Pfam 10 years on: 10 000 families and still growing

Stephen John Sammut, Robert D. Finn and Alex Bateman

Submitted: 15th October 2007; Received (in revised form): 6th February 2008

Abstract
Classifications of proteins into groups of related sequences are in some respects like a periodic table for biology, allowing us to understand the underlying molecular biology of any organism. Pfam is a large collection of protein domains and families. Its scientific goal is to provide a complete and accurate classification of protein families and domains. The next release of the database will contain over 10 000 entries, which leads us to reflect on how far we are from completing this work. Currently Pfam matches 72% of known protein sequences, but for proteins with known structure Pfam matches 95%, which we believe represents the likely upper bound. Based on our analysis a further 28 000 families would be required to achieve this level of coverage for the current sequence database. We also show that as more sequences are added to the sequence databases the fraction of sequences that Pfam matches is reduced, suggesting that continued addition of new families is essential to maintain its relevance.

Keywords: Pfam; protein families; classification; coverage; hidden Markov model

INTRODUCTION
Ernest Rutherford famously and contemptuously said, ‘All science is either physics or stamp collecting’. But ‘stamp collecting’ or classification is of central importance in science. The periodic table of the elements in chemistry and the Linnaean binomial classification are both exceptional examples where classifying entities by certain characteristics allows a better understanding of those entities and further enables the prediction of new entities. For example, Mendeleev was able to use the periodic table to predict the existence and properties of Gallium and Germanium before they were discovered. The advent of high throughput sequencing and bioinformatics has enabled the classification of the proteins through the identification of sequences similarities they contain [1]. These similarities are often characteristic of shared protein domains, which can be considered as the common currency of protein structure and function. Classifications of proteins into groups of related sequences and domains are in some respects like a periodic table for biology allowing us to understand the underlying molecular biology of any organism.

Pfam is a large and widely used collection of protein domains and families [2]. The Pfam database is accessible via the Web (http://pfam.sanger.ac.uk) and available in several different downloadable formats. It was founded in 1995 by Erik Sonnhammer, Sean Eddy and Richard Durbin as a collection of common domains that could be used to annotate the complex proteins of multicellular animals [3]. Their work was partly motivated by Cyrus Chothia’s publication entitled ‘One thousand families for the molecular biologist’, in which the author hypothesized that there are around 1500 different protein families, and the majority of proteins come from no more than 1000 families [4]. Our scientific goal in the Pfam
group is the complete and accurate classification of protein families and domains. In this article we examine whether this is a realistic goal.

Since Pfam’s conception the content of a Pfam entry has remained largely unchanged. Briefly, for each entry in the Pfam database there are two multiple sequence alignments called 1) the seed alignment, which consists of a set of representative members of the family, and 2) a full alignment, which lists all the protein sequences that form part of the family. The sequences that are found in the full alignment are identified by searching a profile hidden Markov model (HMM), which is generated using the seed alignment and the HMMer program, against the UniProt sequence database [5]. Each entry also contains annotation and references to the structure and function of the family if known. Pfam is composed of two parts: when we normally talk about Pfam domains and families we mean those entries in Pfam-A that are of high quality and manually curated. However, there is a supplementary set of Pfam families called Pfam-B that are automatically generated from the ProDom database of protein domain families [6]. The Pfam-B families are generally of lower quality, but they can give hints towards the existence of new domains that might be added to Pfam.

Given that Chothia suggested that about 1500 families would be needed to catalogue biology it is somewhat surprising that the next release of Pfam (version 23.0) will contain over 10 000 families. Many of the families within Pfam are related to each other and represent only a single evolutionary origin. For example the ANTH domain (Pfam accession PF07651) and the ENTH domain (PF01417) have structural and functional similarities that argue their relatedness [7]. Pfam clans are sets of related Pfam families that have arisen from a single evolutionary origin, which has been confirmed by structural, functional, sequence and HMM comparisons (see http://pfam.sanger.ac.uk/clan?acc=CL0009). To date, Pfam (release 22.0) contains 283 clans that group together 20% of Pfam families. The Pfam clan classification is as yet incomplete and we can expect many of the families that are not yet in a clan to actually belong to one of the existing clans or to be part of new clans.

Using known structures to group families together into larger superfamilies, the SCOP database [8] has now identified almost 1600 distinct evolutionary families, and growth appears to be increasing in a linear fashion. The CATH structural database [9], as of version 3.1.0, contains 2091 homologous superfamilies (H level). This, combined with the fact that there will soon be over 10 000 families in Pfam, suggests that Chothia’s estimate of 1500 families is likely to be an underestimate. But with 10 000 families, how close is Pfam to reaching its scientific goal of a complete classification?

**PROGRESS IN COVERING SEQUENCE SPACE**

Pfam has always aimed to increase the number of families it contains over time, as well as finding as many homologous sequences to each family as possible. Progress has been monitored by looking at a measure called sequence coverage, which is the fraction of protein sequences listed in UniProt that have at least one Pfam domain and is currently 73.23%. A second measure that can also be used is called residue coverage, which is the fraction of the amino acid residues in UniProt that are within Pfam domains and is currently 50.79%.

The graph in Figure 1A shows how the sequence coverage has grown as we have added families to Pfam. As the first families were added to the database the coverage increased rapidly. These families were well-known, ubiquitous and important domains such as EGF, SH2, SH3, Fn3 and Ig. After the most common domains were added the new additions became smaller and smaller. A total of <2000 families are required for 50% sequence coverage, and a law of diminishing returns can be seen in Figure 1A. By extrapolation, we estimate that 184 472 or 519 940 families are required for 100% sequence or residue coverage, respectively with the current sequence database. However, there are good reasons to suppose that complete coverage is neither a realistic nor sensible goal.

When the 3-dimensional (3D) structure of a sequence has been determined, the domain organization of that sequence can often be clearly defined. However, there are also many cases where an unambiguous definition of domains is not possible and several alternatives exist. Particular complications arise when domains are composed of discontinuous segments of sequence. Sequences of domains are also frequently well biochemically characterized. Consequently, Pfam has always had a good coverage of the Protein Data Bank (PDB) [10], a database of known 3D protein structures. A recent drive to maximize
the Pfam coverage of sequences found in the PDB database provides a good estimation of the maximum coverage of the sequence database that is achievable. At Pfam release 22.0 there were 11,338 distinct protein sequences with at least some parts with a known 3D structure (and with a ‘mapping’ between the sequence in the PDB entry and a UniProt entry). Of these 10,519 (94.7%) match a Pfam domain. The residue coverage of PDB sequences by Pfam is 77.5%, much higher than that found across the whole of UniProt. We were unable to achieve 100% sequence coverage of the PDB sequences primarily due to one of the following three reasons: (i) some PDB sequences are absent from the UniProt sequence database upon which Pfam is built and therefore can not be represented (for example PDB:2od4 is a sequence from the Global Ocean Survey metagenomics dataset); (ii) some of the structural genomics targets have been focused on sequences that show no sequence similarities to any other known sequence (e.g. PDB identifier 1xpv); (iii) The PDB contains some very short sequences that do not represent domains and hence are not in Pfam (e.g. the inhibitor in PDB:2cn7). Regardless of the underlying sequence database, similar technical and philosophical difficulties would prevent us from reaching 100% sequence coverage. However, we estimate that 95% probably represents the current upper boundary of sequence coverage (explored further in the ‘Recalculating sequence coverage’ section). This study also suggests that, at most, we can expect three-quarters of residues to be matched by Pfam. One of the problems with estimating coverage across either the UniProt sequence database or the PDBs is that it is very biased to certain classes of sequences. For example, there are over 60,000 nearly identical sequences of the HIV GP120 glycoprotein in UniProt and many hundred structures of lysozyme in the PDB. The PDB also tends to contain smaller globular proteins with few transmembrane domains. These biases all have the effect of inflating our estimated coverage.

An alternative way to measure coverage is to calculate the sequence coverage on a variety of complete genomes that are stable. This measure of coverage has been used by other groups trying to estimate the number of protein domains [11, 12]. We find that about 65% of most genomes are matched by about 1000–2000 Pfam families as predicted by Chothia and subsequently confirmed by Orengo and Thornton [12]. Table 1 gives the sequence and residue coverage figures for a group of model organisms. Intuitively, the larger the proteome the more Pfam families are

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence Coverage (%)</th>
<th>Number of families in Pfam database</th>
<th>Number of Sequences in Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human - Homo sapiens</td>
<td>67</td>
<td>3506</td>
<td>38,453</td>
</tr>
<tr>
<td>Worm - Caenorhabditis</td>
<td>63</td>
<td>2631</td>
<td>22,509</td>
</tr>
<tr>
<td>Yeast - Saccharomyces</td>
<td>73</td>
<td>2166</td>
<td>5799</td>
</tr>
<tr>
<td>Yeast - Schizosaccharomyces</td>
<td>78</td>
<td>2140</td>
<td>5013</td>
</tr>
<tr>
<td>Escherichia coli K12</td>
<td>86</td>
<td>2070</td>
<td>4329</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td>75</td>
<td>1674</td>
<td>4105</td>
</tr>
<tr>
<td>Methanococcus jannaschii</td>
<td>73</td>
<td>942</td>
<td>1782</td>
</tr>
<tr>
<td>Pyrococcus abyssi</td>
<td>81</td>
<td>956</td>
<td>1786</td>
</tr>
<tr>
<td>Buchnera aphidicola subsp.</td>
<td>97</td>
<td>636</td>
<td>562</td>
</tr>
<tr>
<td>Schizaphis graminum</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Sequence coverage per genome of several model organisms

Figure 1: (A) Cumulative increase of sequence coverage with increasing number of families for Pfam release 22.0 against UniProt, composed of Swiss-Prot 5.1.7 and SP-TrEMBL 34.7. The x-axis orders the addition of families to Pfam over time. (B) Change in sequence coverage and number of families with respect to time.
required to provide the same level of coverage. However, as the size of a proteome increases there appears to be an increase in the amount of domain reuse. Proportionally fewer families are required to give the same level of coverage (compare *Saccharomyces cerevisiae* and *Methanococcus jannaschii* in Table 1). Overall, the average proteome coverage for eubacteria, archaeabacteria and eukaryotes is 74, 68 and 66%, respectively. It is noteworthy that sequence coverage does not correlate to residue coverage. In archaeabacteria and eukaryotes the level of sequence coverage is comparable, but the average residue coverage is only 35% in eukaryotes compared with 51% in archaeabacteria. The Pfam residue coverage of eubacteria is the highest at 56%, but this is far from complete. Both measures of coverage are significantly lower than that achieved for the PDB, indicating that our knowledge of proteins structures is far from complete.

**KEEPING UP WITH THE DELUGE OF NEW SEQUENCE**

Despite sequence data is growing exponentially, Pfam coverage appears to be tracking this growth. Looking at the sequence coverage over the last 8 years (Figure 1B) we see that it has remained between 73 and 75% since 2002 despite the addition of about 5000 new families. The estimated number of families required to obtain around 100% coverage (based on curve fitting on Figure 1A) estimates 184 472 families. A similar, more simplistic, curve extrapolation was carried out by us 4 years ago and we estimated that only 24 000 families would be required to achieve the same level of sequence coverage. Thus, as more sequences are added to the database more novel families are discovered. By taking the current version of Pfam (22.0) and searching against UniProt database at different points in time, it is possible to assess the change in cumulative coverage over time (Figure 2). From this analysis, it is evident that the apparent number of Pfam families needed for complete coverage is growing over time. It is interesting to note that the current Pfam database matches to over 90% of sequences that were available in 1988. But perhaps the most surprising result is that if Pfam or other databases stopped building new entries, their coverage would drop over time. We have simulated this, and found out that if Pfam had stopped building families in 2003, our current sequence coverage would only be at 68.4% (about 5% lower). Thus, continued curation is essential to ensure that the database remains current and hence useful.

The idea that the number of sequence families is growing with the increase in sequence data is not new. It was noted that the rate of increase in new families was approximately linear with the number of newly sequenced genomes [11]. In this paper Kunin and colleagues plotted the total number of families identified using the TRIBE-MCL algorithm against the number of genomes included. Their analysis included the first 83 complete genomes known (total of 311 256 proteins) and identified 56 667 families. This number is consistent with subsequent genome coverage analyses that have indicated that ~20 000–60 000 families are required, but the numbers vary according to the protein families data applied [13–17]. Kunin et al. [11] suggested that the protein universe was still greatly undersampled. The graph shown in Figure 3 corroborates their results, showing that new sequences are resulting in an increase in the number of families required to achieve the same level of coverage. A sequence coverage cut-off value of 95% was chosen as this
**Figure 3**: Growth in the number of families estimated for complete sequence coverage. Although the number of families needed doesn’t appear to grow rapidly the exponential growth in sequence numbers will mean that number of families continues to grow.

is the expected maximum sequence coverage inferred from the current level of PDB coverage. Figure 3 was obtained by extrapolating the curves shown in Figure 2, using a modified exponential curve [Equation (1)].

\[ y = V_{\text{max}} + A_1 e^{-x/t_1} + A_2 e^{-x/t_2} \]  

(1)

Where \( V_{\text{max}} \) specifies a boundary condition (i.e. curve flattens off at 100% coverage), \( A_1, A_2, t_1 \) and \( t_2 \) are curve specific constants, \( x \) the number of pfam families and \( y \) is sequence coverage. \( R^2 \) values for all curves plotted using this fitting model were \( >0.992 \). Solving Equation (1) for the number of families required to give 95% coverage based on different sized sequence databases (which determine the curve specific constants) demonstrates that there is an increase in the number of families as the sequence database increase, we are beginning to see some saturation, as shown by the latter part of the graph in Figure 3, where the rate of increase of the number of families required for a corresponding increase in sequences is decreasing. The graph in Figure 3 shows that 38,062 families are required for 95% coverage with the current size of UniProt (4.1 million sequences). If 100% coverage were attainable, the number of families would be far greater, exceeding 180,000 families. This further reinforces the concept of the law of diminishing returns mentioned previously.

In addition to the curated Pfam-A families, Pfam provides an automatically generated supplement, termed Pfam-B. At release 22.0, the 182,493 supplemental Pfam-B families provided an additional 4% sequence coverage and 7% residue coverage. As these lower quality Pfam-B families provide a source of new Pfam-A families, we recently performed a survey to understand how many families appeared to be novel. To do so, a profile-HMM was built for all Pfam-B families containing 10 or more sequences (17,127 Pfam-B families). Each of these profile-HMMs was compared with the library of Pfam-A profile-HMMs using the profile-profile comparison tool PRC. A match with an \( E \)-value \( <10^{-3} \) was used as a conservative threshold for relatedness between a Pfam-B and Pfam-A family. Interestingly, using this approach it was found that 50% of Pfam-Bs appear to be related to existing Pfam-A families. This indicates that some Pfam-A HMMs are insufficiently sensitive to detect all known members and/or that some families are too divergent to model with a single profile-HMM. Consequently, we expect that many of the sequences that remain to be classified will fall into existing families.

**RECALCULATING SEQUENCE COVERAGE**

So should the ‘complete coverage’ component of the Pfam scientific goal equate to 100% sequence and residue coverage? There are good reasons to argue that this should not be the case and that these are unrealistic goals. First, we consider sequence coverage. Some sequences contained within the UniProt database are only partial sequences, termed fragments. As some of these fragments are only short or only contain a poorly conserved region they fail to score significantly above the background noise to any Pfam model and will therefore not be covered by Pfam. Furthermore, in Pfam we only build families where the seed alignment contains two or more sequences and consequently, all Pfam entries have two or more sequence members. Although it is possible to build a profile HMM from a single sequence, there is little to be gained from it. Philosophically, a classification of one sequence does not improve our understanding of protein space.

The value of \( \sim 180,000 \) families for 100% sequence coverage is comparable to the figure obtained using the Automatic Domain Decomposition Algorithm (ADDA) [18]. In this article, an automated approach was used to cluster large sets of non-redundant protein sequences, split them into domains and then further classify them into families. This approach
resulted in the identification of 123,000 families required for 100% coverage, a figure of the same order of magnitude as that obtained by us.

Given the law of diminishing returns it seems likely there will be a large number of singleton families. We calculate that of the 184,472 families required for 100% coverage, 115,784 of them would currently be singletons. But where do all of these singleton families come from? It seems entirely likely that many of these could be incorrect translations that are never expressed as a protein in the cell. There is also evidence that alternative reading frames have been incorrectly annotated in some cases in bacterial genomes [19]. Analyses of these proteins have demonstrated that they are typically short (<200 amino acids in length) and are predicted to contain little or no secondary structure [20]. Even in well-known families of proteins, some sequences appear to have incorrect initiating methionines that can give rise to spurious N-terminal extensions or deletions in the subject proteins. We also see retained introns that seem unlikely to be expressed as part of a functional protein product. Only a minute fraction of proteins have ever been investigated experimentally and most are the product of gene prediction software. Thus, how many of these proteins are real proteins? Recently, UniProt added protein existence tags to each entry, which provides an indication to the amount of evidence for the existence of a protein. We have taken the simple 1–5 scale and divided it into two gross classes: evidence (protein existence values 1–3) and no-evidence (protein existence values 4–5), and assessed the sequence coverage of each class (Figure 4). Interestingly, we have 91% coverage of sequences where there is some evidence for its existence, but only 55% coverage where there is no-evidence. There appears to be no correlation between date of entry and the protein existence value, so we are not simply observing the same date/coverage effect as in Figure 2. The early, large important families are still being found on sequences that are being predicted, but the contribution to coverage is proportionally lower compared with those proteins that have evidence. More recent Pfam families provide proportionally higher coverage on sequences that are predicted, but these latter families or domains can still be found on characterized sequences, indicating that they may well be real sequences. However, the large gulf in coverage could be due to one or both of the following factors: there is a fraction (which is difficult to quantify) of proteins that are spurious translations and hence do not appear like any other proteins and/or the predicted proteins contain previously unobserved domains that are typically scarcely distributed.

But not all singletons are spurious. There has been some work on proteins that appear to share no similarity to any other known proteins, termed ORFans [21]. A few of these proteins, such as ORFans with PDB identifiers 1of5B (mRNA export factor Mex67-Mtr2), 1oqlA (hypothetical protein Apc1120) and 1m5uA (hypothetical protein Hi1480), have been expressed and purified and recently their 3D structure has been determined. It has been shown, that although many ORFans show no significant sequence similarity to other proteins, the 3D structures of the majority of the ORFans have previously observed folds, and fall into existing superfamilies [22]. There have been a number of attempts to explain why ORFans have no homologues: it has been suggested that they could be the product of lateral gene transfer from as yet unknown organisms [23]. Alternatively, they could potentially be due to gene losses in all actual sampled organisms or de novo generation, and are thus unique to the organisms that produce them [24].

Some proteins evolve extremely rapidly and it can be very hard to discern any similarity to known proteins, with only a few key functional residues and the basic 3D structures retained compared with the ancestral sequence [24]. For example, some bacterial restriction endonucleases represent an extreme case of rapidly evolving proteins. These proteins evolve rapidly due to the fact that they act as a defence
mechanism against bacteriophage infection. Thus, another explanation of ORFans is that these proteins are rapidly evolving and hence have diverged beyond recognition with its homologues using traditional sequence analysis. To enable the classification of these rapidly evolving proteins, it may be necessary to employ more sophisticated techniques for homologue detection, such as fold assignment methods and profile–profile HMM comparisons. Other suggestions for the presence of ORFans include that they play a role in signalling or regulatory events that are highly specific to an individual organism [12].

**RECALCULATING RESIDUE COVERAGE**

Intuitively, if 100% sequence coverage is unachievable then 100% residue coverage must also be unachievable. But, even if singletons were modelled and included in Pfam, 100% residue coverage would still be an unrealistic goal. This is because not every residue in a polypeptide chain will form part of a folded domain. Indeed, domains are connected together by linker regions (Figure 5), segments of residues that play an essential role in maintaining cooperative interdomain interactions [25], and also affect protein stability, folding and domain–domain orientation [26]. Such regions, together with non-conserved terminal regions (NCTRs), low complexity regions (LCRs) and signal peptides should also, in theory, be outside Pfam-A domains.

Calculations on the Pfam 22.0 database revealed that from a total of 1.3 billion residues, 0.91% fall within linker regions. A linker was defined as a region between two Pfam-A domains with a minimum length of one amino acid, and a maximum length of 30 residues (since very few globular domains are shorter than this), ensuring that the region was small enough such that there would be no unidentified domains within it. It is worth noting, however, that the rigid linker definition used resulted in a number of potentially longer linkers being missed. In fact, altering the linker definition to that suggested by George et al. [27] (maximum linker length being 80 residues instead of 30), increases the number of residues that fall within linker regions to 2.62%.

NCTRs at the N/C termini of peptides also contribute to a decrease in maximum residue coverage that can be obtained. Further calculations on the database showed that if maximum NCTR size is defined as 30 residues, then an additional 2.72% amino acids fall within these regions. Increasing the maximum size of NCTRS to 80 residues results in a further increase: up to 7.74% of residues fall within such regions. It is worth noting that 0.61% of all residues also fall within signal peptide sequences—such regions are accounted for in the N-terminal non-conserved regions.

**PROSPECTS FOR COMPLETING THE Pfam CLASSIFICATION**

In the previous sections, we have outlined the current state of coverage and what the ‘complete and accurate classification of proteins families and domains’ may actually equate to, in terms of sequence and residue coverage. Based on our current PDB coverage, we estimate that the maximum sequence coverage that can be attained is ~95%. Interestingly, the graph in Figure 4 also shows that we currently have around 91% coverage of all proteins that have a PE level of 1–3. Assuming that maximum sequence coverage is around 95% and that we are going to need over 38,000 families to reach this target, what are the prospects of reaching this goal in the near future given that it has taken Pfam 10 years to achieve 73.23% and 10,000 families?

We know from both structural classifications, such as SCOP and CATH, and our own analysis of Pfam-B families that some of our existing families are insufficiently sensitive to detect all known members of a large divergent family. Thus, one strategy for improving sequence coverage is to try and improve the sensitivity of existing models. A more sensitive model can be achieved by improving the seed alignment either by optimizing the alignment and/or including more divergent sequences. Sequence analysis tools have improved greatly over the
last 10 years. Multiple sequence alignment software, such as MAFFT [28] and Muscle [29] are fast and accurate to the point where manual curation of alignments is rarely employed for Pfam seed alignments. Such alignment software makes it possible to align more distantly related sequences than before and this has allowed us to build more inclusive families and in some cases merge older families together. Although the deluge of sequence data does cause problems, it does bring benefits. As sequence space fills up it can make it much easier to identify similarity between two clusters of similar sequences, as there are more intermediate sequences to enable the link to be made. Identifying more divergent, yet homologous sequences, makes the seed alignments more comprehensive, thereby allowing the profile HMM to model more of the events that have occurred during the evolution of a protein family. Furthermore, an improvement of the HMMer software, which is used to generate profile HMMs, would also improve sequence coverage. The second generation of the HMMer (version 2) was a significant improvement over the first generation and a new improved third generation of HMMer is currently in development (Eddy S, personal communication). There are a wide variety of other techniques that could be used to identify distant relationships between proteins and expand Pfam families. For example secondary structure prediction has been used for this purpose [30, 31] as well as fold recognition methods such as threading [32].

As established above, as the number of sequences in the underlying database increases, there is also a requirement for an increase in the family numbers so that the database remains current. Although the speed of HMM searches is important, this currently does not pose the largest bottleneck in producing new families. The rate-limiting step in building a Pfam family is, without doubt, the manual steps involved in curation. For each family, the domain boundaries need to be manually defined, a threshold chosen and each entry subsequently annotated. The solution to this problem is not trivial. Currently, Pfam is comprised of a team of four people, with only one dedicated annotator. However, manual intervention is essential. If we blindly added automatically generated families, such as those found in Pfam-B, this would in theory give us enough families to give us 95% sequence coverage. However, automatic generation of protein families is dependent on the domain definition used within the algorithms, and currently cannot deliver high quality families consistently. Different domain definitions will result in different classifications, which may not accurately model the sequence space. Human input is required to ensure accuracy and subsequent high quality annotation. Various automatically generated protein domain databases may be found online. These include ProDom [6], which at release 2004.1 had a sequence coverage of 59.74%, EVEREST, which had 70 029 families at release 2.0, with 64% sequence coverage and 59% residue coverage, as well as 88% PDB sequence coverage and 84% PDB residue coverage [33]. This shows that Pfam has higher sequence coverage than both automatically generated classification resources. Another resource, ADDA [18], referred to previously in the ‘Recalculating coverage’ section, has obtained a 100% coverage of more than 1.5 million sequences with 123 000 families.

An alternative approach to achieving complete coverage is to change some of the underlying principles of the Pfam database. Currently, thresholds are chosen such that there are no known false positives. Occasionally, false positives are identified; these families are immediately remediated ensuring that error is removed at the next Pfam release. Having such a rule means that our specificity is extremely high, but it has a detrimental effect on sensitivity. Thus, one way of improving sensitivity would be to reduce our specificity, maybe allowing a false positive rate of 1%, such as that employed by the SUPERFAMILY database [34]. Another different, but related, rule is that no overlaps are allowed between Pfam families. This means that no residue in the underlying sequence database can appear in more than one family. Overlaps can occur as a result of clashes between the C-terminus and N-terminus of consecutive domains or simply that the two families are related and hence the two different models are matching the same region of sequence. In order to comply with this rule, thresholds on some families can be artificially increased to remove overlaps. Alternatively, the domain boundaries can be redefined to prevent the overlap. Relaxing either of these rules would allow us to improve coverage. However, it would mean that users would have less confidence in Pfam entries. Ambiguity in annotations would mean that Pfam would become less useful for genomic annotation. Furthermore, the classification would become less precise and hence less informative. Therefore, we do not intend to increase coverage at the cost of quality.
Another approach would be to increase the amount of community input in the curation process. Many collaborators from various laboratories are submitting new families for inclusion in the database, defining the function of entries labelled as having an unknown function, and providing further annotation to supplement existing families. However, the process of community input is currently very *ad hoc*. A more systematic method of annotation could involve the use of common tools such as version control (CVS) and/or wikis. However, there are many problems associated with community annotation. The first is making the interface simple enough for community uptake, yet the interface has to be sophisticated enough to ensure that there is data provenance and that the quality of the database does not deteriorate. Nevertheless, community annotation is most likely the most effective mechanism to ease the curation bottleneck.

From the analysis presented within this article it is clear that 100% sequence coverage is unachievable with Pfam in its current form. We estimate that an additional 28,000 families are required for 95% sequence coverage of the current sequence database, and 174,472 families for 100% coverage. This may seem a distant goal, given that it has taken us 10 years to add 10,000 families. However, through the use of state-of-the-art-software, increased community input, we believe that this is an achievable goal for the database.

**Key Points**

- Pfam is a comprehensive database of protein families, containing 9318 families in the latest release (22.0) with 73.23% sequence coverage.
- Inferring from PDB coverage, maximum sequence coverage is probably around 95%.
- Curve fitting models based on the current UniProtKB release estimate around 38,000 families required to achieve 95% sequence coverage.
- Improvements in algorithms and software used to build families, as well as community input have, and will continue to result in an increase in the rate of family creation.

**Acknowledgements**

S.J.S., R.D.F. and A.B. are funded by The Wellcome Trust.

**References**


