BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 21 2012, pages 2763–2772
doi:10.1093/bioinformatics/bts533

Structural bioinformatics

Advance Access publication August 31, 2012

Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB

Qifang Xu and Roland L. Dunbrack, Jr*

Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Automating the assignment of existing domain and protein family classifications to new sets of sequences is an important task. Current methods often miss assignments because remote relationships fail to achieve statistical significance. Some assignments are not as long as the actual domain definitions because local alignment methods often cut alignments short. Long insertions in query sequences often erroneously result in two copies of the domain assigned to the query. Divergent repeat sequences in proteins are often missed.

Results: We have developed a multilevel procedure to produce nearly complete assignments of protein families of an existing classification system to a large set of sequences. We apply this to the task of assigning Pfam domains to sequences and structures in the Protein Data Bank (PDB). We found that HHsearch alignments frequently scored more remotely related Pfams in Pfam clans higher than closely related Pfams, thus, leading to erroneous assignment at the Pfam family level. A greedy algorithm allowing for partial overlaps was, thus, applied first to sequence/HMM alignments, then HMM–HMM alignments and then structure alignments, taking care to join partial alignments split by large insertions into single-domain assignments. Additional assignment of repeat Pfams with weaker E-values was allowed after stronger assignments of the repeat HMM. Our database of assignments, presented in a database called PDBfam, contains Pfams for 99.4% of chains >50 residues.

Availability: The Pfam assignment data in PDBfam are available at http://dunbrack2.fccc.edu/ProtCid/PDBfam, which can be searched by PDB codes and Pfam identifiers. They will be updated regularly.

Contact: Roland.Dunbracks@fccc.edu

Received on February 10, 2012; revised on August 19, 2012; accepted on August 24, 2012

1 INTRODUCTION

Clustering proteins of known structures into families or superfamilies is a long-standing task of particular importance in understanding structure–function relationships and for protein structure prediction by homology. Usually, protein classification in the PDB is accomplished at the level of domains—substructures that recur as functional units in different protein contexts.

Structure-based domain classifications of the PDB, such as SCOP (Murzin *et al.*, 1995) and CATH (Orengo *et al.*, 1997), are constructed by comparing the available protein structures in the PDB and creating classifications of new folds and superfamilies manually. Existing structure-based classifications cover only a portion of the PDB. The most recent SCOP release (v. 1.75A) is 2 years behind the PDB and only covers 61% of current PDB entries. CATH was last updated in November 2011 and covers 64% of the current PDB.

Sequence-based approaches, such as Pfam (Sonnhammer *et al.*, 1997), ProDom (Servant *et al.*, 2002), InterPro (Hunter *et al.*, 2009), SMART (Schultz *et al.*, 1998) and SUPfam (Pandit *et al.*, 2002), can be readily applied to new structures because most new structures fit into existing sequence clusters. In this way, they are well suited for rapid and automated classification of new structures in a way that structure-based classifications are not. Some differences between sequence-based and structure-based classifications occur when a single sequence domain is structurally two or more domains with separate hydrophobic cores or a single structural domain is two sequence domains (Zhang *et al.*, 2005). Structure-based methods are often superior in recognizing remote relationships between families, as these relationships may be apparent only from structural similarity and in the absence of any recognizable sequence similarity. Even with structural information, it may be difficult to distinguish between divergent and convergent evolutionary relationships (Tress *et al.*, 2005).

Our aims in this article are two fold: first, to develop a general procedure that can be used to make rigorous assignments of existing protein family classification systems for any set of protein sequences, and second, to perform this task for the entire PDB. For the PDB, we wish to define a method that can be run automatically on a weekly or monthly basis. We have thus chosen the sequence-based domain classification given by Pfam, as new proteins in the PDB can be readily assigned to existing Pfams, without manual intervention required by structure-based classification systems.

Pfam is a database of protein families, in which each family is represented by a hidden Markov model (HMM) created from manually curated multiple sequence alignments (Sonnhammer *et al.*, 1997). The Pfam classification of protein families has gained widespread acceptance among biologists because of its wide coverage of proteins and a sensible naming convention related to protein functions and commonly used names (Pkinase, SH2, etc.). Pfam was recently used in the Protein Structure Initiative to select targets and divide them among different high-throughput centres (Dessailly *et al.*, 2009). Some Pfam families are seeded by structures in the PDB (Finn *et al.*, 2010). Two or more related Pfam families are grouped into a Pfam clan (Finn *et al.*, 2006). Such relationships are often

*To whom correspondence should be addressed.

identified through structural similarity, as they are in the structural classification systems.

Several assignments of Pfam domains to the PDB are currently available, including Pfam itself (Punta *et al*., 2012), SIFTS (Velankar *et al*., 2005) and the RCSB (Research Collaboratory for Structural Bioinformatics), covering 45, 87 and 94% of unique sequences in the PDB, respectively. Each of the currently available sources of Pfam assignments to the PDB suffer from one or more of a number of problems. First, because they use only the original PDB or UniProt sequences against the Pfam HMMs, they miss many potential assignments that occur when the sequence is not closely related to any single Pfam family. It is likely that sequence methods based on profile–profile comparison may identify these relationships and provide higher levels of statistical significance. Second, in some cases, these sources also provide completely overlapping assignments, sometimes when two different Pfams in the same clan align to the same region of a PDB sequence with good E-values. If we want to cluster at the family level and then the superfamily level, this produces discrepancies. Third, some proteins have long insertions relative to the Pfam HMM definition, and HMMER may produce two alignment segments, one on either side of the insertion. These two segments cover non-overlapping regions of the HMM, and together should comprise a single Pfam assignment. However, the publicly available sources simply list these separately, and they cannot easily be distinguished from repeated domains in the same protein. Fourth, some protein structures are composed of two chains that together comprise a single Pfam domain, i.e. two non-overlapping regions of the same Pfam HMM. Pfam, SIFTS and the RCSB do not properly account for domains split by insertions or split between different chains.

We overcome some of the deficiencies of other Pfam databases using several strategies. The first is to use consensus sequences derived from PSI-BLAST profiles and to run these through the Pfam HMM library. Such sequences can be fed to the Pfam HMMs like any protein sequence, and they usually produce more complete alignments with better E-values than the original sequences. Similar techniques have been used by us previously (Kahsay *et al*., 2005) and by others (Przybylski and Rost, 2008). Secondly, we utilize HHblits (Remmert *et al*., 2012) to produce HMMs for PDB sequences and their parent UniProt sequences, and then apply HMM–HMM alignment of these HMMs against the Pfam HMMs using HHsearch (Söding, 2005). The third approach is to utilize structure alignment of statistically confident and complete structures in Pfam families with weak hits in the same Pfam families—either those with poor E-values and/or alignments that cover only a portion of the Pfam HMM. This allows us to verify whether a weak assignment is correct and to extend short alignments.

Finally, we have developed a procedure for optimally combining assignments from these multiple sources into Pfam architectures for each protein in the PDB. The procedure combines non-overlapping or minimally overlapping partial assignments to the same Pfam into single assignments, thus, accounting for large insertions or domains split across multiple protein chains. We assign additional repeat domains at weaker E-values if the same repeat family is assigned earlier in the procedure at an E-value better than the general cut-off.

We explore the properties of the regions and proteins that cannot be assigned to a Pfam domain and the interactions between Pfam domains in the biological assemblies of structures in the PDB, according to our Pfam assignments. Regions not assigned to Pfams have a greater tendency to be disordered in protein structures and to have lower rates of regular secondary structure than regions assigned to Pfam domains. The number of Pfam–Pfam interactions is increased by the number of assignments made using the methods described here but are also critically dependent on the usage of biological assemblies from crystal structures rather than the asymmetric units.

The Pfam assignments can be searched through the ProtCID server (http://dunbrack2.fccc.edu/protcid) by PDB codes, Pfam codes and sequences. Downloadable files of the Pfam assignments and those proteins that cannot be assigned are also available on the website http://dunbrack2.fccc.edu/ProtCID/PDBfam.

Our procedure is general and can be applied to other domain classification systems and other target sequence sets. Even if the target sequence set is not the PDB, structural information may still be used for proteins in the sequence set that can be readily aligned with proteins of known structure.

2 METHODS

2.1 Searching Pfam through PSI-BLAST consensus sequences and HHsearch

Pfam v26 files Pfam-A.hmm and Pfam-B.hmm were downloaded from the Pfam website and were used as HMMER3 profile databases (Finn *et al*., 2010). The PDB sequences (Berman *et al*., 2000) were parsed from pdbx_seq_one_letter_code records in the PDB XML files (Westbrook *et al*., 2005). UniProt sequences were downloaded from the UniProt website (Bairoch *et al*., 2005). The XML files from the SIFTS database (Velankar *et al*., 2005) were used to find the residue correspondence between the UniProt and PDB sequences.

For each unique PDB sequence, we used one iteration of our modified PSI-BLAST (Altschul *et al*., 1997) from MolIDE (Wang *et al*., 2008) to generate a profile from sequences in the UniRef90 database (Li *et al*., 2000). The parameters for PSI-BLAST were '-e 10 -h 0.0001 -v 5000 -b 5000 -N 25 -f 16'. A PSI-BLAST profile is a position-specific scoring matrix (PSSM), which provides a log-odds score and percentage of occurrences for each of the 20 amino acid types at each position in the query sequence. A consensus sequence is a 1D simplification of a PSI-BLAST profile obtained by substituting the 20-dimensional vector in each residue position by the highest scoring or most common amino acid observed at that position. In this article, a 'percentage consensus sequence' is composed of the most frequent residues in each column, whereas a 'PSSM consensus sequence' is composed of the highest scoring amino acid at each position. We also applied the same procedure to the full UniProt sequences from which PDB sequences are derived, as identified by SIFTS. We, thus, have the following six sets of sequences: PDB sequences, PDB percentage consensus sequences, PDB PSSM consensus sequences, UniProt sequences, UniProt percentage consensus sequences and UniProt PSSM consensus sequences. In this article, we denote those sequences as PDB, PDB-percent, PDB-pssm, UNP, UNP-percent and UNP-pssm, respectively. We ran HMMER3 on all six sets of sequences against Pfam A and Pfam B HMM models. We refer to these six sets of alignments as 'HMMER hits'.

We ran HHblits on unique sequences in the PDB and UniProt sequences to generate HMMs on database uniprot20_29Mar11, which is a database of HMMs created from a clustering of UniProt sequences

at 20% identity (Remmert *et al.*, 2012). We searched the Pfam HMMs with the HHblits-derived PDB and UniProt HMMs with HHsearch (Söding, 2005) to generate Pfam to PDB alignments through HMM–HMM alignments. We refer to these as 'HH hits'.

2.2 Pfam E-value and FATCAT *P*-value cut-offs

To determine the cut-off of HMMER E-values and structure alignment *P*-values for each Pfam A present in the six sets of alignments, we collected those Pfam hits with HMMER E-value of $<10^{-5}$, HMM coverage >0.9, and then selected the alignment with the largest number of match states assigned to residues with Cartesian coordinates in the PDB structures as the representative hit. A total of 5134 Pfams were selected. With HMMER3, we aligned each PDB sequence of these representative hits to all of the 5134 Pfam HMMs. The resulting data points were divided into the following two classes: same clan and different clan, depending on whether the two Pfams were in the same or different clans according to the clan definitions in the Pfam v26 MySQL database.

Smoothed density function curves were calculated using kernel density estimates (Sheather and Jones, 1991) in the R project (http://www.r-project.org/) by calculating probability density estimates of same clan and different-clan prediction as a function of $\log_{10}$(E-value):

$$f(A) = \frac{1}{N}\sum_{i=1}^{N} K_h(A - A_i)$$

where A_i is $\log_{10}$(E-value), and K_h is a Gaussian kernel with bandwidth h:

$$K_h(x) = \frac{1}{\sqrt{2\pi}h}\exp\left(\frac{-x^2}{2h^2}\right)$$

The probability at A is calculated using Bayes' rule:

$$P(same|A) = \frac{P(A|same)P(same)}{P(A|same)P(same) + P(A|diff)P(diff)}$$

where $P(A|same)$ and $P(A|diff)$ are calculated from $f(A)$ using the same clan and different-clan sets of E-values, respectively. $P(same)$ and $P(diff)$ are the percentages of data points from the same-clan class and the different-clan class. From a value of A such that $P(same|A) > 95\%$, we selected a threshold for the Pfam E-values of 10^{-5} (see 'Results' section).

To select a Pfam E-value threshold for HH hits, we applied the same procedure on the HH alignments, which contains 5387 Pfam hits. The threshold of HHsearch E-value for $P(same|A) > 95\%$ is 10^{-4} (see 'Results' section).

We performed structure alignment with the FATCAT program (Ye and Godzik, 2003) of each structure with every other structure in the 5134 Pfam set. The data points consisting of $\log_{10}$(*P*-values) were defined as either same clan or different clan. Kernel density estimates and Bayes' rule were used to obtain $P(same|A)$ where A is the $\log_{10}$(*P*-value) from FATCAT. From the value of A such that $P(same|A) > 95\%$, we selected a threshold for the FATCAT *P*-values of 10^{-3} (see 'Results' section).

2.3 General greedy algorithm

From any set of alignments of PDB sequences to Pfam HMMs, we use the same general procedure based on a simple greedy algorithm to create a unique assignment of a Pfam to each residue in a PDB sequence. Such an assignment constitutes a Pfam 'architecture' or arrangement of domains in the PDB sequence, allowing only for short overlaps.

For a given PDB sequence, we start by assigning the hit with the best E-value. If there is any region in the query of >30 amino acids that occurs within the boundaries of the alignment to the best HMM but which is not aligned to HMM match states, we create a 'split assignment'. A split assignment indicates that match states in the HMM align to separate non-contiguous regions of the query sequence. The residues in the inserted region of the query are then 'unassigned', which means they are available for subsequent assignments. For each additional hit in order by E-value (best to worst), we check whether it overlaps the current Pfam assignments by >10 residues on either end. If it does not, then an assignment is made. Again, long insertions in the query result in split assignments and the insertions are unassigned.

If at any time, the same Pfam model aligns more than once to a query sequence, we check if the HMM match states align only once to the query and in order allowing short overlaps of <10 amino acids in the HMM. If yes, then we combine them into one assignment to the HMM. The assignment is split if there are >30 residues between the assigned regions, and the intervening residues are left unassigned. If the assignments to the Pfam cover the HMM match states more than once, then there is more than one copy of the Pfam in the sequence (e.g. repeated domains), and multiple assignments of the Pfam are made.

We also check whether the same Pfam aligned to different sequences within the same PDB entry. In some cases, these hits do not overlap in the HMM by >10 amino acids, and they are then combined into a single assignment.

In our procedure, we always used HMMER hits first, then HH hits.

2.4 Using structure alignments to improve Pfam assignments

We use structure alignment to verify whether Pfam–PDB alignments with weak E-values are correct and to extend short alignments to Pfam HMMs. To do so, we need to identify structures (or domains within structures) that cover Pfams in their entirety with good E-values. We call such structures *exemplars* for their Pfams. Only a subset of Pfams in the PDB has such high-quality alignments.

To identify exemplars, we first applied the greedy algorithm on all Pfam alignments in the six sets of sequences and consensus sequences with a conservative HMMER E-value of $\leq 10^{-5}$, obtaining split and combined Pfam assignments. Some split assignments may be possible where one component has significant E-value, whereas the other is weaker. Therefore, we continue the greedy algorithm with alignments with E-value of $>10^{-5}$ if the same Pfam has already been assigned to the PDB sequence, up to an E-value of 1.0. We continued the greedy algorithm with the HH hits with an E-value cut-off of 10^{-4}. The reason for applying the HMMER alignments before the HH alignments is discussed in the 'Results' section. For Pfams assigned in this procedure, we identify an *exemplar structure*, defined as the structure with the largest number of match states assigned to residues with Cartesian coordinates in the PDB entry, with a coverage of the Pfam HMM of at least 80%. HMM coverage is the number of the sequence residues with coordinates aligned to a Pfam HMM match state divided by the length of the model. In the event of a tie, the structure with the best E-value is used.

We divided the HMMER Pfam hits of all six sets into two non-overlapping sets: {Strong Hits} and {Weak Hits}. Strong hits are those hits with E-value of $\leq 10^{-5}$ and <10 residues missing from the N or C terminal end of the HMM, whereas weak hits comprise the remaining alignments. For each hit in {Weak Hits}, we checked whether there are exemplar structures for that Pfam and/or other Pfams in the same clan. If there are, we perform structure alignments with the FATCAT program (Ye and Godzik, 2003) on the region(s) of the weak hit structure not previously aligned to the {Strong Hits}. We performed this procedure separately for HH Pfam hits with E-value of $\leq 10^{-4}$.

If the FATCAT *P*-value is better than 10^{-3}, we create an alignment of the PDB query to the Pfam HMM through the exemplar structure through a transitive alignment. For residue pairs AB and BC, (A to B) + (B to C) = (A to C). Here, A to B is the HMM to exemplar alignment, B to C is the structure alignment of the exemplar to the weak assignment and A to C is HMM to the weak assignment. Once this alignment is created, we move the alignment from {Weak Hits} to a new set {Struct Hits}.

2.5 The full algorithm for assigning Pfams to PDB sequences

The full procedure of creating Pfam assignments to PDB sequences is as follows. We have in hand six sets of alignments, {HMMER Strong Hits}, {HH Strong Hits}, {HMMER Struct Hits}, {HH Struct Hits}, {HMMER Weak Hits} and {HH Weak Hits}, the last two containing those weak hits (too short and/or too weak an E-value) for which structure alignment was not possible or did not produce a significant alignment. We use the {HMMER Strong Hits} first in the greedy algorithm until no more assignments can be made, and then continue with the {HH Strong Hits}. Second, we continue the greedy algorithm with the alignments in the {HMMER Struct Hits} and {HH Struct Hits} sets in that order until no more assignments can be made. Third, we apply the greedy algorithm to the remaining {HMMER Weak Hits} and {HH Weak Hits} with E-value of $\leq 10^{-5}$ (HMMER) or $\leq 10^{-4}$ (HH). These hits have strong statistical significance but >10 residues missing from the N or C terminal end of the HMM. Fourth, we proceed with the remaining {HMMER Weak Hits} and {HH Weak Hits} up to a value of 1.0, but we only add these if the same Pfam has already been assigned in one of the earlier steps. Some of these will be combined with earlier assignments to produce split assignments. Some will be repeated domains. Pfam B assignments are treated as weak hits and added only if the E-value is better than the appropriate threshold.

3 RESULTS

3.1 Establishing E-value and *P*-value cut-offs

We investigated the HMMER3 E-value level at which Pfam assignments are likely to be reliable. We created a set of 5134 Pfams with E-values to unique PDB sequences $\leq 10^{-5}$ and HMM coverage ≥90%. The associated PDB for each sequence was the one with the largest number of match states assigned to residues with Cartesian coordinates in the PDB entry. We aligned the PDB sequences against the other 5133 Pfams in the set with HMMER3 and classified the resulting alignments and E-values depending on whether the PDB sequence and the Pfam belonged to the same clan or different clans, according to Pfam v. 26. The probability density functions and classification functions versus $\log_{10}$(E-value) are shown in Figure 1a. The classification function refers to how likely the Pfam of the query sequence and the Pfam of the hit HMM belong to the same or different clan as a function of $\log_{10}$(E-value). A hit has equal probability of being in the same clan as a different clan when the E-value is 0.01 ($\log_{10} = -2$). When the E-value is 10^{-5}, the probability that a sequence belongs to the same clan is >95%. In this article, we define a Pfam assignment to be a strong assignment when its E-value is $\leq 10^{-5}$.

The same analysis was performed for the HHsearch alignments, using a set of 5387 Pfams with E-value of $\leq 10^{-4}$. Figure 1b shows the classification functions for the HH hits. When HHsearch's E-value of $\geq 10^{-4}$, the probability of being in the same clan is >95%.

FATCAT provides a *P*-value for the significance of the structural similarity between two proteins. We ran FATCAT on all pairs of structures in the set of 5134 PDB structures used for the evaluation of HMMER. The *P*-values were also divided into two classes: same clan and different clan. The probability density and classification functions are shown in Figure 1c. When the *P*-value is <0.001, the probability that two structures are in the same clan is >95%. FATCAT suggests a *P*-value cut-off of 0.05 for two similar structures with 95% confidence interval. Our cut-off is more restrictive because we are trying to identify not only similar structural patterns but also probable homology. In this article, we use the more strict *P*-value cut-off of 0.001.

3.2 Comparison of HMMER and HHsearch

To determine the relative utility of HMMER3 and HHsearch for assigning Pfams to the PDB, we performed alignments of PDB sequences and UniProt sequences against the Pfam HMMs using both programs. We first calculated PSI-BLAST profiles using one round of search on UniProt90 for all unique protein sequences in the PDB. From these profiles, we determined consensus sequences using the most common amino acid in each position (given in the PSI-BLAST profile output in percentage terms) and the highest PSSM scoring amino acid. The means and SDs of the sequence identities between PDB and PDB-percent and PDB-pssm are $65.1 \pm 10.0\%$ and $63.3 \pm 11.2\%$, respectively. We ran HMMER3 with the original PDB and UniProt sequences and their consensus sequences as queries against the HMMs in Pfam-A. A probability density estimate of the E-values from the original sequences demonstrated a maximum in the density at an E-value of 10^{-20}, whereas the consensus sequences were shifted to a mode at 10^{-25}. At a poor E-value of 10^{-5}, the original sequences have almost twice as many hits as the consensus sequences, which have all been shifted to higher statistical significance. A total of 38% of the consensus alignments were longer than the original-sequence alignments, whereas only 10% were shorter. Most of the shorter assignments occurred when the alignments of the consensus sequences are broken down into two or more fragments, when the original PDB sequence alignment was not. These fragments will be joined in the application of the greedy algorithm.

We applied the greedy algorithm on the alignments to Pfam from the original PDB and UniProt sequences alone and a set combining these alignments with those from the consensus PDB and Uniprot sequences. The set of assignments from the combined consensus and original sequence alignments contains 371 more Pfam-As than the original PDB and UniProt sequences and increases residue assignments by 4%.

It is often assumed that HMM–HMM alignments should be better than sequence-HMM alignments; therefore, we compared the Pfam assignments from the consensus sequences with HMMER3 (described earlier in the text) and those from HMM–HMM alignments produced by HHsearch using the general greedy algorithm by HMMER3 given these E-value cut-offs. HHsearch produced assignments to 2% more entries and 2% more sequences than HMMER3. HHsearch did produce a much larger number of weak hits, by >65% compared with HMMER3. This indicates that HHsearch may be most useful when HMMER3 fails to make any assignment. HHsearch produced 60% fewer assignments of repeats than HMMER3 did (1881 versus 5006). HMMER and HHSearch assignments to the PDB are compared in Table 1.

However, the most significant drawback of the HHsearch assignments was the tendency to assign more remotely related Pfams in a Pfam clan compared with the HMMER3 assignments. We compared Pfam assignments from HMMER3 and

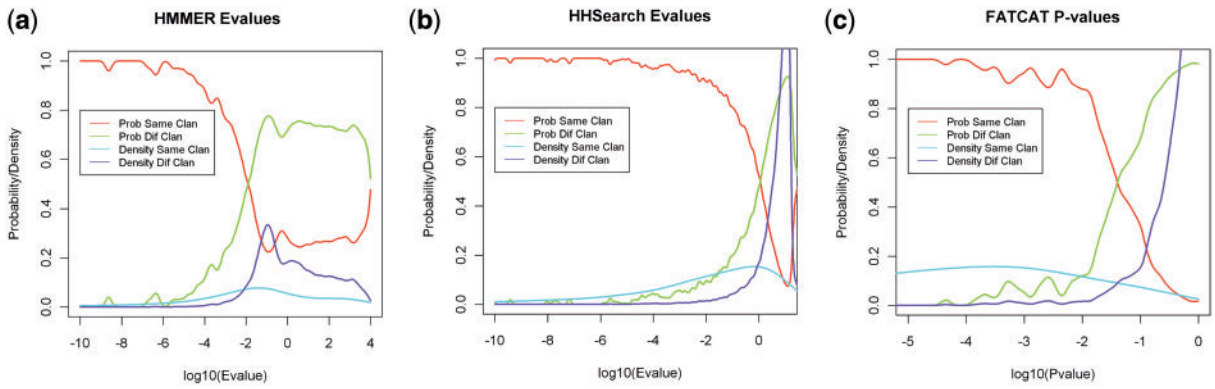

Fig. 1. Probability density functions and classification functions of Pfam E-values by HMMER and HHsearch and FATCAT *P*-values. (**a**) Pfam E-values from the exemplars and Pfam A v26 profile database by HMMER3. Only log10(E-value) from −10 to 5 are shown. (**b**) Pfam E-values by HHsearch. Only log10(E-value) from −10 to 1 are shown. (**c**) FATCAT *P*-values. Only log10(*P*-value) from −5 to 0 are shown

Table 1. Pfam assignments from consensus sequences and profile HMMs

	PDB–HMM	Cons-HMM	HMM–HMM
No. of Entries with assignments	68 489	71 051	72 781
No. of Entities[a] with assignments	91 870	95 772	97 885
No. of Domains	118 425	128 462	128 645
No. of Residues	17 494 555	18 620 323	18 991 436
No. of Pfams	5744	6016	6134
No. of Repeat Pfams	52	75	71
No. of Repeats	3128	5006	1881
No. of Weak hits (E-value of ≤10)	73 211	114 027	189 664

[a]Entities are unique sequences in a PDB entry. So an asymmetric unit that is a homooligomer of any size has a single protein entity sequence.

HHsearch with ≥80% overlap on the PDB sequence. A total of 120 517 (91.6%) of 131 585 domain assignments were with the same Pfam in the two sources, whereas 10 629 assignments (8.1%) were to different Pfams within the same clan. Only 439 assignments (0.3%) belonged to different clans. Because HHsearch is expected to find more remote hits than HMMER, it seems likely that the HMMER assignment is correct, whereas the HHsearch assigns a more remotely related Pfam. As we want to make correct assignments at the Pfam level and the clan level, we prefer the HMMER assignments over the HHsearch assignments, when both are statistically significant.

3.3 Structure alignments

Both HMMER3 and HHsearch produce many alignments to PDB sequences with weak E-values and/or alignments shorter than the Pfam model definition. We investigated whether we could confirm some of the weak hits and extend short alignments by comparing structures. We define exemplars as structure/Pfam pairs with good HMMER E-values ($\leq 10^{-5}$) or HH E-value ($\leq 10^{-4}$) to the Pfam and HMM coverage of at least 80%. A total of 81% of Pfams in the PDB have exemplars. The structures of weak Pfam hits were aligned to the exemplar structures in the same clan, including the Pfam of the weak hit. A total of 7381 structure assignments were added to our Pfam assignments by replacing the original alignment to the HMM by a transitive alignment through the structure alignment. The number of PDB residues aligned to Pfam HMMs for these sequences rises by 36%. An example is shown in Figure 2.

The ability of structure alignments to verify weak Pfam assignments varies with the statistical significance of the Pfam alignment. At E-values better than 10^{-10}, >80% of structure alignments are statistically significant (*P*-values of <0.001); these alignments are used solely to extend the Pfam domain assignments. At E-values of >0.01 and <10, about one-third of assignments are confirmed by structure alignment.

3.4 Pfam architectures for the PDB

Several domain assignments through Pfam and other classifications are publicly available. In Table 2, these assignments are compared with our Pfam assignments ('PDBfam') computed with the full procedure described in the 'Methods' section. Our method combines HMMER alignments to PDB and UniProt sequences and their PSI-BLAST consensus sequences, HHsearch alignments of HMMs of the PDB and UniProt sequences, FATCAT structure alignments for weak and/or short hits and a greedy algorithm. We provide the number of entries with at least one domain assignment by each method and the number of unique sequences with assignments and residues within unique sequences with assignments. In our Pfam assignments, there are 6379 Pfam-As. Our PDBfam assignments cover part or all of 98% of unique PDB sequences and 99.4% of all unique sequences with length >50. Pfam itself provides a file, pdbmap, which has domains listed for only 34 188 PDB entries. For these same 34 188 entries, we have made 64 832 domain assignments, whereas Pfam has only 38 927 assignments. We also detected about three times as many repeats. The coverage of PDB sequences from the other well-known data resources are also given in Table 2. SIFTS covers 87% of PDB sequences by mapping PDB sequences to UniProt sequences. As not all PDB sequences in the PDB map to UniProt, SIFTS does not cover a significant portion of the PDB. Pfam assignments from the RCSB website contain a list of Pfam hits from HMMER3

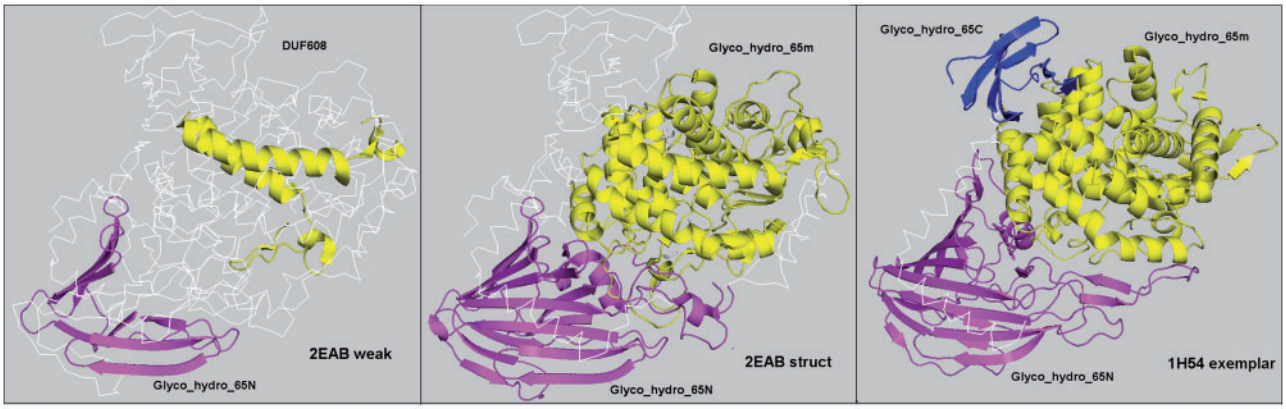

Fig. 2. Structure alignments verify and expand the Pfam assignments for the PDB entry 2EAB. Left: the initial Pfam assignments from the consensus sequences. Right: the Pfam assignments to the exemplar 1H54. Middle: the Pfam assignments of 2EAB after structure alignment

based on an E-value cut-off. Many of its assignments overlap completely in the PDB sequence, usually because they are different Pfams from the same clan.

To get a fair assessment of the RCSBs coverage, we used the same criterion we applied to our data—no >10 residues of overlap between Pfam assignments. The structure protein classification systems CATH and SCOP have much lower coverages because they are built manually and updated infrequently. Our Pfam assignments to the PDB cover a larger percentage of entries, unique sequences (entities) and residues than other available assignments. Perhaps more important than the number of entries or sequences with at least one assignment is the number of domain assignments made. Each assignment may lead to the identification of new interactions of known structure available in the PDB. The RCSB makes 68 767 domain assignments after removing overlapping assignments. In PDBfam, we have 79 201 domain assignments to unique sequences in the PDB, for an increase of 15%.

3.5 Split Pfam architecture assignments

One feature of structures that we have accounted for in our Pfam assignments is the presence of large insertions, relative to the multiple sequence alignments that define the Pfam models. Such insertions often result in separate alignments from HMMER3 or HHsearch covering different parts of the PDB sequence and different parts of the Pfam HMM. These are not accounted for on the Pfam website, where they are often listed as distinct architectures containing two copies of the Pfam (e.g., "IMPDH × 2") rather than one that is split by an insertion. In our current dataset, we have 5023 split domains (1.9% of the total) of which 966 domains are multichain domains.

Our split assignments come in a number of forms because of the ways that domains can be inserted or split up in the PDB sequences. Table 3 displays the different formats of split domains in our assignments, where X and Y are two Pfam IDs. The format of the chain Pfam architecture for proteins with an inserted domain is given on Line 1 as *Domain1[Start-End]_Domain2_Domain1[Start-End],* where *Domain1* is a split domain and *Start* and *End* are positions within the HMM. Line 2 shows the format when there is a long insertion that is not assigned to a Pfam. Line 3 represents those structures where two portions of the HMM are in reverse order in the PDB structures. Line 4 of the table denotes those structures where a Pfam is split between two different chain sequences in the structure (e.g. in this case, entity_id 2 and 3 in the PDB XML file).

3.6 Unassigned sequences and regions

There are 2043 unique sequences (from 4405 sequences including redundancy) that can not be assigned to any Pfam by our procedure. A total of 945 of these unique sequences (46%) have weak Pfam assignments—HMMER E-value of $>10^{-5}$ or HHsearch E-value of $>10^{-4}$ and no structure alignment with FATCAT *P*-value of $<10^{-3}$ up to E-values of 10. The remaining 1098 sequences (54%) are short peptides with length ≤21, of which 263 sequences are part of UniProt sequences, but not covered by Pfam strong hits. We investigated the properties of the 18% of the ~13 million residues in unique protein sequences in the PDB that are not assigned to any Pfam by our method. Figure 3 shows the histograms of the lengths of unassigned regions [either N or C terminal regions (Fig. 3a) or internal regions (Fig. 3b) and completely unassigned sequences (Fig. 3c)]. More than 90% of unassigned regions/entities are short peptides with length <50. Figure 4 shows the secondary structures for the unassigned regions. The last bar in each figure is for proteins and protein regions with Pfam assignments. The percentage of N and C terminal unassigned regions that is disordered is more than that for Pfam-assigned regions (10%). The secondary structures of the internal unassigned residues are closer to that of Pfam assignments but still have a higher proportion of coil and/or disordered residues. For completely unassigned sequences, the percentages of residues in coil or disordered are somewhat higher for shorter sequences than for Pfam assignments, especially in the amount of disorder.

3.7 PFAM interactions

Pfam assignments have been used previously to catalogue the physical interaction of different domain families within the PDB (Finn *et al.*, 2005; Stein *et al.*, 2011). With a larger and more accurate set of assignments, we have investigated the number of such interactions that are now present in the PDB.

Table 2. The coverage of the PDB in various sources compared with PDBfam

Data Source	No. of entries Entries[a] (%)	No. of unique sequences[b] (%)	No. of entries Residues[c] (%)	No. of Pfams[d]	No. of repeats
PDB	80 575 (100)	54 437 (100)	13 289 255 (100)	—	—
Pfam (PDBfam)	79 600 (99)	53 494 (98)	10 930 394 (82)	6311	3828
Pfam (v.26)	34 188 (42)	24 708 (45)	4 925 947 (37)	4874	1173
Pfam (SIFTS)	73 901 (92)	47 293 (87)	9 458 200 (71)	5643	855
Pfam (PDB)[e]	77 712 (96)	51 122 (94)	9 901 386 (75)	6073	2869
CATH (v.3.5)	51 334 (64)	30 862 (57)	6 906 806 (52)	—	—
SCOP (1.75A)	49 217 (61)	30 527 (56)	6 912 489 (52)	—	—

[a]The number of PDB structures with polypeptide sequences. [b]Number of unique PDB sequences (no two sequences of 100% identity and same length), excluding those with all Xs or ≤5 distinct amino acid types (943 sequences). [c]The number of residues in the domain regions of unique sequences. [d]The number of Pfam-A HMMs in each dataset. Number of entries, unique sequence coverage and number of residues includes Pfam-B assignments. [e]After removing those Pfam hits with >10 residues overlap to the one with best E-value.

Table 3. The split domains in Pfam in PDBfam

Format	No. of Domains	No. of Pfams	Pfam example (HMM)
X[s1-e1]_Y_X[s2-e2]	1087	X = 96 Y = 116	(ADK[1-122])_(ADK_lid)_(ADK[123-150])
X[s1-e1]_(#)_X[s2-e2]	2519	422	(Hpt[1-69])_(35)_(Hpt[70-88])
X[s2-e2]_X[s1-e1]	451	75	(CIMR[46-145])_(CIMR[5-48])
Multichain domains	966	100	((2)Trypsin[1-132]) (3)(Trypsin[135-220])

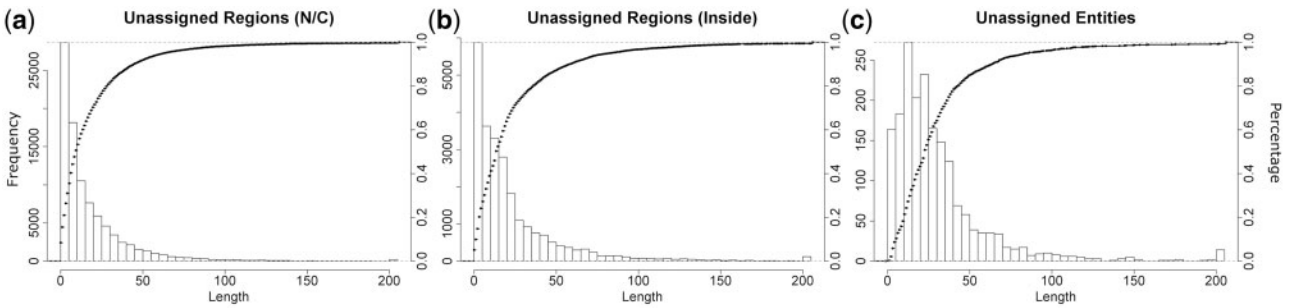

Fig. 3. Histograms of the lengths of the unassigned regions/entities in PDBfam. (a) Unassigned regions at N/C terminals. (b) Unassigned regions between two Pfam assignments. (c) Entities with no Pfam assigned. The curves represent the cumulative percentage as a function of increasing length (right axis)

Less than 50% of the asymmetric units of crystal structures correspond with the PDBs annotated 'biological assemblies' (Xu *et al.*, 2006) that either come from the authors of each structure or are assigned using the PISA software (Krissinel and Henrick, 2007). Although these biological assemblies are not 100% accurate (Xu *et al.*, 2008), they provide a better dataset to tally the number of Pfam interactions than the asymmetric units. For example, an asymmetric unit for a structure may be a monomer, whereas its biological assembly is a dimer. Such an interaction would be missed if only the asymmetric unit is considered. The converse may also be true—an asymmetric unit may consist of multiple chains, whereas the biological assembly is a monomer. An incorrect interaction is counted in this case.

The numbers of interchain and intrachain interactions of Pfams are given in Table 4. An interaction is defined if there are at least five pairs of residues with any atomic distance <5 Å or at least 10 pairs of residues with Cβ/Cα distance <12 Å and at least one atomic distance <5 Å. The column 'PDBfam (all BA)' gives our results for the biological assemblies in the entire current PDB. For instance, there are 3499 Pfams involved in homodimeric interactions between chains (or as heterodimers of two proteins containing the same Pfam domains). There are 3958 pairs of Pfams in physical interactions across two protein chains. If we consider interchain and intrachain relationships together, there are 3576 same-pfam interactions, and 6132 Pfam pair interactions involving 3982 Pfams.

The 3DID database also uses Pfam to determine the prevalence of domain interactions in the PDB (Stein *et al.*, 2011). We parsed the 3DID Pfam interactions directly from the text file (3did_flat_Apr_3_2011.dat) available from 3DID. This file

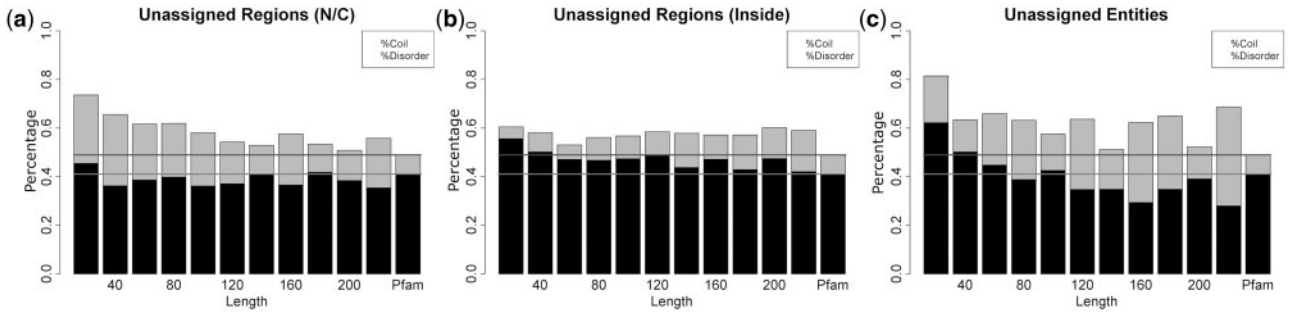

Fig. 4. Secondary structure of the unassigned regions/sequences and Pfam assignments in PDBfam. Disorder means the residues do not have coordinates in the PDB file, whereas coil includes all residues that are not α-helix or β-strand. The *y*-axis is the number of coil plus disordered residues divided by the total unassigned residues at that length range. A bar is 20 amino acids long, i.e. 1–20, 21–40, 41–60, etc. The last bar in each figure shows the values for Pfam assignments. The next to last column in each figure shows the values for all regions with length >200. The light grey horizontal line is the rate of coil in the Pfam assigned regions, and the dark grey horizontal line is the rate of coil plus disorder in the Pfam assignments

Table 4. Number of Pfams and Pfam pairs involved in interchain and intrachain interactions in the PDB in PDBfam and 3DID

	PDBfam (all BA)	PDBfam (BA)	PDBfam (ASU)	3DID (ASU)
Interchain				
Same-Pfam	3499	2839	3338	3087
Diff-Pfam pairs	3958	3267	3455	2306
Intrachain				
Same-Pfam	383	293	293	249[a]
Diff-Pfam pairs	3240	2236	2236	1271
Same-Pfams (inter and/or intra)	3576	2909	3382	3127
Diff-Pfam pairs (inter and/or intra)	6132	4720	4796	3049
No. of Pfams in diff interactions	3982	3336	3338	2658
No. of Pfams in same or diff interactions	5263	4445	4713	4265
No. of Entries[b]	47 458	35 449	35 449	35 449

BA, Biological assembly; ASU, asymmetric unit. [a]After removing those Pfams that are split domains in our Pfam assignments. [b]47 458 entries in current PDB ('all BA') that have one or more Pfam/Pfam interactions in biological assemblies. 35 449 PDB entries listed in 3did_flat_Apr_3_2011.dat after removing obsolete entries.

contains Pfam interactions for 35 449 entries. 3DID contains interactions present in asymmetric units and does not utilize the biological assemblies of the PDB. Table 4 compares our results with theirs. To accomplish this, we calculated the number of Pfam interactions in the same set of 35 449 PDB entries in both the asymmetric units [PDBfam(ASU)] and in the biological assemblies [PDBfam(BA)]. The 3DID results are given in the last column [3DID(ASU)]. Because we have more Pfam assignments for these entries, we have more Pfam interactions in the asymmetric units. However, when using the biological assemblies, the number of interchain Pfam interactions (both same-pfam and diff-pfam) are reduced from those of the ASU.

The interactions in biological assemblies provide more accurate estimates of our structural knowledge of how proteins interact with each other, and how this information might be used to investigate protein–protein interaction networks (Aragues *et al.*, 2005). In Table 5, we show the Pfams that interact with ≥30 other Pfams within the PDB. Several of these are domains that typically bind short peptides from other proteins, including MHC I, Pkinase, WD40 and efhand domains. Many are also repeat modules that serve as protein–protein interaction scaffolds.

4 DISCUSSION

One potential use for the assignments in PDBfam is in programs that perform searches of the PDB with a query sequence or structure and return lists of structures that contain similarities to the query above some threshold. Examples include fold recognition servers, such as FFAS (Jaroszewski *et al.*, 2005), and structure similarity search servers, such as FATCAT. Often the list of hits returned consists solely of the PDB entry and chain, residue numbers and sequence or structure alignments. It is, therefore, difficult to know whether the hits are related to each other or whether they share any functional relationship with the query. SCOP and CATH designations are sometimes provided, which solves the first problem, but SCOP and CATH represent less than two-thirds of the PDB, and their utility for this purpose is, therefore, limited. We believe that our Pfam domain assignments and associated clan information may be used to provide both relationships among the hits (shared Pfams or clans) and functional information of the hits (e.g. Pkinase).

Protein domain classification is also useful in the analysis of the interactions of protein domains with each other. Our ProtCID server provides information on clusters of similar interfaces between homologous proteins or protein pairs in multiple crystal forms in the PDB (Xu and Dunbrack, 2011). Assignments that are as accurate and complete as possible allow for better identification of biologically relevant interfaces in multiple crystal forms. Each Pfam in the PDB has a webpage in ProtCID listing all PDB entries that contain that Pfam, the chain architecture of the protein containing the Pfam, and the chain architectures of any other protein sequences in the same entry. Thus, looking up a Pfam in ProtCID provides information on other domains that the Pfam may interact with, either within the same chain or in protein complexes.

Table 5. Pfams with largest number of intrachain and interchain interactions with other Pfams in PDBfam

Pfam	No. of Interacting Pfams[a]	No. of Entries[a]	No. of Intrachain Pfams[a]	No. of Interchain Pfams[a]
V-set	178	1575	8	177
Ras	77	206	2	75
MHC I	69	516	2	69
Crp	67	86	57	30
Pkinase	52	480	17	41
Ubiquitin	47	99	13	36
Trypsin	46	519	9	43
EF_hand_5	42	209	12	37
WD40	40	166	15	27
fn3	36	70	16	28
efhand	35	314	23	22
I-set	35	152	18	25
Arm	35	62	12	26
HEAT_2	33	61	19	16
C1-set	32	1756	5	32
Fer4_7	30	66	12	21

[a]Pfams includes same Pfam–Pfam interactions within and between chains. All interactions are involved in annotated biological assemblies in the PDB.

Our assignments using Pfam have some limitations. We have defined a pipeline using a combination of available methods that is relatively efficient and can be applied automatically on a regular basis. It is possible that different choices might increase the representation slightly (e.g. different ways of defining the consensus sequences or the use of several structure alignment programs). But the coverage is high as it is, and exploring further avenues is likely to be a case of rapidly dwindling returns.

Pfam does not have models for some kinds of proteins. For instance, about half of the entities >250 amino acids that do not have Pfam assignments in our results are virus proteins, especially virus capsids. Pfam is organized into clans, and these may in most cases be useful for inferring structural relationships at the superfamily level. But many clans have HMMs of very different lengths; therefore, the structural domains are not likely to be completely consistent across clans. Also, it is likely the case that some superfamily relationships evident in SCOP and CATH are missed by the Pfam clans. We cannot make assignments to many peptides, either because the parent UniProt sequence is not known or does not exist or because the peptide belongs to the region of a protein that is not within a Pfam domain. It is likely that intrinsically disordered protein regions have less coverage in Pfam than folded domains, and many peptides bound to proteins come from these regions. Nevertheless, our assignments cover 99.4% of unique sequences in the PDB >50 amino acids. We believe they will be useful in numerous applications in structural bioinformatics and structure prediction.

Finally, the procedure we have outlined for PDBfam is sufficiently general that it may be used to make assignments of other domain classification systems to any set of target sequences. Because of the power of HMMs to represent profile families, such a classification of domains should be represented as a set of HMMs. With the program HHblits, this is straightforward to do even if the classification system is based on a small number of sequences. Further, structural information also significantly improves the results. If the target sequences are not of known structure, then they may be assigned to known folds through the sequence search methods used here. This should be done at stringent levels of statistical significance so as not to add another layer of uncertainty to the assignment process. If the domain classification system is structure based, then the other side of the equation is already satisfied. We can imagine a number of further applications that will be presented later, including Pfam assignments to human proteins and assignment of SCOP domains to the entire PDB on an ongoing basis.

Funding: NIH (grant R01 GM84453) and the Pennsylvania Department of Health (in part). The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of database programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Aragues,R. *et al.* (2007) Characterization of protein hubs by inferring interacting motifs from protein interactions. *PLoS Comput. Biol.*, **3**, 1761–1771.

Bairoch,A. *et al.* (2005) The universal protein resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Dessailly,B.H. *et al.* (2009) PSI-2: structural genomics to cover protein domain family space. *Structure*, **17**, 869–881.

Finn,R.D. *et al.* (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Finn,R.D. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.

Finn,R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, 211–222.

Hunter,S. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.

Jaroszewski,L. *et al.* (2005) FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Res.*, **33**, W284–W288.

Kahsay,R.Y. *et al.* (2005) Quasi-consensus-based comparison of profile hidden Markov models for protein sequences. *Bioinformatics*, **21**, 2287–2293.

Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.

Li,W. *et al.* (2000) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Orengo,C.A. *et al.* (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pandit,S.B. *et al.* (2002) SUPFAM—a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. *Nucleic Acids Res.*, **30**, 289–293.

Przybylski,D. and Rost,B. (2008) Powerful fusion: PSI-BLAST and consensus sequences. *Bioinformatics*, **24**, 1987–1993.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Schultz,J. *et al.* (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. USA*, **95**, 5877–5884.

Servant,F. *et al.* (2002) ProDom: automated clustering of homologous domains. *Brief. Bioinform.*, **3**, 246–251.

Sheather,S.J. and Jones,M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Series B Stat. Methodol.*, **53**, 683–690.

Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Sonnhammer,E.L. *et al.* (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.

Stein,A. *et al.* (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **39**, D718–D723.

Tress,M. *et al.* (2005) Domain definition and target classification for CASP6. *Proteins*, **61** (**Suppl. 7**), 8–18.

Velankar,S. *et al.* (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.

Wang,Q. *et al.* (2008) SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat. Protoc.*, **3**, 1832–1847.

Westbrook,J. *et al.* (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, **21**, 988–992.

Xu,Q. and Dunbrack,R.L.Jr. (2011) The protein common interface database (ProtCID)—a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic Acids Res.*, **39**, D761–D770.

Xu,Q. *et al.* (2006) ProtBuD: a database of biological unit structures of protein families and superfamilies. *Bioinformatics*, **22**, 2876–2882.

Xu,Q. *et al.* (2008) Statistical analysis of interface similarity in crystals of homologous proteins. *J. Mol. Biol.*, **381**, 487–507.

Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19** (**Suppl. 2**), II246–II255.

Zhang,Y. *et al.* (2005) Comparative mapping of sequence-based and structure-based protein domains. *BMC Bioinformatics*, **6**, 1–16.