BMC Bioinformatics

Research article

Open Access

A comparison of Pfam and MEROPS: Two databases, one comprehensive, and one specialised.

David J Studholme*, Neil D Rawlings, Alan J Barrett and Alex Bateman

Address: Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

Email: David J Studholme* - ds2@sanger.ac.uk; Neil D Rawlings - ndr@sanger.ac.uk; Alan J Barrett - ab9@sanger.ac.uk;
Alex Bateman - agb@sanger.ac.uk

* Corresponding author

Published: 9 May 2003

BMC Bioinformatics 2003, **4**:17

Received: 20 March 2003
Accepted: 9 May 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/17

Abstract

Background: We wished to compare two databases based on sequence similarity: one that aims to be comprehensive in its coverage of known sequences, and one that specialises in a relatively small subset of known sequences. One of the motivations behind this study was quality control. Pfam is a comprehensive collection of alignments and hidden Markov models representing families of proteins and domains. MEROPS is a catalogue and classification of enzymes with proteolytic activity (peptidases or proteases). These secondary databases are used by researchers worldwide, yet their contents are not peer reviewed. Therefore, we hoped that a systematic comparison of the contents of Pfam and MEROPS would highlight missing members and false-positives leading to improvements in quality of both databases. An additional reason for carrying out this study was to explore the extent of consensus in the definition of a protein family.

Results: About half (89 out of 174) of the peptidase families in MEROPS overlapped single Pfam families. A further 32 MEROPS families overlapped multiple Pfam families. Where possible, new Pfam families were built to represent most of the MEROPS families that did not overlap Pfam. When comparing the numbers of sequences found in the overlap between a MEROPS family and its corresponding Pfam family, in most cases the overlap was substantial (52 pairs of MEROPS and Pfam families had an intersection size of greater than 75% of the union) but there were some differences in the sets of sequences included in the MEROPS families versus the overlapping Pfam families.

Conclusions: A number of the discrepancies between MEROPS families and their corresponding Pfam families arose from differences in the aims and philosophies of the two databases. Examination of some of the discrepancies highlighted additional members of families, which have subsequently been added in both Pfam and MEROPS. This has led to improvements in the quality of both databases. Overall there was a great deal of consensus between the databases in definitions of a protein family.

Background

As the ever-growing number of genome sequencing projects reach completion, numbers of protein sequences in the primary sequence databases such as SWISSPROT [1] and GenBank [2] grow at an increasing rate. However, many of the newly deposited sequences are clearly homologous to previously known sequences, resulting in the need for strategies to classify sequences into clusters or families related by sequence similarity. This need has led to the proliferation of so called 'secondary' databases, derived from the primary sequence databases but with value-added curation. These databases are invaluable for predicting the function of new sequences based on homology to previously characterised proteins.

Secondary databases such as InterPro [3] and Pfam [4] aim to be comprehensive in their coverage of protein families and classify proteins on the basis of sequence relationships. Similarly, databases such as SCOP [5] attempt to provide a comprehensive classification of all known three-dimensional structures. However, there are also several databases that specialise in a particular subset of protein families, for example GPCRDB [6] and CAZy [7]. The MEROPS database [8] provides a catalogue and a structure-based classification of peptidases (*i.e.* proteolytic enzymes or proteases). Peptidases are a large group of proteins, representing around 2% of all gene products, and are of particular importance in medicine and biotechnology [9].

Previously, Pfam was compared to SCOP [10]. The aim of that study was to investigate the similarities and differences between a protein family database based on structural similarity and another based on sequence similarity. In the present study, we wished to compare two databases based on sequence similarity, one of which (Pfam) aims to be comprehensive in its coverage of known sequences, and the other (MEROPS) specialises in a relatively small subset of known sequences.

The two databases use different methods to identify family members. MEROPS selects a type example, and identifies the peptidase unit within it, and then makes pairwise matches using any number of transitive relationships. In contrast, Pfam stores a hidden Markov model (HMM) profile constructed from a seed sequence alignment. Using the HMMER computer package [11], Pfam searches for matches to the HMMs. The threshold values used in the HMMER searches are chosen manually by the Pfam curators.

One of motivations behind this study was quality control. These secondary databases are used by thousands of researchers worldwide and influence their work, yet the database contents are not peer reviewed. Therefore, we hoped that a systematic comparison of the contents of Pfam and MEROPS would highlight missing members and false-positives leading to improvements in quality of both databases. An additional reason for carrying out this study was to explore the extent of consensus in the definition of a protein family. The InterPro project is helpful in this respect because it simultaneously displays content from several different databases. However, InterPro does not contain data from any of the specialised databases such as MEROPS.

Results and Discussion

We analysed the contents of Pfam release 7.8 (December 2002) and MEROPS release 6.1 (January 2003). The secondary databases Pfam and MEROPS attempt to classify protein sequences from the primary databases into families. Therefore they have to draw from an underlying primary database of protein sequences. The underlying primary sequence database for MEROPS6.1 was the NCBI non-redundant database (NR) as released on 1st November 2002. Pfam7.8 used an underlying sequence database, pfamseq7, which consisted of SwissProt 40 plus trEMBL 18 (released in September 2001). Before we could embark on a meaningful comparison between the two protein family databases, we had to determine the set of common sequences shared by both underlying primary databases. Therefore we attempted to map each sequence in pfamseq7 to a sequence in NR, using crc64 checksums [12]. We found that there were 571,017 (537,646 unique) sequences common to both pfamseq7 and NR. Only these sequences were included in our subsequent analyses. We excluded 13,923 sequences from pfamseq7 that were absent from NR and excluded a further 649,578 in NR that were absent from pfamseq7.

Overall correspondence between MEROPS and Pfam

Of 174 families described in MEROPS, 121 overlapped at least one Pfam family and 53 did not. About 51% (89 out of 174) MEROPS families overlapped exactly one Pfam family (see Table 1 Additional file: 1 and Figure 1) and 18% of MEROPS families (32 out of 174) overlapped more than one Pfam family (Table 2 Additional file: 2 and Figure 2). The remaining 53 MEROPS families (30%) did not overlap any Pfam family. Of the 5,049 protein families in Pfam, 157 overlapped exactly one MEROPS peptidase family (Figure 1, panel B). A further five Pfam families overlapped multiple MEROPS families.

Several MEROPS families had only a fairly small degree of overlap with a Pfam family (see Table 2), so we investigated these relationships further and made improvements to MEROPS and/or Pfam family where appropriate. For example, family S41 shared six of its 49 unique sequences with PF02692 (Interphotoreceptor retinoid-binding protein), which contained a further 125 unique sequences. In all of those six shared sequences, Pfam reported a short fragment match to PF02692. On closer inspection, it was clear that these matches to PF02692 represented false positives. This proposition that they were false positives was further supported by the fact that fragment matches overlapped matches to Smart [13] domain TSPc (tail specific protease). Therefore we raised the cut-off values in Pfam for PF02692 such that the false positives would be

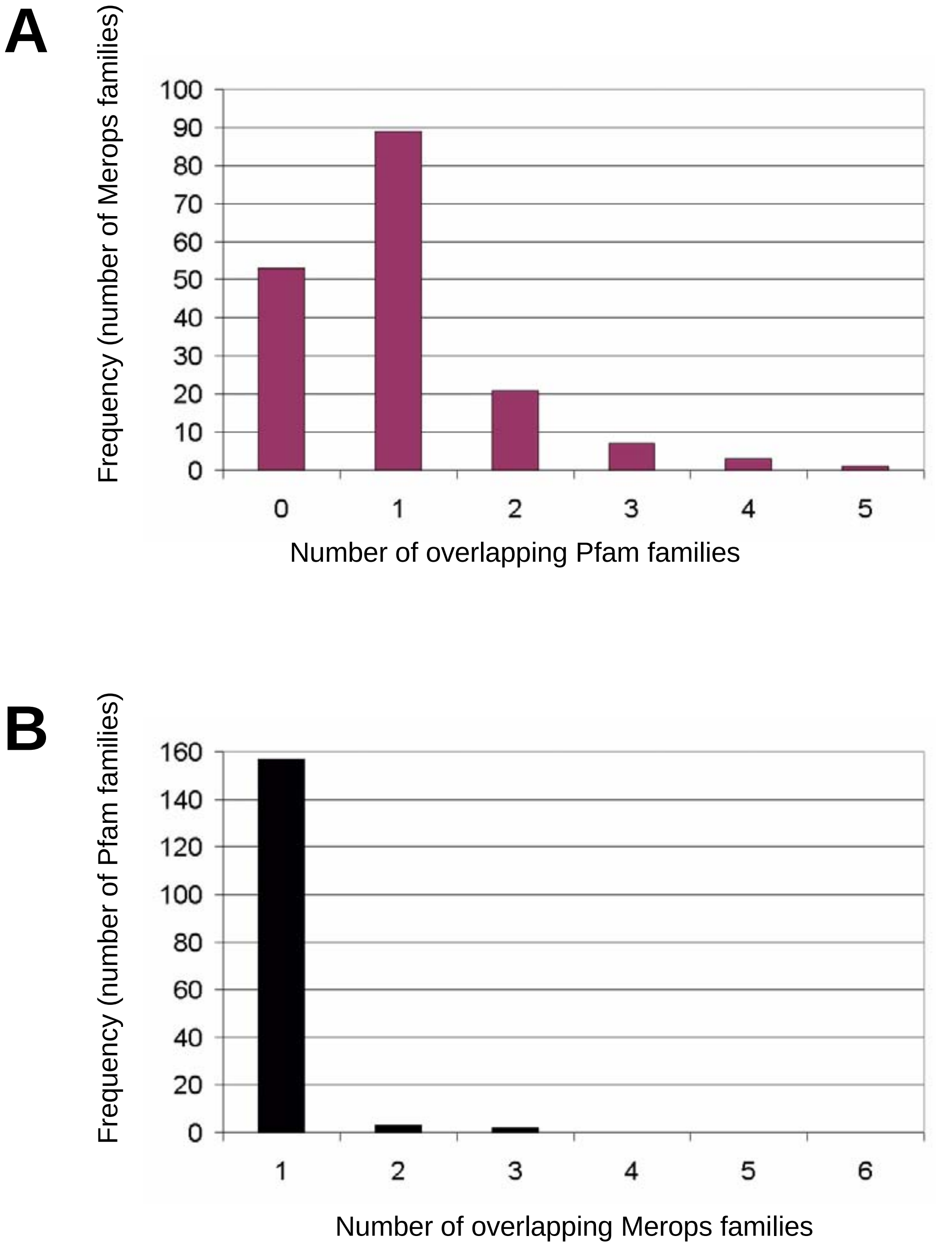

Figure 1
Frequency distribution of numbers of overlaps between MEROPS families and Pfam families. Panel A: For each of the 174 MEROPS families, we counted the number of Pfam families that the MEROPS family overlapped. See the text for definition of an overlap. The frequency (*i.e.* number of MEROPS families) was plotted for each number of overlaps. Panel B: For each of the 162 Pfam families that overlapped at least one MEROPS family, we counted the number of MEROPS families that the Pfam family overlapped.

excluded. Also, we extended the seed alignment for another Pfam family, PF03572, such that it will now correspond closely with MEROPS family S41.

MEROPS family A2 shared 648 unique sequences with PF00077 (Retroviral aspartyl protease), and it is clear that these two families are attempting to represent the same entity. However, there were a further 8008 sequences included in PF00077 but apparently absent from MEROPS (Table 1). These 8008 accession numbers have not been listed in MEROPS because they represent only minor variations (*i.e.* varying by only one or a few residues) of se-

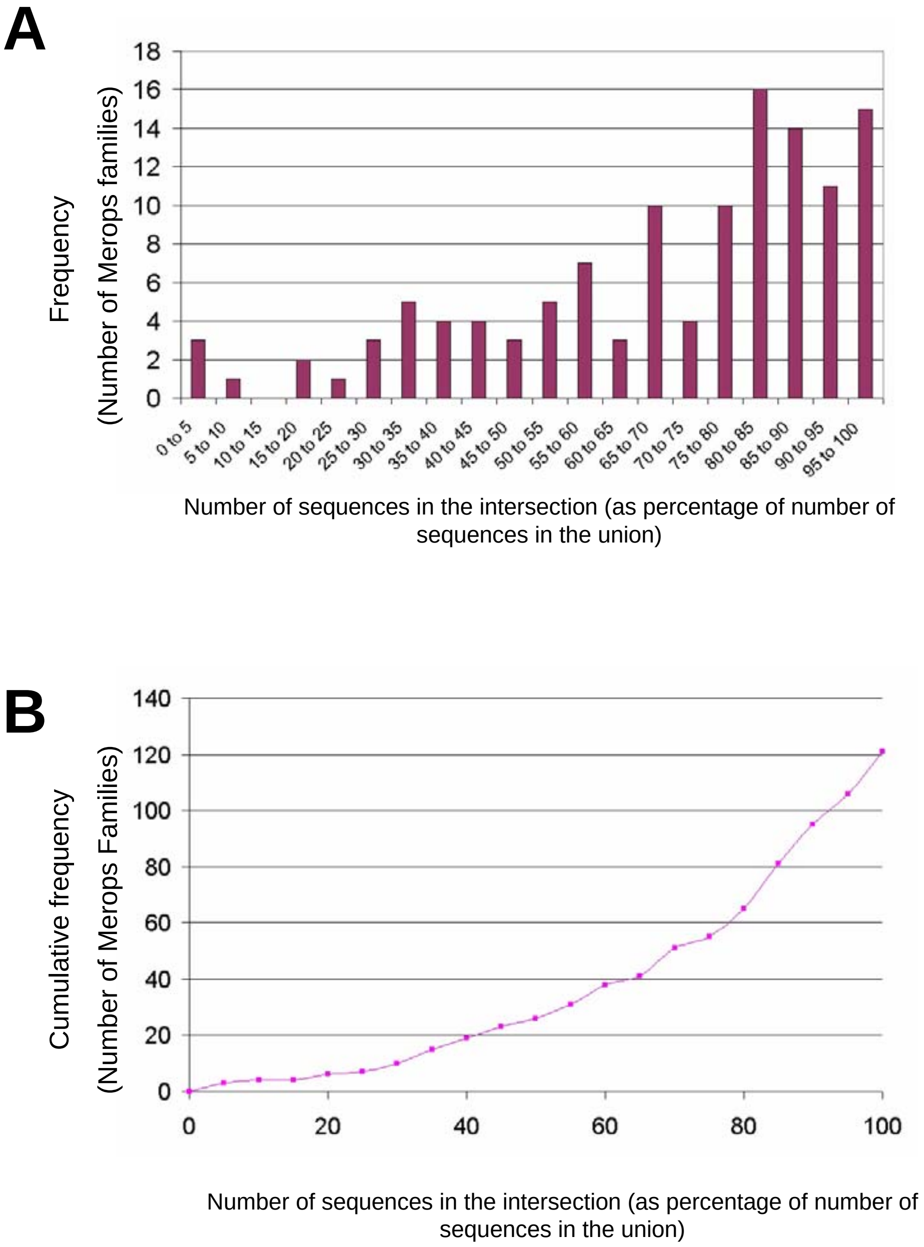

Figure 2

Frequency distribution for sizes of overlaps between MEROPS families and their corresponding Pfam families. For each of the 121 MEROPS families that overlapped at least one Pfam family, the number of unique sequences in the intersection was counted and expressed as a percentage of the total number of unique sequences in the union of the two sets. Panel A shows the frequencies (*i.e.* numbers of MEROPS families) for each 5% interval. The cumulative frequencies are shown in Panel B.

quences that are included in MEROPS family A2. This discrepancy highlights an important difference between the two databases in what they 'regard' as a distinct protein. Whereas Pfam relies entirely on SWISSPROT/trEMBL and treats each sequence accession as a distinct sequence entity, in MEROPS the curators usually use the following criteria: splice variants and sequences with greater than 95% identity are not considered to be separate proteins unless there is evidence that they are encoded by different genes (or come from different species). For RNA virus sequences, a lower threshold identity is often used.

We attempted to build HMMs to represent the 53 MEROPS families that did not overlap a Pfam family. We were able to build new HMMs to represent most (36) of these families in Pfam (see Table 3 Additional file: 3). However, in a few cases it was not possible to build a new Pfam family because the number of member sequences was very small (*e.g.* A18) or because sequence similarity to existing Pfam families (*e.g.* A11) led to violation of Pfam's rule against allowing overlaps between families [14].

Pfam contained several families annotated as having peptidase activity that did not have corresponding families in MEROPS. Most of these, including PF02338, PF04096, PF03926, PF04228 and PF04298 were deliberately excluded from MEROPS because there is insufficient experimental evidence for peptidase activity in these putative proteases. Another example was PF00905 (Penicillin binding protein transpeptidase domain). This was excluded because MEROPS does not attempt to comprehensively include transpeptidases. These differences reflect the different aims of the two databases: whereas Pfam attempts to comprehensively include all closely-related sequences within its families, in MEROPS the emphasis is much more upon biological significance of the member sequences.

How well do MEROPS families match Pfam families?

Using set theory as our approach, we quantified the closeness of a match between each MEROPS family and its corresponding Pfam family. For each MEROPS family and its closest matching Pfam family, we compared the number of members in the intersection (*i.e.* those sequences belonging to both the Pfam family and to the MEROPS family) against the number of members of the union (*i.e.* all those sequences belonging to either family) and expressed this ratio as a percentage (Tables 1 and 2). On average, the number of sequences in the intersection was 70% of the number of sequences in the union (Figure 2). The distribution was clearly skewed towards larger intersections, *i.e.* good matches between Pfam and MEROPS; out of 121 MEROPS families that intersected a Pfam family, 52 had an intersection size of greater than 75% of the union.

Family and sub-family levels in MEROPS

As illustrated in Figure 1 and listed in Table 2, thirty-two MEROPS families overlapped more than one Pfam family. MEROPS uses a hierarchical classification system. One aspect of this hierarchy is that families are grouped together into structurally related 'clans'. Also some families are further sub-divided into subfamilies. Among the families divided into subfamilies are A22, C1, C2, M10, M12, M14, M15, M28, M50, S1, S8 and S9, each of which overlaps multiple Pfam families. Therefore we investigated whether in these cases the Pfam families more closely corresponded to subfamilies or to families in MEROPS.

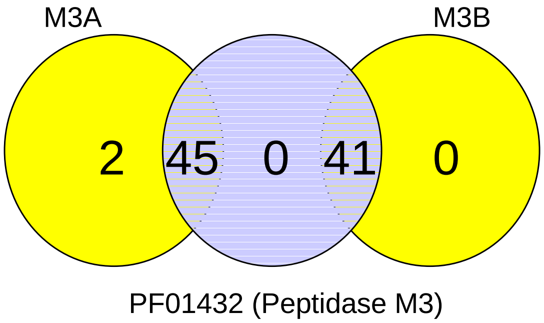

Figure 3
The relationships between Pfam family PF01432 and the MEROPS subfamilies M3A and M3B. MEROPS peptidase family M3 contains two subfamilies, M3A and M3B, and corresponds closely to Pfam family PF01432. There are 86 unique sequences that belong to both M3 and to PF01432. These sequences are roughly evenly distributed between the two subfamilies, with 45 belonging to MEROPS subfamily M3A and 41 belong to subfamily M3B. In other words the Pfam family PF01432 best matches MEROPS at the family level (M3) rather than at the subfamily (M3A, M3B) level.

Again, we used set theory and quantified the unions and intersections between each MEROPS subfamily and each Pfam family. In almost all cases we saw a poorer relationship at the subfamily level than at the family level. A typical example is illustrated in Figure 3. The members of PF01432 were roughly evenly spread between MEROPS subfamilies M3A and M3B, the two subfamilies that comprise family M3. Clearly PF01432 matched MEROPS M3 at the level of family rather than subfamily.

There were three exceptions to this generalisation, which are illustrated in Figures 4, 5, and 6. MEROPS family A22 overlaps both PF01080 and PF04258. Interestingly, when we look at the subfamily level, subfamily A22A overlaps PF01080, but not PF04258 (Figure 4). Conversely, subfamily A22B overlaps PF04258, but not PF01080. In other words PF01080 and PF04258 best match MEROPS at the subfamily level rather than at the family level. The second exception involves MEROPS family M12 (Figure 5). Family M12 intersects with both PF01400 and PF01421. At the subfamily level, M12A matches PF01400 better than PF01421, whilst M12B matches PF01421 better than PF01400. However, the relationships between the two

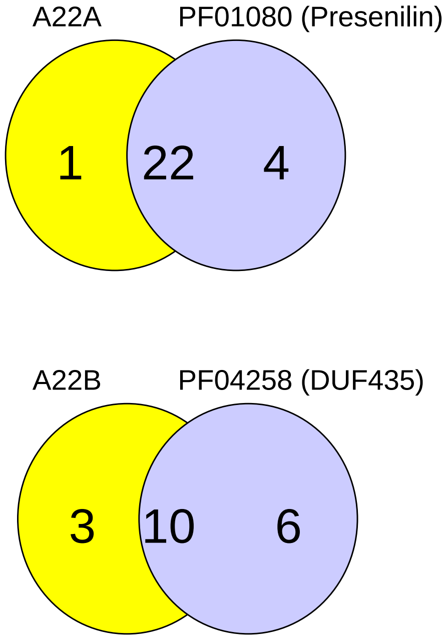

Figure 4
The relationships between Pfam families PF01080 and PF04258 with MEROPS subfamilies A22A and A22B. MEROPS peptidase family A2 contains two subfamilies, A22A and A22B, which exclusively overlap Pfam families PF01080 and PF04258 respectively.

MEROPS subfamilies and the two Pfam families are not clearly delineated; there are also small intersections between M12A and PF01400 and between M12B and PF01421. A similar scenario was observed in the comparison of MEROPS subfamilies C1A and C1B with PF00112 and PF03051 (Figure 6). Notwithstanding the examples of M3, A22 and C1, the overwhelming majority of MEROPS families corresponded most closely to Pfam at the family level rather than at the subfamily level.

Families with multiple overlap relationships

Several MEROPS families overlapped two or more Pfam families. With the exceptions of A22, C1 and M3, these multiple relationships could not be explained by a closer match at the subfamily level, so we investigated them further. For example, MEROPS family A2 overlaps five Pfam families (Table 2). On browsing the Pfam entries, it is clear that PF00077 (retroviral aspartyl protease) attempts to represent a domain that corresponds closely to A2, whilst the other four Pfam families represent other domains that are often found along with the peptidase unit in retrovirus polyproteins. Overlaps were reported due to discrepancies between the domain boundaries in Pfam and the peptidase units in MEROPS. For example, MEROPS6.1 reported the peptidase unit in O92805 to be between residues 30 and 121. According to Pfam7.8, however, the match to PF00077 was at residues 583 to 699. Furthermore, residues 2 to 87 matched PF02813 (Retroviral M domain) so that A2 and PF02813 were reported to overlap. On inspection of the sequence it was clear that the peptidase unit had been wrongly assigned in MEROPS, leading to erroneous reporting of the overlap.

In many cases, the discrepancies between Pfam domain boundaries and MEROPS peptidase units were not erroneous, but reflected differences in the design of the databases. Where there are several fragment sequences from viral polyproteins, MEROPS records the coordinates of the peptidase units with respect to the parent sequence rather than the fragment sequences. In contrast, Pfam maps domain boundary positions onto the individual fragment sequences, thus leading to discrepancies between the two databases with respect to domain boundaries. This scenario explains most of the multiple overlaps between families of viral proteases.

Nevertheless, we have now begun to introduce a process of checking and correcting the peptidase unit assignments in MEROPS. Similar situations explain most of the remaining cases where a MEROPS family has multiple overlaps. The exception is S49, which overlaps PF01343 and PF01957, the former of which represents MEROPS family S49 (formerly U7), and the latter includes a group of poorly characterised bacterial proteins of unknown function. Judging by alignments of these sequences in MEROPS, it appears that there is a close evolutionary relationship between these two Pfam families.

Five Pfam families (PF00004, PF02225, PF00851, PF00863 and PF00680) overlap multiple MEROPS families. In all five cases the reported overlaps could be explained by discrepancies in assignment of the peptidase unit in MEROPS in a similar manner to that involving MEROPS families with multiple overlaps.

Pfam family members absent from corresponding MEROPS family

As is clear from Tables 1 and 2, many Pfam families contained additional member sequences that were not found in the corresponding MEROPS families (Table 4 - See Additional file: 4). Most of this discrepancy could be explained by MEROPS and Pfam having different criteria for what they consider to be a different sequence, as discussed above. After eliminating discrepancies due to this difference, the majority of remaining sequences were clearly confirmed as members of existing peptidase families (on the basis of FastA and Blast searches) and have been subsequently added to MEROPS for inclusion in future releases. Most of these sequences would probably

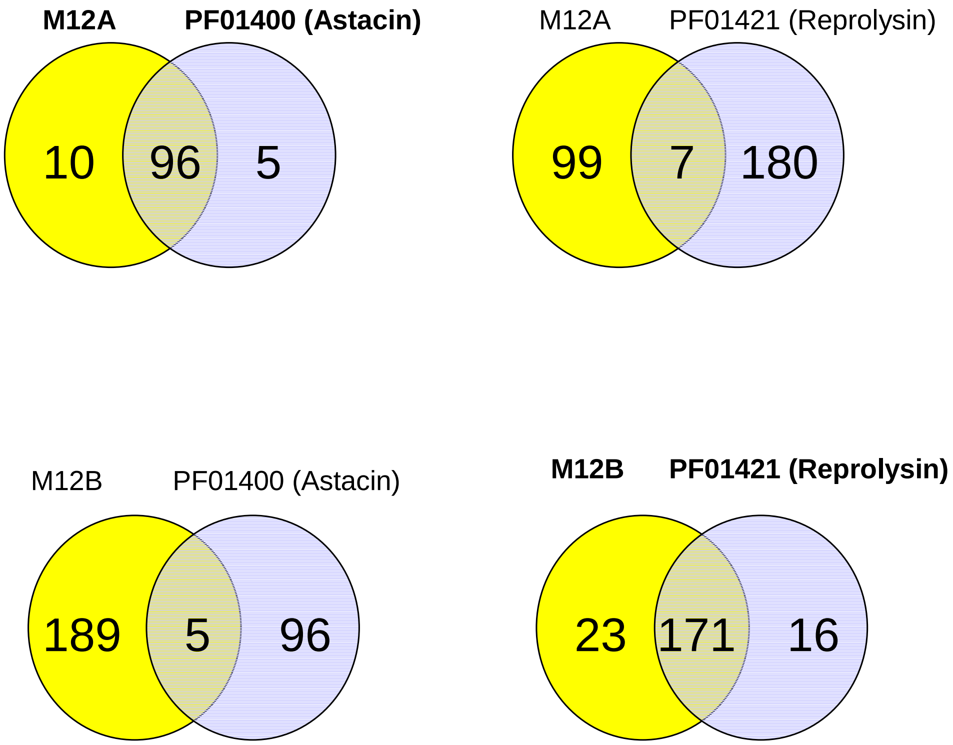

Figure 5
The relationships between Pfam families PF01400 and PF01421 with MEROPS subfamilies M12A and M12B. MEROPS peptidase family M12 contains two subfamilies, M12A and M12B, which both overlap Pfam families PF01400 and PF01421. The extent of the overlap between M12A and PF01400 is much larger than that between M12B and PF01400 and that between M12A and PF01421. Furthermore there is a very large degree of overlap between M12B and PF01421. In other words the Pfam families PF01400 and PF01421 approximately correspond to the subfamilies M12A and M12B rather than at the level of family M12.

have been picked up for inclusion in MEROPS by the curators' routine similarity searches. However, some of the sequences were previously not detected by MEROPS because they had a transitive relationship to the family type examples (see Methods). In other words there was no statistically significant direct relationship to the family type example identifiable by Blast or FastA. The sequences were indirectly linked to the type example *via* similarity to an intermediate sequence. About 180 sequences were added to MEROPS as a direct result of this study. However, a few sequences could not be identified as homologues by pairwise similarity searches against the MEROPS library of peptidase units, even by transitive links.

Sequences Q66541, Q66620, and Q20521 gave only partial or fragment matches to Pfam peptidase families. On closer inspection it is apparent that these sequences are fragments of multi-domain proteins lacking regions of the sequence defined as the peptidase unit in MEROPS. For this reason they were excluded from MEROPS.

Sequences O44472, O45151, O45157, P91466, P91467, P91515, P91519, Q09539, Q09393 and Q9N566 from *Caenorhabditis elegans*, and Q9VN01 from *Drosophila* belonged to PF01431 (Peptidase_M13) and contain the characteristic HEXXH zinc-binding motif [15]. Although PF01431 significantly overlapped MEROPS family M13, MEROPS did not include these sequences in family M13

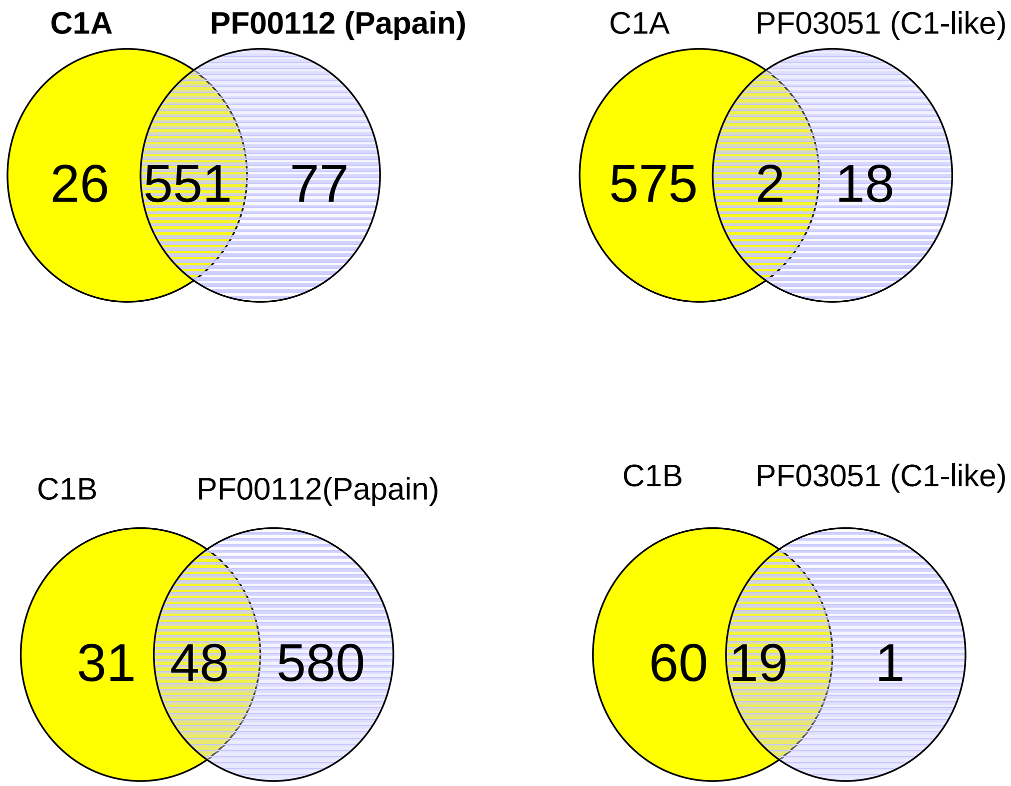

Figure 6

The relationships between Pfam families PF00112 and PF03051 with MEROPS subfamilies C1A and C1B. MEROPS peptidase family C1 contains two subfamilies, C1A and C1B, which both overlap Pfam families PF00112 and PF03051. The extent of the overlap between C1A and PF00112 is much larger than that between C1A and PF03051 and that between C1B and PF0012. In other words the Pfam family PF00112 approximately corresponds to the subfamily C1A rather than at the level of family C1.

since no similarity could be detected using FastA or Blast searches. It appears that the HMM representing Pfam family PF01431 is more sensitive and able to find additional homologues not detectable by Blast and FastA searches. Q9I304 belonged to Pfam family PF00227 (Proteasome), yet no similarity could be found between this sequence and members of the overlapping MEROPS family T1. This may be another example of a case where Pfam's HMM method for finding family members is more sensitive than MEROPS' Blast and FastA-based method.

Sequence P71878 shows a partial or fragment match to PF00814 (Peptidase_M22). Blast searches revealed that this *Mycobacterium tuberculosis* protein is related to 3-ketoacyl-CoA thiolase and acetyl-CoA C-acyltransferase, but not to any peptidases. It is possible that the match to PF00814 was a false positive in Pfam.

MEROPS family members absent from their corresponding Pfam families

There were 214 unique sequences that were classified into families in MEROPS but were absent from the corresponding Pfam families (Table 5 - See Additional file: 5). Just over half of these were sequences of protein fragments where the region or domain containing peptidase activity was missing. These fragments were treated differently in MEROPS as compared to Pfam. Whereas Pfam recognises only those sequence features that are actually present in the sequence, MEROPS assigns the fragments to families according to the properties of the complete parent sequence. A further 23 sequences were found to have been erroneously classified in MEROPS, and have subsequently been removed or moved to the correct family. Several SwissProt/trEMBL accessions that have had their sequences updated recently and so different versions of the sequences were found in Pfam *versus* MEROPS; this accounted for a further nine of the discrepancies.

Aside from the fragment sequences and trivial errors, 72 sequences were included in MEROPS but not in the corresponding Pfam families. In a few cases, such as Q9A9N9 and Q9PFX5, although no statistically significant similarity could be found between these sequences and the rest of the family, the MEROPS curators decided that these sequences should be included on the basis of expert knowledge.

In some cases, MEROPS identified statistically significant sequence similarities that Pfam had failed to detect. For example, four sequences (Q9A748, Q9RDK4, Q9PAC4, and Q9KM08) are included in MEROPS family M48 but were not included in PF01435. Inspection of the sequence alignment for M48 in MEROPS confirmed that these sequences were *bona fide* members of the family. As a result of this discrepancy we expanded the alignment for family PF01435 in Pfam and used this to rebuild the HMM and searched for new members of the family. As a result of this, the new HMM successfully identified Q9A748, Q9RDK4, Q9PAC4, and Q9KM08 as members of the expanded PF01435.

Family sizes and search sensitivity

Although most MEROPS families substantially overlapped Pfam families, there were some differences between the sets of MEROPS family members and the intersecting Pfam family members. These discrepancies have been examined in some detail in the previous paragraphs. One reason for there being an imperfect match between the Pfam family and the MEROPS family could be differences in sensitivity of family member-detection, in some cases at least due to the differing methods used to curate the two databases. This might be reflected in the relative sizes of the families between them. Therefore we compared the sizes (*i.e.* numbers of members) of each MEROPS family against each overlapping Pfam family. We found that in 19 cases, the MEROPS family was the same size as the Pfam family. In 95 cases, the Pfam family was larger than the MEROPS family. In the remaining 55 cases, the MEROPS family was larger than the Pfam family. In 34 cases, the MEROPS family was a proper subset of the Pfam family, and in 8 cases the reverse was true. This does not reveal any significant bias in the relative sizes of Pfam and MEROPS families. In other words there is no evidence that the HMM-based method is much more sensitive than Blast and FastA pairwise similarity search methods for detecting family members, at least in the context of Pfam and MEROPS.

The fact that MEROPS and Pfam are similarly successful at identifying family members is surprising at first sight, given that MEROPS is based on pairwise similarity searches whilst Pfam uses HMMs. There are two major factors that contribute to high sensitivity in detecting MEROPS family members, beyond that which might be expected from a simple pairwise search procedure. Firstly, a candidate sequence is searched against every existing peptidase sequence in the MEROPS database, not just the type examples. This enables identification of family members that are outliers in sequence space. Thus families can be iteratively expanded to include sequences that are only transitively linked to the homologous type examples via one or more intermediate members. Secondly, catalytic residues are often highly conserved within families of peptidases. This frequently helps the curators to confidently make a judgement about whether or not a particular sequence belongs to a given family when the degree of sequence similarity is relatively low. It should be noted that there is a significant amount of human intervention in the curation of MEROPS families. This certainly improves the quality and coverage over what could be achieved by completely relying on automated similarity searches.

This study revealed that Pfam and MEROPS are largely consistent with each other in terms of their classification of proteins into families. Nevertheless, MEROPS also contains additional features and data not present in Pfam. These features include a facility for BLAST searching against the peptidase sequence database, systematic data on active sites and substrates of peptidases and inhibitors, and a very comprehensive literature database. Perhaps the most important difference in this context is that MEROPS uses a hierarchical classification whilst Pfam uses a flat classification. To implement all of these features in Pfam would not be feasible given that Pfam aims to be comprehensive in its coverage of proteins. The fact that MEROPS is currently accessed by about 10,000 academic users per month and has previously sold several commercial licenses confirms the value of MEROPS to the scientific community as more than merely a subset of any more general database.

Conclusions

Since there is no peer review process for assessing the contents of the protein family databases, we carried out a systematic comparison of the contents of two widely used protein family databases with the intention of checking and improving the quality of data in both. As a result, about 35 new families have been added to Pfam. The numbers of members (*i.e.* sequence coverage) has been increased for several families by identifying false negatives. Furthermore, accuracy has been improved as a result of identifying false positives highlighted by this comparison.

Notwithstanding the differences identified between the contents of the two databases, overall there was a high degree of consensus between the two databases, despite their being independently curated using different methodologies and with different objectives. In particular the families defined in Pfam corresponded closely to the family level in the MEROPS hierarchy in most cases. This suggests that the families are accurately curated and that the lack of peer-review has not led to gross errors in family assignments. These databases also receive feedback from users with suggestions and improvements, which help to keep the quality of data high.

The methods developed here for systematically comparing the contents of the two databases will be used in future as routine quality control procedures in production of Pfam and MEROPS to highlight errors and to help refine family boundaries. Thus we hope that close cooperation yet independence will lead to continuing benefits to both databases.

Methods

Additional information about the Pfam and MEROPS databases can be found at their respective websites [16,17].

Blast and FastA searches for detection of MEROPS family members

Pairwise similarity searches were carried out using the routine MEROPS procedures. Data collection for the MEROPS database is done as follows. The best-characterised member in each peptidase subfamily is designated as the "type example", and the peptidase unit (that part of the sequence that bears the residues important for proteolytic activity, usually corresponding to one or two consecutive structural domains) of each type example was used as the query sequence in a BlastP search of the NR database from NCBI. In order to minimise redundancy, and to attempt to find known orthologues, each significant hit (e <= 0.001) from each BlastP search is used as a query sequence against the MEROPS collection of peptidase unit sequences. A sequence is appended to the MEROPS collection if it matches all the following criteria: (1) it is a significant hit from the FastA analysis (e <= 0.001), (2) it is less than 95% identical to an existing sequence in the collection, unless it is from a different species (but not a subspecies, strain or isolate) or is known to be the product of a different gene, and (3) it is not derived from an mRNA splice variant or alternative initiation. For any sequence failing to meet these criteria, only the database cross-references are added to the MEROPS database.

Comparison of Pfam versus MEROPS

We used a series of Perl [18] scripts to analyse overlap relationships between MEROPS families and Pfam families. We treated each MEROPS family and each Pfam family as a set of member protein sequences. We then compared every MEROPS family set against every Pfam family set, calculating the number of members in the union and the intersections of each combination. In order to qualify as an overlap, the relationship between the Pfam family and the MEROPS family had to satisfy both of two criteria:

(1) The intersection between the MEROPS family set and the Pfam family set must contain at least one member sequence, and (2) for at least one of the member sequences in the intersection, the matches to the MEROPS and the Pfam families must be co-linear. Co-linearity is defined as follows. When Pfam finds that a protein sequence matches a given HMM, it reports which region of the sequence contains that match. For example on the 309-residue long sequence of *Bacillus subtilis* sporulation σ^E factor processing peptidase SP2G_BACSU (P13801), Pfam identifies a match to family PF03419 covering residues 1 to 300. MEROPS identifies residues 148 to 309 to be the peptidase unit (family U4). Since the 1–300 and 148–309 regions overlap by more than 50% of each of their respective lengths, we consider these matches to be co-linear.

Building Pfam families

HMMs representing Pfam families were built using the standard Pfam procedures [4]. Pfam stores a hidden Markov model (HMM) profile constructed from a seed sequence alignment. Using the HMMER computer package [11] Pfam searches for matches to the HMMs. The threshold values used in the HMMER searches are chosen manually by the Pfam curators.

Abbreviations

HMM: Hidden Markov Model.

NR: The NCBI non-redundant sequence database

NCBI: National Centre for Biotechnology Information

Authors' contributions

DJS carried out the comparison between MEROPS and Pfam, including development of Perl scripts and examining discrepancies between the two databases, and drafted the manuscript. DJS and AB implemented changes and additions to Pfam arising from this study. NDR performed the Blast and FastA searches and implemented changes and additions to MEROPS arising from this study. AB conceived of the study, and participated in its design and coordination. AB, AJB, NDR and DJS all contributed to writing the final manuscript and interpretation of data.

Additional material

Additional file 1

Table 1 – Relationships between the two databases for MEROPS families that overlap a single Pfam family.

For each of the MEROPS families that overlapped exactly one Pfam family, the number of unique sequences in the intersection was counted and expressed as a percentage of the total number of unique sequences in the union of the two sets. The numbers of unique sequences found only in the MEROPS family and the numbers found only in Pfam family were also counted and expressed as percentages of the numbers of unique sequences in the union of both families.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-4-17-S1.doc]

Additional file 2

Table 2 – Relationships between the two databases for MEROPS families that overlap multiple Pfam families.

For each of the MEROPS families that overlapped more than one Pfam family, the number of unique sequences in the intersection was counted and expressed as a percentage of the total number of unique sequences in the union of the two sets. The numbers of unique sequences found only in the MEROPS family and the numbers found only in Pfam family were also counted and expressed as percentages of the numbers of unique sequences in the union of both families.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-4-17-S2.doc]

Additional file 3

Table 3 – MEROPS families for which there was no corresponding Pfam family.

We attempted to build new Pfam families to represent those peptidase families in MEROPS that did not substantially overlap any family in Pfam. Where we successfully created a new family, the Pfam accession number and long name is given.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-4-17-S3.doc]

Additional file 4

Table 4 – Protein sequences found in Pfam peptidase families but absent from the corresponding MEROPS family.

The sets of member sequences found in each family were compared as described in the text. Those sequences identified as members of a Pfam family but not included in the overlapping MEROPS family were examined in detail to determine the reasons for the discrepancies. As a result of this, many of the sequences were then considered to be bona fide *members of the family and were then added to the MEROPS database.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-4-17-S4.doc]

Additional file 5

Table 5 – Protein sequences found in MEROPS families but absent from the corresponding Pfam family.

The sets of members sequences found in each family were compared as described in the text. Those sequences identified as members of a MEROPS family but not included in the overlapping Pfam family were examined in detail to determine the reasons for the discrepancies. Several sequences had been wrongly included in MEROPS as a result of trivial errors (column 2). Some discrepancies could be explained by differences in the underlying sequence data where SwissProt/trEMBL where Pfam7.8 used an older (out of date) version of a SwissProt/trEMBL sequence. The remaining discrepancies could be explained by differences in the methods of building families between the two databases.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-4-17-S5.doc]

Acknowledgements

AB is funded by the Wellcome Trust. AJB, NDR and DJS are funded by the Medical Research Council, UK.

References

1. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003** *Nucleic Acids Res* 2003, **31:**365-370
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL **GenBank** *Nucleic Acids Res* 2003, **31:**23-27
3. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R and Zdobnov EM **The InterPro Database, 2003 brings increased coverage and new features** *Nucleic Acids Res* 2003, **31:**315-318
4. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer EL **The Pfam protein families database** *Nucleic Acids Res* 2002, **30:**276-280
5. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C and Murzin AG **SCOP database in 2002: refinements accommodate structural genomics** *Nucleic Acids Res* 2002, **30:**264-267
6. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE and Vriend G **GPCRDB information system for G protein-coupled receptors** *Nucleic Acids Res* 2003, **31:**294-297
7. Coutinho PM and Henrissat B **Carbohydrate-active enzymes: an integrated database approach** *In Recent Advances in Carbohydrate Bioengineering (Edited by: Gilbert HJ, Davies G, Henrissat B, Svensson B) Cambridge: The Royal Society of Chemistry* 1999, 3-12
8. Rawlings ND, O'Brien E and Barrett AJ **MEROPS: the protease database** *Nucleic Acids Res* 2002, **30:**343-346
9. Rawlings ND and Barrett AJ **MEROPS: the peptidase database** *Nucleic Acids Res* 1999, **27:**325-331
10. Elofsson A and Sonnhammer EL **A comparison of sequence and structure protein domain families as a basis for structural genomics:** *Bioinformatics* 1999, **15:**480-500
11. Eddy SR **Profile hidden Markov models** *Bioinformatics* 1998, **14:**755-663
12. Press WH, Teukolsky SA, Vetterling WT and Flannery BP *Numerical recipes in C Cambridge: Cambridge University Press* 1992,
13. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP and Bork P **Recent improvements**

to the SMART domain-based sequence annotation resource *Nucleic Acids Res* 2002, **30:**242-244

14. Sonnhammer EL, Eddy SR and Durbin R **Pfam: a comprehensive database of protein domain families based on seed alignments** *Proteins* 1997, **28:**405-420
15. Turner AJ, Isaac RE and Coates D **The neprilysin (NEP) family of zinc metalloendopeptidases: genomics and function** *Bioessays* 2001, **23:**261-269
16. **Pfam Home Page** [http://www.sanger.ac.uk/Software/Pfam/]
17. **MEROPS: the Protease Database** [http://merops.sanger.ac.uk]
18. Wall L, Christiansen T and Orwant J *Programming Perl Sebastopol: O'Reilly & Associates* 2000,