQAGVKLDQNRQSVVWLNQPISLTSREYKLLLELFMLNKDRVLSRSSIEEKLSSWDEEISSGALDVHIYNL 198
YXDJ_BACSU 131 KVVEYAGVQLFVERFELRFQDEKSELKSKKLELVLLERGEKVTSRDRLEKMTWDTDIFIDDNTLVVITRL 203
YYCF_BACSU 134 NEIHIGSLVIFPDAYVVKRDETTELTHREFFELHYLAKHIGQVMTREHLLQTVWGYDYFGDVRVTVIRRL 206

Figure 5

APMU_PIG	1062	CKEESP...VNVTV	RYNGCT..IKVEMARCVGEKKTIV..TYDYDIFQIKNSCL	COEEDYERFDIVLD	CPD	STL	PRY	RHTA	QGLD	PC	1145	
CE10_CHICK	281	CTTKTKSPSPVRF	TYAGSSVKKYRPKYC.GSNV	CTPQOTRIVKIRFR	CDDE	ETFTK	SVMM	IOSCR	NY	NC	354	
CGHB_HUMAN	29	CRPIN..ATTAAV	EKEGCVCIITVNTTII	ENYRDVRESIRLPG	CPRE	VNVP	SYAV	ADSC	QA	LC	113	
CGHB_PAPAN	29	CRPIN..ATTAAV	EKEGCVCIITVNTTII	CNYREVRRESIRLPG	CPRE	VNVP	SYAV	ADSC	QA	LC	113	
CTGF_HUMAN	256	LRTPKISKPKF	EISGCTSMKTYRAKFC	CTPHRTITLLPVEFK	CPDE	VWKK	NMF	IKTCA	CHY	NC	329	
CTGF_MOUSE	255	LRTPKIAKPKF	EISGCTSVKTYRAKFC	CTPHRTITLLPVEFK	CPDE	IMKKN	NMF	IKTCA	CHY	NC	328	
CYR6_MOUSE	284	SKTKKSPPEVRF	TYAGSSVKKYRPKYC	CTPLQTRIVKMRFR	CEDE	EMF	SKNVM	IQSC	CNY	NC	357	
FSHB_BOYIN	21	ELTNN..ITLTV	EKEEGFCISIMNTTW	CTFEKELVETVKVP	CAHADS	SLY	TPV	ATE	CHG	KG	105	
FSHB_HORSE	3	ZLTNN..ITLAV	EKEGCRFCITLNTTW	CTFEKELVETVKVP	CAHADS	SLY	TPV	ATE	CHG	KG	87	
FSHB_HUMAN	21	ELTNN..ITLAI	EKEEGRFCISIMNTTW	CTFEKELVETVKVP	CAHADS	SLY	TPV	ATE	CHG	KG	105	
FSHB_PIG	21	ELTNN..ITLTV	EKEENFCISLNTTW	CTFEKELVETVKVP	CAHADS	SLY	TPV	ATE	CHG	KG	105	
FSHB_RAT	22	ELTNN..ITLSV	EKEENFCISLNTTW	CTFEKELVETVKVP	CAHADS	SLY	TPV	ATE	CHG	KG	106	
FSHB_SHEEP	21	ELTNN..ITLTV	EKEESFCISLNTTW	CTFEKELVETVKVP	CAHADS	SLY	TPV	ATE	CHG	KG	105	
GTH1_COBAU	32	RLINN..MTITV	EREDCHG..SITITTCAGL	CNFKEMSKEVYLEG	CPSE	GNP	FFIP	VAK	SDCI	KG	113	
GTH1_ONCKE	32	RLINN..MTITV	EREDCHG..SITITTCAGL	CNFKEMSKEVYLEG	CPSE	GNP	FFIP	VAK	SDCI	KG	113	
GTH1_ONCMA	32	RLINN..MTITV	EREDCHG..SITITTCAGL	CNFKEMSKEVYLEG	CPSE	GNP	FFIP	VAK	SDCI	KG	113	
GTH1_THUOB	8	HEPKN..ISLSV	ES..GITEFLITTI	CNG.DMS	EVKHI	EG	CPV	...	VITV	ARN	ECT	AC
GTH2_ONCKE	29	COPIN..QVYSL	EKEGCVCIITVNTTII	CTYRDVREKEMIRL	CPD	PWD	PHV	TPV	VAL	SCD	GS	IC
GTH2_ONCMA	29	COPIN..QVYSL	EKEGCVCIITVNTTII	CTYRDVREKEMIRL	CPD	PWD	PHV	TPV	VAL	SCD	GS	IC
GTHB_MURCI	6	COPIN..ETLSV	EKDGCPKLVQTSICSG	CTYRDVREKEMIRL	CPD	PWD	PHV	TPV	VAL	SCD	GS	IC
GTHB_ONCTS	29	COPIN..QVYSL	EKEGCVCIITVNTTII	CTYRDVREKEMIRL	CPD	PWD	PHV	TPV	VAL	SCD	GS	IC
LSHB_COJTA	56	CRPIN..VTVAV	EKEEPCIMAVTITTA	CTYRDLRERWDLV	CPD	ESD	PVIL	PVAL	SC	GA	RC	140
LSHB_EDUAS	29	CRPIN..ATTAAV	EKEAPICITLTTTII	CTYRDLRERWDLV	CPD	ESD	PVIL	PVAL	SC	GA	RC	140
LSHB_HUMAN	29	CRPIN..ATTAAV	EKEAPICITLTTTII	CTYRDLRERWDLV	CPD	ESD	PVIL	PVAL	SC	GA	RC	140
LSHB_MELGA	48	CRPIN..VTVAV	EKDEPCIMAVTITTA	CTYRDLRERWDLV	CPD	ESD	PVIL	PVAL	SC	GA	RC	132
LSHB_PIG	29	CRPIN..ATTAAV	EKEAPICITLTTTII	CTYRDLRERWDLV	CPD	ESD	PVIL	PVAL	SC	GA	RC	113
LSHB_SHEEP	29	COPIN..ATTAAV	EKEAPICITLTTTII	CTYRDLRERWDLV	CPD	ESD	PVIL	PVAL	SC	GA	RC	113
MUB1_XENLA	301	CKEVPV..AIVGIG	qeydyqph	CKADRVEREKRAH	LV	QDNG	KKK	IYK	HIT	SC	CT	SC
MUC2_HUMAN	2170	GVTV..VITLV	SYAGCT..KTYIMNH	CKEKEKTSQRE	VLS	EN	SS	L	TH	Y	THE	SC
MUC5_HUMAN	917	GAVYH..RSLII	QOQSSSSSEPVRLAY	CKEELRTSLRN	VTLH	CTD	SS	RAF	SY	TE	VE	EG
MUC5_RAT	732	GSAID..VMKEI	SYNGCA..KNI	CKREERTSV	RMS	SLD	CPD	ES	KL	SH	S	Y
MUCS_BOYIN	471	GRSSS..VNTVY	NYNGCK..KKVEMAR	CKOENYER	E	RE	ID	LD	CPD	ES	KL	SH
NDP_HUMAN	39	QWRHHY.VDSISH	PLYKSS.KMWILLAR	CKRPTSKIKAL	RLR	CSG	EM	RL	T	AT	Y	RI
NDP_MOUSE	37	QWRHHY.VDSISH	PLYKSS.KMWILLAR	CKRPTSKIKAL	RLR	CSG	EM	RL	T	AT	Y	RI
NOV_CHICK	258	CIQTKKSMKAVRF	EYKNCSTSVQTYKPRYC	CTPHNTKI	IQVE	FR	CPD	ES	KL	SH	S	Y
NOV_COTJA	260	LRTRKSMKAVRF	EYKNCSTSVQTYKPRYC	CTPHNTKI	IQVE	FR	CPD	ES	KL	SH	S	Y
NOV_HUMAN	264	LRTRKSLKAHHL	EYKNCSTSLHTYKPRFC	CTPHNTKI	IQAE	FPQ	CS	EO	IV	KK	P	VM
SLIT_DROME	1409	GRKEQ...VREYY	TENDCRSRQPEKRYAK	CTPHNTKI	IQAE	FPQ	CS	EO	IV	KK	P	VM
TSHB_BOYIN	22	CLPTE..YMMHY	ERKECAYCLTINMTTV	CTYRDI	IVRR	KV	RMV	SN	NR	KY	IK	NL
TSHB_HUMAN	22	CLPTE..YTMHL	ERKECAYCLTINMTTI	CTYRDI	IVRR	KV	RMV	SN	NR	KY	IK	NL
TSHB_ONCMA	22	CYPTD..YTLYE	ERRECDVVALNMTTII	CTYDQVE	XR	TV	IL	PG	CL	H	AN	P
TSHB_PIG	22	CLPTE..YMMHY	ERKECAYCLTINMTTI	CTYRDI	IVRR	KV	RMV	SN	NR	KY	IK	NL
TSHB_RAT	22	CLPTE..YMMHY	ERKECAYCLTINMTTI	CTYRDI	IVRR	KV	RMV	SN	NR	KY	IK	NL
VWF_HUMAN	2724	QNDIT..ARLQYV	LVGSSCKSEVEV	CSPTIRTE	BMQ	VAL	HL	CTN	ES	VV	YH	E

Figure 6 A

7LES_DROVI	1917	S.YA.P.LPPLQILIEL.NAYGMWTA	PGT.....PDALSSLTTEC.QSIREQ.	IQFN.VAGNHT.QMRLAFLQKTRSCR	LAAYAATP....GAPI	1997	
APU_TRETY	1165	P.TAP.V.LOOPGI.ESSRVTINWSPSA...	DDVAIFGEIYK.SSEETGPF.....IKAT.	VSDSVY.NNYVDITVNGNYYKVV	ZVDTSYN....RTAS	1248	
AXO1_RAT	914	PRRP.GNISWTF..SSSLSLTKMDPVP	PLNENSTVTKMLY.QNDLHP	TPtLhltskmWIEP.VPEDIG.	HALVDIRTTGGGDDGIP.A	EHAIYRN....GTS	
CHIT_STRLI	142	P.SAP.GTPTASNT.TDTSVKISWSAAT...	DDKGVKNNDV.LR....DGA.KYAT.VTGT...	YDDNGITTKGAYSYSK	ARDTADQ....TGAS	
CPSF_CHICK	491	P.DPE.QSVRVTSY.GEDWAVTSWEAP	PF.dggMPITGELMER.KKKGSMRW...MKNLPE.VEPDT.	TYESTKMIEGVEMRF	AVNAIGV....SOPS	
CPSF_CHICK	784	P.GPP.QAVRWVME.WGSNNALIFEP	PKd.dGNALISGVTIÖK.ADRITMEWFVTL.EHSEPT.	RCVIVELVMGNRRFRVY	SNVCGT....SOP	
FAS2_SCHAM	530	P.SAV.LÖVKMVM.TATVTFRKFFG	Pgn.dggLPAIKYAVÖK.KODSÖGWEDALN.RTWVDS	PYLLENKIPQTRNFRFA	ÄONEVGF....GPM	
FINC_BOVIN	689	P.VVA.TSEVTEI.TASSFVSVSA...	SDTVSGRVEY.EISEEDE.....PÖYLD.LPSTAT	SVNLPLIPGRYIVNVY	EISEE....GEO	
FINC_BOVIN	780	P.DAP.PDPTVDQY.DDTSIVRWSRP...	RAPITGRIVY.SPSEVGSSTELN.LPEPAN	SVTLSDIQGVONITTY	AVERN....OES	
GUNB_CELFI	1511	I.DKE.SÖMQVTDY.ÖDNSISVRLPS...	SSPVTCRVTI.APKNGPGP.....SKTKT.VGPDQT	EMTEGLQPIVAVSVY	ÄQONÖN....GES	
GUNB_CELFI	651	P.TTP.GTPVATGY.TTVGASISWAAS	td.AGGSVAGVEI.YR.VÖGTTÖTLVGT.TTAA...	YIIRDITPGRAYSVYK	ÄKDVAÖN....VSAAS	
I12B_HUMAN	235	P.DPE.KNLÖLKLPLKNSRÖVEVSE	WEPDc.WSTPHSIESTE.CYOYÖK.SKREK.KDRYFT.	DKTSATVIRKANASISR	ÄDRRYS....SSWS	
IDUA_CANFA	547	P.GPV.TRIRALPL.TRGÖVLVWSD	EDERV..GSKCLMTYEIQF.SADGEVYDPIS.RKPSTFN	LHFVSPDITGAVSGSYR	ÄVDYWAR....GPF	
IDUA_HUMAN	548	P.GÖV.TRIRALPL.TÖGÖVLVWSD	EDERV..GSKCLMTYEIQF.SÖDGEKYDPVS.RKPSTFN	LHFVSPDITGAVSGSYR	ÄVDYWAR....GPF	
IL7R_HUMAN	1129	P.EAP.FDLVSVYREGANDFVVTINT	SHLqkRYVVKVLEMDVAYRÖEKDENKWTHVN.LSSTKL.	TLLÖRKIQPAAVYIQR	SIPDHYIKGfWSEWS	
ITB4_HUMAN	1127	L.GAP.ONPNAKA.GSRKIHNM	LP...SGKPMGVRKY.WIQDSESEAHLL.DSKVP.	SVELTNIYPCDYMKC	ÄYGAÖGE....GYS	
ITB4_HUMAN	1581	P.DIP.TRLVFSAL.GPTSLRWSQ	EP...CERPLÖGVSVEY.ÖLNGGELHRLN.IPNFAÖL	SVAVEDLIPNHSYVRV	ÄQSÖGEW....GRER	
KECK_HUMAN	436	O.TEP.PKVRIEGR.STTSLSVS	SIPp.QOSRVMKVEITR.KKGDSNSYN.VRRTEGF	SVTLDDIAPDITLVÖQ	ÄLTOEGQ....GAGS	
KEK5_CHICK	444	P.SAV.SIMHÖVSR.TVDSITTS	SÖPdq..PNGVILD.ELOQ.YEKNISELNSTA.VKSPTN.	TIVVÖNKAGITVÖQR	ÄRTVAGY....GRYS	
KMLC_CHICK	60	P.DPE.AGTPCASDI.RSSSLTSM	YGSSY.dgGSVAVOSYVEI.WNSVDNKWTDILT.TCRST.	SFNVODIQADREKFRV	ÄANVYGI....SEPS	
KSEK_MOUSE	441	P.SSI.ALVÖAKVEY.TRYSVATA	LEPDr..PNGVILEYVKY.YEKDÖNERSYR.IVPTAAR	NTDIKINPLTSTVFRV	ÄRTAAGY....GDFS	
LAR_DROME	322	P.TAP.TDVOISEV.TATSVRIEWS	YK...GPEILOVYVIOQ.KPKNANÖARSEI.SGILTM.	YVYVRAISPYTEEFYI	ÄVNNIGT....GPPS	
MPSF_CHICK	371	P.GAP.MDVKCHDA.NRDYVIVY	TKPnt.tsÖNPVIGYFVK.CEYGLENWÖQCN.DÄPVKIC	KYVYGTVEGRSYIFR	ÄVMSAGI....SRPS	
NCAL_BOVIN	509	P.SSP.SIDÖVER..YSSTAÖQ	CEPEa.tGGVPIIRYKAEMR.AMGEVWhskWYD.AKESMS	egivIVGLKPEITVAVR	ÄLANGKGL....GEIS	
NCAL_BOVIN	928	P.SPP.SELKITNP.TLDSITLEM	GSPTh..PNGVLTSTILKE.OPINNTHeLgplVHIR.IPANES	SLIKKINISTRYKFEYN	ÄÖTSV....GSGS	
NRC4_CHICK	344	G.SAP.TGLAVIAT.TSTVSIS	WNAV...ANASSYGV.YR....NGSKVGS.AIATA.	YDSSGLIAGITYSYVT	ÄADPTAG....esÖPS	
PHB_ALCFA	123	P.DPE.SNLSVÖVR.SGKNAILI	WSPt...OGSYTAEKIKV.LGISEASSSYNRTFQ.VNDNTE.	ÖHSVKEILTPGATVÖVAY	ÄTYDG....KES	
PTP1_DROME	554	P.AÖV.TDLHVANÖGMTSSLF	TNQ.A...ÖGDVEE.ÖVLL.IHENNVIKNES.ISSETS.	RYSHSLKSGLSVVT	ÄTVSSGG....IS	
PTPB_HUMAN	290	P.PREIAPPOLIGV.GPTYILL	ÖLNANSI.IGDGPIIKKEVEYR.MT.SGSWTETh.AVNAP.	TYKLMHIDPDEIEIRV	ÄLTPRGEgg....tGLPG	
PTPK_MOUSE	593	V.SPD.TELTVNV.TDKTVNE	WKEH...NVNVEVLVTV.VPTSSGGLDLOFT.VPQNOT.	SATTHEIEGVYFIRV	ÄILKN....KKS	
TENA_CHICK	446	P.PVPIAAPRLITK.ÖSRÖLV	SPLSFS.GDGPISVRLIHYRPODSTMDWSTIV.VDPSE.	NVTILNLRPKTGSVRVÖL	ÄSRPGEgg....eGAGH	
TIE1_HUMAN	444	L.PKPLINAPNVIDT.GHNF	AVINISSEPY.fGDGPIKSKLLYKPVNHYYEa	WÖHIÖ.VTNEI.	VTLNLYEPTREYELCQ	ÄIVRGEgg....GFGH
TIE2_HUMAN	639	P.POP.ENIKISNTI.THSSAVI	SWTILD...GYSISSITIRY.KVÖGKNEDÖHVDVKIKNA	TII.ÖYÖLKGLEPEITAY	ÄÖVDIF.ÄNNIGS....SNPÄ	
UFO_HUMAN	327	L.GPP.ENISATR..NGSÖAF	HVHÖEPRa.pLÖGTLIGYRLAYÖGÖDTPEVLMDI.GLRÖEV.	TLELÖGGDGSVSNLTVCA	ÄYTAAGd....GPM

Figure 6 B

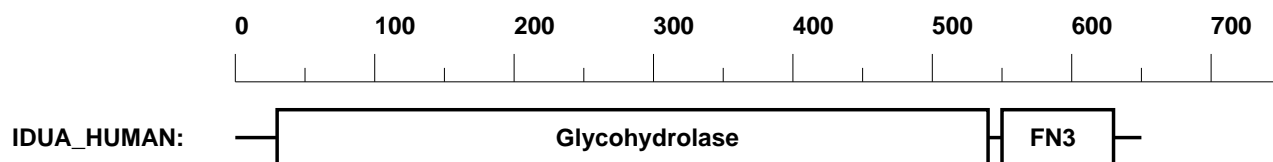


Figure 7

AGRI_CHICK	154	CVC PAS	CS	Gva . ESI VCGS D G K D Y R S E C D L N K H A C	DK	QEN V F K K F D C A C	201	
AGRI_RAT	165	CLC P T T	CF	Gap . D G T V C G S D G V D Y P S E C Q L L S H A C	AS	Q E H I F K K F N C P C	212	
FSA_HUMAN	116	CVCAPD	CS	NitwKG P V C G L D G K T Y R N E C A L L K A R C	KE	Q P E L E V Q Y Q C R C	164	
FSA_PIG	116	CVCAPD	CS	NitwKG P V C G L D G K T Y R N E C A L L K A R C	KE	Q P E L E V Q Y Q C K C	164	
FSA_RAT	116	CVCAPD	CS	NitwKG P V C G L D G K T Y R N E C A L L K A R C	KE	Q P E L E V Q Y Q C K C	164	
FSA_SHEEP	109	CVCAPD	CS	NitwKG P V C G L D G K T Y R N E C A L L K A R C	KE	Q P E L E V Q Y Q C K C	157	
IAC1_BOVIN	14	CKVYTEA	CT	RE . Y N P I C D S A A K T Y S N E C T F	CNEKM . NN	D A D I H F N H F G E C	61	
IAC2_BOVIN	7	CAEFKDP	KVYCT	RE . S N P H C G S N G E T Y G N K C A F	CKAVM . KS	G G K I N L K H R C K C	57	
IACA_PIG	7	CNVYRSH	LFFCT	RQ . M D P I C G I N G K S Y A N P C I F	CSEKG . LR	N Q K F D F G H W C H C	57	
IACS_PIG	12	CDVYRSH	LFFCT	RE . M D P I C G T N G K S Y A N P C I F	CSEKL . GR	N E K F D F G H W C H C	62	
IAC_MACFA	33	CARYQLPG	CP	RD . F N P V C G T D M I T Y P N E C T L	CMKIR . ES	G Q N I K I L R R C P C	81	
IOV7_CHICK	94	CSPYLQVVRDGNtMVA	CP	RI . L K P V C G S D S F T Y D N E C G I	CAYNA . EH	H T N I S K L H D G E C	150	
IOVO_ABUPI	8	CSDHPKP	ACL	QE . Q K P L C G S D N K T Y D N K C S F	CNAV V . DS	N G T L T L S H F C K C	56	
IOVO_ALECH	6	CSEYPKP	ACT	LE . Y R P L C G S D S K T Y G N K C N F	CNAV V . ES	N G T L T L S H F C K C	54	
IPSG_VULVU	68	CTEYSDM	CT	MD . Y R P L C G S D G K N Y S N K C I F	CNAV V . RS	R G T I F L A K H G E C	115	
IPST_ANGAN	12	CGEMSAMHA	CP	MN . F A P V C G T D G N T Y P N E C S L	CFQRQ . NT	K T D I L I T K D D R C	61	
IPST_BOVIN	9	CTNEVNG	CP	RI . Y N P V C G T D G V T Y S N E C L L	CMENK . ER	Q T P V L I Q K S C P C	56	
IPST_PIG	9	CTSEVSG	CP	KI . Y N P V C G T D G I T Y S N E C V L	CSENK . KR	Q T P V L I Q K S C P C	56	
IPST_SHEEP	9	CTNEVNG	CP	RI . Y N P V C G T D G V T Y A N E C L L	CMENK . ER	Q T P V L I Q K S C P C	56	
OATP_HUMAN	439	CNVDCN	CPs	KI . W D P V C G N N G L S Y L S A C L A	GC	E T . S I	G T G I N M V F Q N C S	485
OATP_RAT	439	CNTRCS	CS	TNt . W D P V C G D N G V A Y M S A C L A	GCKKFV . GT	G T N M . V F Q D C S C	486	
PE60_PIG	37	CEHMTESPD	CS	RI . Y D P V C G T D G V T Y E S E C K L	CLARI . EN	K Q D I Q I V K D C E C	86	
PGT_RAT	444	CRRDCS	CP	DSf . F H P V C G D N G V E Y V S P C H A	GC	S S	T N T S S E A S K E P I	488
PSG1_MOUSE	33	CHDAVAG	CP	RI . Y D P V C G T D G I T Y A N E C V L	CFENR . KR	I E P V L I R K G C P C	80	
QR1_COTJA	466	CICQDPA	ACPs	tKD . Y K R V C G T D N K T Y D G T C Q L F G T K C Q L E G T K M	GRQLHLDYMGAC	521		
SC1_RAT	424	CVCQDPET	CPp	aKI . L D Q A C G T D N Q T Y A S S C H L F A T K C M L E G t K K	GHQLQLDYFCAC	479		
SPRC_BOVIN	93	CVCQDP . TS	CPap . iGE	FEK V C S N D N K T F D S S C H F F A T K C T L E G t K K	GHKLHLDYICPC	149		
SPRC_CAEL	74	CECISK	CPeldgDP	MDK V C A N N N Q T F T S L C D L Y R E R C L C K R . K S k e c s k a f N A K V H L E Y L C E C	135			
SPRC_MOUSE	92	CVCQDP . TS	CPap . iGE	FEK V C S N D N K T F D S S C H F F A T K C T L E G t K K	GHKLHLDYICPC	148		
SPRC_XENLA	90	CVCQDPST	CPts . vGE	FEK I C G T D N K T Y D S S C H F F A T K C T L E G t K K	GHKLHLDYICPC	146		