BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 6 2010, pages 745–751
doi:10.1093/bioinformatics/btq034

Structural bioinformatics

Advance Access publication January 29, 2010

A fast and automated solution for accurately resolving protein domain architectures

Corin Yeats*,†, Oliver C. Redfern† and Christine Orengo
Department of Structural and Molecular Biology, UCL, Darwin Building, Gower Street, London WC1E 6BT, UK
Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Accurate prediction of the domain content and arrangement in multi-domain proteins (which make up >65% of the large-scale protein databases) provides a valuable tool for function prediction, comparative genomics and studies of molecular evolution. However, scanning a multi-domain protein against a database of domain sequence profiles can often produce conflicting and overlapping matches. We have developed a novel method that employs heaviest weighted clique-finding (HCF), which we show significantly outperforms standard published approaches based on successively assigning the best non-overlapping match (Best Match Cascade, BMC).

Results: We created benchmark data set of structural domain assignments in the CATH database and a corresponding set of Hidden Markov Model-based domain predictions. Using these, we demonstrate that by considering all possible combinations of matches using the HCF approach, we achieve much higher prediction accuracy than the standard BMC method. We also show that it is essential to allow overlapping domain matches to a query in order to identify correct domain assignments. Furthermore, we introduce a straightforward and effective protocol for resolving any overlapping assignments, and producing a single set of non-overlapping predicted domains.

Availability and implementation: The new approach will be used to determine MDAs for UniProt and Ensembl, and made available via the Gene3D website: http://gene3d.biochem.ucl.ac.uk/Gene3D/. The software has been implemented in C++ and compiled for Linux: source code and binaries can be found at: ftp://ftp.biochem.ucl.ac.uk/pub/gene3d_data/DomainFinder3/

Contact: yeats@biochem.ucl.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 27, 2009; revised on January 12, 2010; accepted on January 22, 2010

1 INTRODUCTION

An active field of bioinformatics pursues the identification and classification of protein domain families. This is critical to our understanding of how proteins evolve new functions and how domain family content can explain phenotypic variation between organisms. From an annotation perspective, newly sequence genomes can be scanned against libraries of domain families in order to predict the function of novel genes (Schug *et al*., 2002). However, due to both the huge number of sequences in the protein databases, and also the large number of domain families, it is essential to use fast robust sequence analysis and comparison techniques. Furthermore, unless manual validation is used when building profiles of families, a given query sequence can generate many conflicting and overlapping matches. These are the result of complex structural rearrangements and gene fusion events that have occurred throughout evolutionary history. Resources such as InterPro and Pfam (Finn *et al*., 2008; Hunter *et al*., 2009) apply extensive curation of their family models and the resulting predictions. In contrast, fully automated protocols remain useful in helping identify novel families that are then fed into the manual curation pipelines, rather than directly used for biological analyses (i.e. Pfam-B, ADDA and CHOP; Heger and Holm, 2003; Liu and Rost, 2005). However, using manual verification is very expensive and time consuming, and will become increasingly so in the era of high throughput sequencing.

Gene3D (Yeats *et al*., 2008), a member of the InterPro consortium, provides domain annotations based on the CATH structural domain classification (Cuff *et al*., 2009). CATH provides a manually validated set of domains derived from structures in the PDB (Berman *et al*., 2003), and these domains are grouped into superfamilies where there is evidence of common evolutionary ancestry from shared sequence, structural and functional features. Gene3D uses a similar approach to the SUPERFAMILY resource (Wilson *et al*., 2009), which is based on the SCOP structural classification (Andreeva *et al*., 2008). A set of structural representatives are chosen from each superfamily in CATH and used to seed an automatic iterative search process (SAM Target-2K; Karplus *et al*., 2003). The resulting multiple sequence alignment is then converted into a Hidden Markov Model (HMM). Each superfamily may be represented by more than one model, with an average of four per superfamily. These models are then used to identify probable homologous domains in protein sequences, and the resulting predictions merged into a final set of assignments. Gene3D aims to provide high quality structural domain annotations for the major genome and protein sequence databases, including Ensembl (Hubbard *et al*., 2009), UniProt (UniProt Consortium, 2009) and RefSeq (Pruitt *et al*., 2007)—see Supplementary Data for example genome coverage.

The annotation pipeline for determining a multi-domain architecture of a given query sequence (Fig. 1) can be considered to consist of three main steps:

(1) Generate a HMM model library to represent structural domains in CATH.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

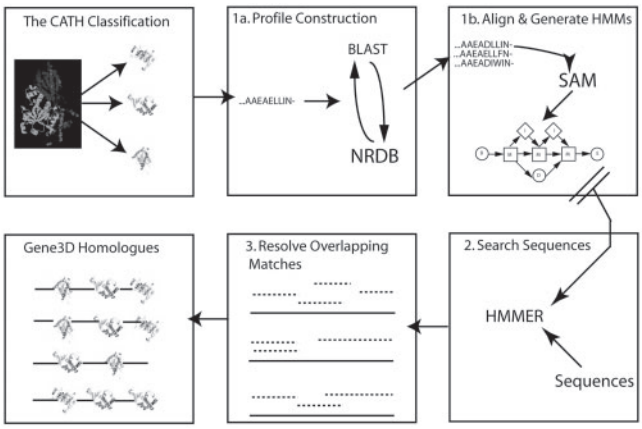

Fig. 1. Domain homologue identification. The mapping of structural superfamilies from the PDB to their homologues in sequence databases currently takes four steps. (1a) First, a representative set of homologues is gathered using the SAM Target T2K procedure to create a large seed alignment. (1b) These are used to generate a library of HMMs with the Hmmer software, which (2) are used to search sequence databases (for full list see Gene3D website). (3) The matches produced by the searches are then assembled into a non-redundant architecture and displayed on the website.

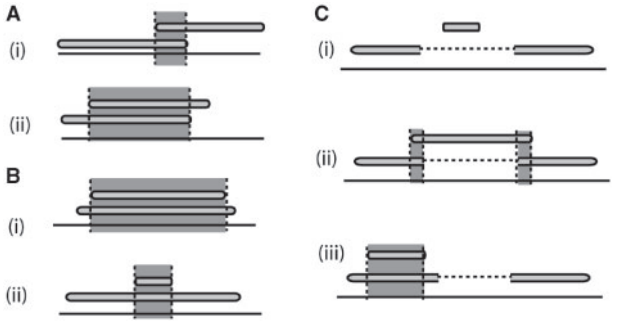

Fig. 2. Three types of match overlap (A, B and C, respectively). A and B show the types of overlap for matches (grey boxes) to a sequence (plain black line) consisting of a single segment. (**A**) Terminal overlaps and (**B**) embedded overlaps. In both the cases we wish to choose whether the overlap is allowed (and hence both matches may co-occur in the final MDA) or disallowed (one or the other or none may occur in the final MDA). Terminal overlaps may be due to two closely proximal domains with imprecise boundary assignments [A(i)], or two related models suggesting different boundaries for the same domain [A(ii)]. We set an overlap length threshold to segregate the two types. Embedded overlaps can be caused by inserted domains [B(ii)] or contrasting predictions [see B(i)]. (**C**) Discontinuous and embedded domains, are frequently identified by CATH. To identify such cases, the alignment of the match to the model is examined and long insert regions (>30 residues) are excised. This results in domains that are made up of more than one discrete segment (gaps indicated by dotted line). Resulting terminal overlaps are resolved as for single segment domains [i.e. C(ii)], while any remaining embedded overlaps are considered 'disallowed' overlaps [C(iii)].

(2) Scan the query against the model library using HMM-sequence search software (e.g. HMMER2).

(3) Resolve significant model matches into a prediction of the domain boundaries and superfamily content for the query sequence.

Step three (referred to as the DomainFinder or 'DF' step) at first appears straight-forward, but in reality a query sequence can match many models with high significance, producing a set of conflicting predictions of domain boundaries and superfamily assignments (for types of overlaps, see Fig. 2). This issue is more trivial for single domain proteins where one can simply take the match with the best score (e.g. *E*-value or alignment score), but for multi-domain proteins there can be hundreds or thousands of putative predictions. Multi-domain proteins are believed to make up more than two-thirds of known proteins (Ekman *et al*., 2005), with some consisting of more than 10 domains. Dealing with these cases is a high priority for resources such as Gene3D.

Another cause of complexity comes from the existence of structural domains that are discontinuous in sequence. This situation arises when one domain is inserted into another, splitting the original domain into multiple sequence segments. Around 20% of domains in CATH, distributed across 15% of superfamilies, are considered to be discontinuous (Redfern *et al*., 2007), which means these cases must be accounted for in order to maximise annotation coverage of the genomes. We employ a simple heuristic that identifies long inserts within the alignment of the HMM to the sequence, splitting the match into multiple segments and permitting other domains to be assigned into the gap (see Fig. 2).

Our previous method (DF2) for resolving HMM model matches into a unified multi-domain architecture (MDA) for protein sequences was both intuitive and simple—we will subsequently refer to this as the 'Best Match Cascade' (BMC) protocol. It also only allowed a single residue overlap between different matches, primarily due to the complexities of resolving discontinuous domains, and a risk of including false positives. However, after examining the performance of this method with a novel benchmark measure, we discovered a significant proportion of missing domains in our resolved predictions, despite the fact that significant matches to these domains were identified by the HMM scanning process. As a consequence, we developed a novel approach that represents sets of potential matches as weighted graphs, and uses the Cliquer (Ostergard, 2002) heaviest clique-finding (HCF) algorithm to assign the most probable MDA to a given sequence. The underlying hypothesis of the approach is that the combination of matches that produces the highest combined score is the most likely to accurately represent the MDA.

We demonstrate here that this HCF algorithm performs better in all benchmarks than the BMC approach. We also show that to account for the inaccuracies of boundary prediction from automatically generated HMMs, it is essential to allow MDAs where the matches overlap on the query sequence. To remove these overlaps, we have implemented a simple, yet accurate, resolution protocol to predict domain boundaries once the optimal set of domain assignments is determined. By improving both aspects of our previous DF2 algorithm, we show that we can achieve a very high degree of accuracy on our benchmarks, and that this performance translates to a substantial increase in domain annotations for sequences in UniProt and Ensembl. While developed for multi-domain architecture prediction by Gene3D, our new algorithm (DF3) can be applied to other sets of overlapping predictions to determine a minimal representative set, such as other domain family resources or template selection for structural modelling.

2 METHODS

2.1 Creating CATH domain predictions

We searched the PDB benchmark, UniProt and Ensembl sequence data sets using the CATH v3.2.0 superfamily representative HMM library and hmmpfam (Hmmer 2.3.2g, default settings and a per-match *E*-value limit of 0.001). The CATH HMM library consists of 8871 HMMs for 2178 superfamilies. The library can be downloaded from http://release.cathdb.info/v3.2.0/hmmer.lib.gz; for details of their construction see Sillitoe *et al*. (2005).

2.2 Pre-processing of CATH HMM matches

The alignments of all model-sequence alignments with an *E*-value < 0.001 were examined for gapped regions (i.e. insertions) longer than 30 residues. These were taken to indicate that the domain is likely to be discontinuous in the query sequence. This may be caused either by natural variation or due to errors in the protein sequence. In these cases, the match is split into multiple segments. For instance, if the model aligns to residues 1–200 in the query sequence but contains a 50-residue insert between positions 70 and 119, we create a single domain prediction with two segments: 1–69 and 120–200.

2.3 Selecting the optimal domain prediction

The BMC and HCF algorithms described below are summarised in Figure 3.

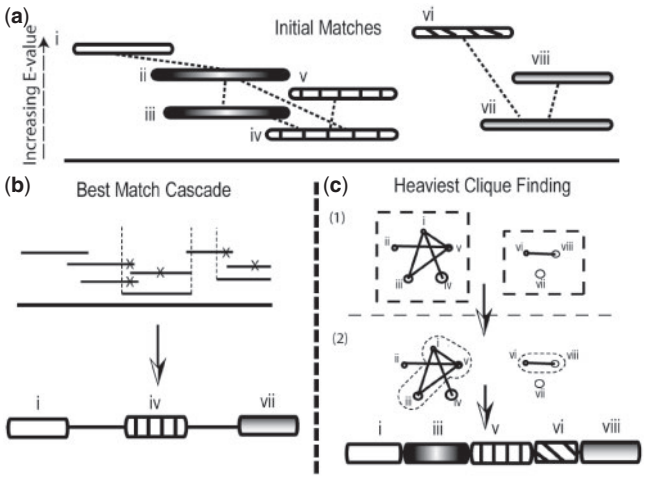

Fig. 3. The match selection methods. (**a**) A hypothetical dataset. Each HMM match is indicated in order along the sequence (*x*-axis), with more significant matches lower on the *y*-axis, and each identified by a roman numeral. The shading indicates the superfamily assignment (homology group). Dotted lines join matches that significantly overlap with each other. (**b**) The BMC process. The most significant match is selected, and significantly overlapping matches removed. This step is repeated until no matches remain, producing the final MDA of three domains (i–iv–vii). (**c**) The HCF process: (i) Matches are assembled into single-linked clusters of significantly overlapping matches. Each cluster is then formed into a weighted graph by making matches nodes, and drawing edges between nodes that do not overlap. Nodes are weighted by the significance of the match (node size in the figure, large is more significant). (ii) The Cliquer algorithm is used to identify the heaviest true cliques in each set (highlighted) and the MDA assembled from the selected matches. This approach yields an MDA of five domains in this instance (i–iii–v–vi–viii). By selecting the less significant but shorter matches, a combination with a greater overall significance is found.

2.3.1 BMC All significant ($E < 0.001$) matches to a given query sequence are assembled into a list, ordered by *E*-value. The top match is assigned to the query and then the list traversed until the next highest non-overlapping (by more than a given number of residues) match is found and this assignment is added. Matches that overlap (above a threshold value) with any previously selected one are ignored. The process continues until the list has been traversed and all matches are either added to the MDA or discarded.

2.3.2 HCF Three steps underpin this approach. Firstly, the matches are assembled into 'chains' of overlapping domains; two matches are linked into a chain if they overlap by more than the assigned threshold (dotted lines in Fig. 3a), and each chain resolved independently [boxes in Fig. 3c(i)]. This initial step essentially splits the problem of finding the optimal match combination into subsets, minimising the search space. Secondly, each subset of matches is then formed into a weighted graph by considering each match to be a node, weighted by the $-\log(E\text{-value})$, and drawing edges between matches that do not overlap by more than the assigned threshold; i.e. if they are both allowed to appear in the final MDA assignment. Thirdly, the Cliquer algorithm (Ostergard, 2002) is then used to identify the maximally weighted clique in the graph, which is the combination of assignments that gives the highest total score. The final MDA is then calculated by combining the optimal sub-graphs (i.e. solutions) from each subset. Overlapping matches, within the allowed threshold, may still exist at this point. These are resolved using the procedure described in section 2.4.

2.4 Dealing with overlapping domain matches

2.4.1 Optimising allowed match overlap Two types of 'allowed overlap' threshold were tested—using a maximum percentage overlap and a maximum fixed length overlap. For the fixed length thresholds, matches are considered as having an allowable overlap, and hence treated as 'non-overlapping' for determining the MDA, when the following criteria is met: if the overlap is less than *n* amino acids in length ($0 < n < 40$ tested; $n = 30$ used where not otherwise indicated), and the overlap comprises <50% of the match. For the percentage thresholds an overlap was allowed if it was less than *n*% of the length of the shortest match (0–50% tested). The performance of both approaches was comparable, so most analyses were carried out only using the fixed length threshold.

2.4.2 Resolving overlapping matches to produce non-overlapping predictions Once the optimal set of domain assignments has been found (either by BMC or HCF), it is necessary to resolve any overlapping matches to provide a single non-overlapping MDA and thus the final domain boundaries for the query. In each case one of the following approaches is used, as appropriate. (i) If a segment of a discontinuous domain is entirely embedded within another match then the segment is removed. (ii) Otherwise the overlap is a straightforward partial overlap, and the overlapping segment divided equally between the two domains; if the segment is an odd number of residues the extra residue is given to the C-terminal domain (see Supplementary Data for a graphical description).

2.5 Creating benchmarks

2.5.1 CATH benchmark data The sequences of all chains in the PDB for which every domain had been classified to a superfamily in CATH were retrieved along with the associated location of each domain—a total of 75 655 chains. However, this dataset is highly redundant at the sequence level, due to the composition of the PDB. Using CD-HIT with default settings and a 90% sequence identity threshold, we obtained 11 394 representative chains, including 2523 multi-domain chains—reducing the dataset size by ~85%. This non-redundant chain set was then scanned against the CATH HMM library as described in section 2.1, creating a list of matches as a starting point for MDA prediction. For the analyses presented in the article, only the multi-domain proteins were considered, since the accuracies on single domain proteins were uniformly extremely high (data not shown).

2.5.2 Distant benchmark One flaw with the benchmark set described above is that every domain in the query sequence could have a 'trivial' match from a model seeded either from itself or a very close relative (>35% sequence identity). To test that the results hold for sequences that are highly diverged from the model seed domain, as is generally true for sequences in UniProt, we constructed a 'distant benchmark'.

Using sequence similarities calculated using the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970) for each domain in the PDB benchmark set, we identified every model that had been seeded from a domain with >35% sequence identity. Matches from those models overlapping the domain were then excluded from the prediction set. If there were two homologous adjacent domains, and the match overlapped both, it was considered to be predicting the existence of the domain it overlapped the most.

For some PDB sequences, the removal of matches meant that it was no longer possible to deduce the correct MDA. These were removed from the benchmark set. This left a total of 1318 multi-domain chains to analyse.

2.5.3 'Gapped' benchmark For the PDB sequences used in the benchmarks here, the structural composition of the chain has been completely defined. However, this is by no means true for sequences in Uniprot, which may contain globular domain families that are not yet represented in PDB or CATH. To test that HCF would still perform as expected, and not over-predict the number of domains by filling in these gaps, we created a set of benchmarks with a subset of superfamilies removed. This was done by randomly dividing the superfamilies into five groups, and then creating five benchmark sets for which one group of superfamilies had been 'removed' from each. To 'remove' a superfamily all instances of that superfamily were removed from both the benchmark and prediction data. Any sequences that were left with no 'known' MDA, because all the domains belonged to excised superfamilies, were also removed. Between 1600–1923 proteins remained in each sample, and the performance averaged over all five samples.

2.6 Measuring prediction accuracy

2.6.1 Quantifying MDA accuracy The predicted multi-domain architecture was considered correct if the superfamily types and the number of occurrences of domains were predicted correctly, irrespective of whether the number of segments or the boundary positions were correct. The overall accuracy is then calculated over the whole dataset. If too many domains of the correct superfamilies were predicted then this was termed a 'semi-false positive', while if the extra domains belong to an incorrect superfamily it was termed a 'false positive'; if the MDA was missing a domain it was a 'false negative'. A single predicted MDA could potentially contain errors in all three classes.

2.6.2 Residue accuracy measure The MDA accuracy gives a good measure of domain content, but does not take into account the accuracy of the domain boundaries. To better quantify the accuracy of our predictions we calculated the percentage of residues that were assigned to the correct superfamily or no superfamily (i.e. linking regions).

2.7 Performance of DF2 and DF3 on UniProt and Ensembl

The final test is to demonstrate that the new method will improve the annotations generated for Gene3D, as compared to previous releases. Domain matches were obtained for UniProt release 14 (5 213 603 unique sequences) and Ensembl v49 (858 815 unique protein sequences) as described in Section 2.1. The matches were resolved using the original protocol of DF2—a BMC method with one residue overlap threshold—and the new DF3 protocol—Heaviest Clique Finding and a 30 residue overlap threshold. A tiny minority of sequences with a large number of matches (<0.005%) generated sub-graphs of more than 6000 edges, normally with extremely high densities.

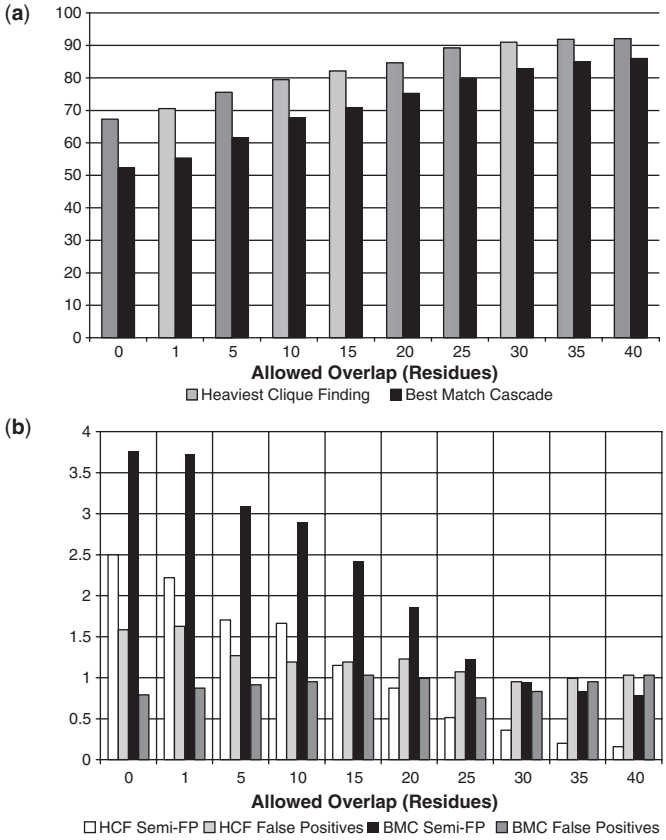

Fig. 4. Comparison of HCF and BMC accuracy. (**a**) Percentage of multi-domain architectures (MDAs) assigned correctly (see section 2.6.1) by Heaviest Clique Finding (HCF, light grey) and Best Match Cascade (BMC, black), as calculated on the main benchmark (only data for multi-domain proteins shown). The allowed match overlap is varied along the *x*-axis. (**b**) Most of the errors were due to false negatives, but there was some variation in the types of false positive at different overlap thresholds. HCF semi-false positives: white; HCF false positives: light grey; BHC semi-false positives: black; BHC false positives: dark grey. Only data for multi-domain proteins is shown. The nature and types of false positives are discussed in more detail in the Supplementary Data.

To maintain a reasonable processing speed these graphs were resolved using the BMC approach instead of HCF.

3 RESULTS

A CATH benchmark dataset was constructed from a representative set of sequences from the PDB, in order to evaluate the relative performance of the BMC and HCF approaches, as well as the effect of varying the allowed overlap threshold. If the domain assignment and boundary resolution process was working accurately, we would expect to be able to reconstruct the manual assignments by scanning the PDB benchmark set against the CATH HMM library. Using the same set of HMM results (the 'prediction set'), we tested the two approaches against each other, along with varying the allowed overlap parameter.

Figure 4a shows a comparison of the accuracy of the two methods with respect to correctly predicting the multi-domain architecture. When HMM matches are not allowed to overlap (i.e. threshold = 0),

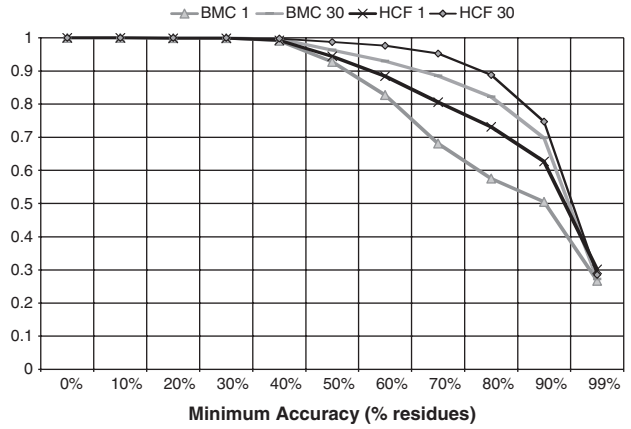

Fig. 5. Residue accuracy comparison. The residue accuracy is the percentage of residues determined correctly as according to the source CATH assignments (only multi-domain data shown). This measure reflects the accuracy of boundary assignments more reliably than attempting to determine equivalence between predicted segments and curated segments. Plotted on the *y*-axis is the percentage of chains for which DF2 or DF3 achieve at least the accuracy indicated on the *x*-axis. Shown are the accuracy distributions for the BMC (grey) and HCF (black) methods with allowed overlaps of 1 (triangles and crosses) and 30 residues (dashes and diamonds). Using HCF with a 30 residue overlap achieves >90% accuracy for 75% of chains, while only 50% of chains are at least that accurate when DF2 is used.

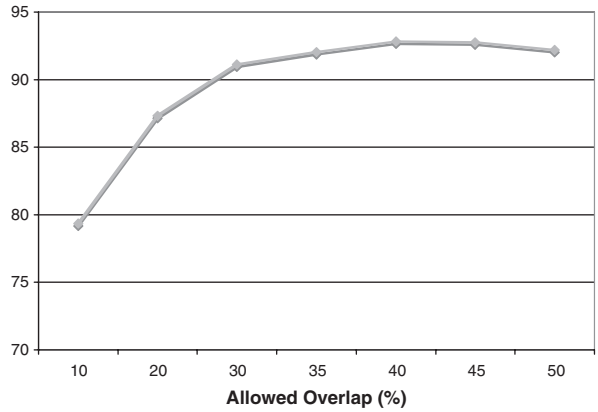

Fig. 6. Percentage allowed overlap accuracy. Using a percentage overlap threshold also performs well on this benchmark set. However, <0.5% (see Supplementary Data) of the matches in this benchmark are to non-homologous families, and so it is not clear that this method would be tolerant of noisy data (see Fig. 7).

the BMC method only achieves 52% accuracy versus 68% for HCF. Increasing the overlap threshold makes a substantial difference to the percentage of correct predictions for both BMC and HCF, but relies on the new overlap resolution mechanism in DF3.

An optimal overlap limit of 30 residues increases MDA prediction accuracy from 68 to 91% for HCF, while BMC accuracy improves from 52 to 84%. The superior performance of the HCF algorithm largely arises from a decrease in false negatives (i.e. missing domain assignments), although there is also some improvement in the false positive rates (Fig. 4b). This improvement is confirmed by examining the residue assignment accuracy (Fig. 5): the original version of DF2 (where overlap = 1) only annotated 50% of sequences with >90% residue accuracy, while DF3 (where overlap = 30) achieved this level of accuracy on 75% of sequences.

An alternative approach for determining the allowed overlap is to use a percentage of the match length. As can be seen in Figure 6, similarly high levels of accuracy are achieved by using a threshold between 30 and 40% (using the HCF approach). However, as noted in the Supplementary Data the true false positive rate for the Gene3D models over this benchmark are very low; we suspect that a percentage threshold will be less tolerant with noisy predictions. If a match can be overlapped by up to 80% of its length (40% at each end) then erroneous matches that do not align with the boundaries of correct ones are more likely to be incorporated.

We have also tested whether limiting the prediction of discontinuous domains to those superfamilies for which they have been observed would improve the accuracy of the algorithm. However, despite our concern that this might in fact over-fit to the benchmark data set, it slightly reduced overall performance (data not shown).

The CATH benchmark, whilst useful for optimising parameters and the core algorithm, does not completely emulate a real-world situation: scanning large, highly divergent, sequence databases for which the nearest structural representative may have a very dissimilar sequence. To better re-create this situation we generated two smaller benchmark sets to capture key aspects of this variation, as described in the section 2.5.2 (the 'Distant Benchmark') and 2.5.3 (the 'Gapped Benchmark'). The results on the multi-domain proteins show a similar distribution and profile to the original benchmark (see Fig. 7a and b), suggesting that the results for the main benchmark can be considered to be generally representative of the relative accuracy of the two approaches.

The benchmark results above give us confidence that using HCF and a 30 residue overlap limit—which we now term DF3—will produce more accurate architectures and allow us to detect many more domains. To test this we analysed their performance on large-scale sequence databases.

To examine whether the new approach will translate into a significant improvement to the Gene3D resource, we searched a sequence database comprising UniProt and Ensembl with Hmmer and the CATH v3.2.0 HMM library. The matches were then resolved using both the old protocol, DF2 (BMC with a one residue allowed overlap) and the newly defined DF3 (heaviest weighted clique with a 30 residue allowed overlap). In a very small number of cases (<0.005% sequences) a graph was produced that contained more than 6000 edges—for these, the HCF approach took at least a day and sometimes much longer to calculate. Hence, matches for these query sequences were instead resolved using BMC to keep the computation time feasible. Both DF2 and DF3 ran across the whole set within two days on two dual core 3.0 GHz Intel Xeon CPUs. The DF algorithms can be easily parallelized, and so can be distributed across multiple CPUs. In practice, the overall speed is limited by the time taken to solve the single most complex graph.

The MDAs predicted by DF3 for UniProt contained 600 000 more domains than those predicted by DF2 (an increase of 15%) and a similar increase for Ensembl (20%) compared to DF2. The majority of new domains were restricted to just a couple of hundred superfamilies, though around half (~1000) of all superfamilies showed a clear increase, and <1% showed any decrease (see Fig. 8). Furthermore, whilst all of the matches to one superfamily were lost

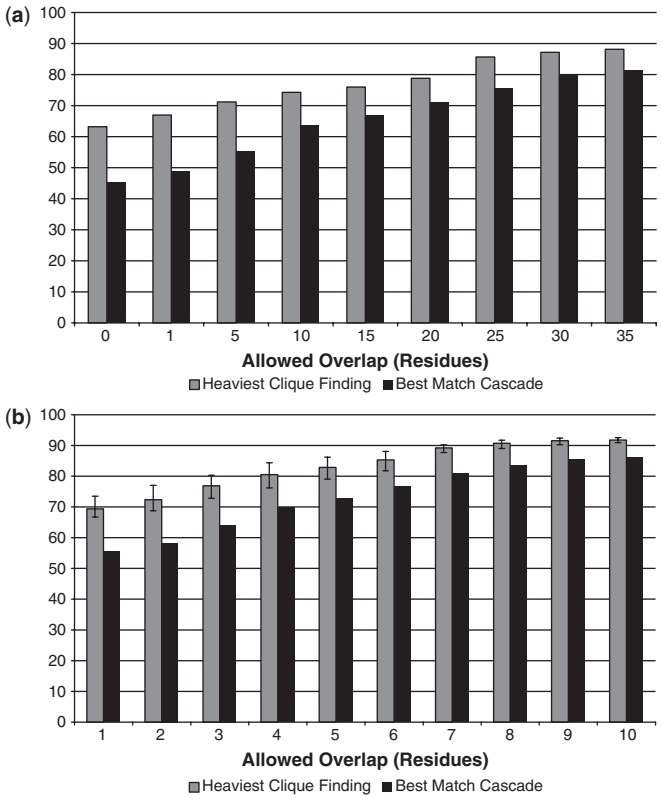

Fig. 7. Accuracy over the 'difficult' benchmarks. (**a**) A second benchmark, allowing only matches from HMMs constructed from distant homologues, shows that heaviest clique finding outperforms a best match cascade with 'distant' data. It also confirms the importance of using an appropriate overlap cut-off. (**b**) Artificial gaps were introduced in the benchmark and prediction set by removing randomly selected superfamilies (see Methods section), and the benchmark re-calculated. The MDA accuracy shown is the mean across the five benchmark subsets. The error bars represent the maximum and minimum accuracies for each method and test set.

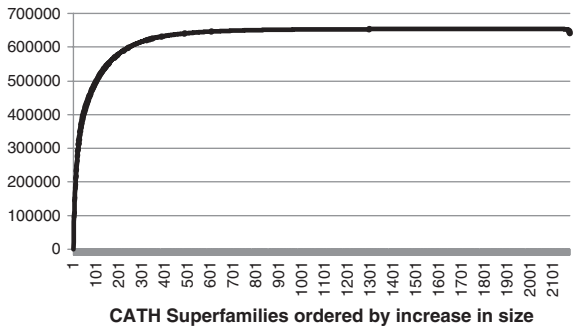

Fig. 8. Extra annotations for UniProt. Ordering the CATH superfamilies by the change in family size (*x*-axis), largest first, and showing a cumulative total (*y*-axis) clearly shows that most of the newly identified domain annotations belong to ∼10% of CATH superfamilies. However, about half of the superfamilies make some contribution to the increase.

by DF3, 26 superfamilies that previously had no matches gained representatives (see Supplementary Data for list).

The superfamilies that gained members showed no obvious shared characteristics, other than that they usually occur in multi-domain proteins. For example, there are a couple of viral structures from polyproteins, some transmembrane families, and some domains that are present in very common/universal protein families. A list of the 'Top 20' increases by numbers of sequences and numbers of domains is shown in the Supplementary Data.

Of the superfamilies that lost members, many belonged to the 'helical orthogonal bundle' architecture (CATH: 1.10), and contain small two helical hairpins. These super-families show significant cross-matching, and it can be hard to separate convergent and divergent evolution. Essentially, in these cases the structural fold is clear, but the precise evolutionary relationships are difficult to determine.

Using the Ensembl data set of eukaryotic genomic proteins, we were also able to see how the distribution of different superfamilies changes across the tree of life. This shows a similar behaviour to the size changes, confirming that a small number of superfamilies have had their identified taxonomic range substantially expanded, while for the majority the change has been slight (data not shown).

4 DISCUSSION

Accurate prediction of multi-domain architectures can be extremely useful for many function and network prediction methods, including phylogenetic profiling, gene fusion detection, protein–protein interaction inheritance, and annotation by homology transfer. However, current automated approaches are of unknown accuracy or are compared with Pfam or InterPro, where manual curation requires substantial investment. Both resources also contain many families that are not based on single domains, which in turn creates problems when benchmarking domain-related methods, such as *ab initio* domain prediction algorithms. Gene3D aims to be a gold-standard provider of domain family assignments, of equivalent accuracy to the manually-curated resources. So we set out in the first instance to define a useful measure of our accuracy in predicting multi-domain architectures, and then to improve on those predictions. Since it is not possible to accurately estimate performance over UniProt, we have generated a PDB-based benchmark that provides a means of comparing different algorithms (i.e. BMC versus HCF). Furthermore, we show that the results are likely to hold true for scanning large protein sequence repositories, by simulating cases of uncharacterised domain superfamilies and the absence of close relatives.

The benchmarks suggest that using DF3—HCF and 30-residue allowed overlap—leads to a 35% increase in accuracy compared to the old DF protocol for predicting multi-domain architecture from a list of matches to a HMM model library. Including single domain proteins, the overall accuracy is ∼95%, with many of the errors due to erroneous structures and assignments in the PDB and CATH. Most of the increase in multi-domain prediction accuracy is due to the loss of false negatives, and so would be expected to translate into an increased number of domains found in the sequence databases. As expected, the number of domains found in UniProt and Ensembl increased by ∼15–20%. As of Gene3D version 8 all annotations have been generated using DF3.

There remains some room for improvement on the specific failures, and many common MDA archetypes, especially transmembrane proteins, are simply not present in the PDB.

Hence, to further characterise and improve the accuracy of our predictive method will require a deeper analysis, and closer manual investigation of predictions, and the construction of artificial sets that better represent the protein world.

However, DF3 provides a robust prediction resolution process, and further significant increases in accuracy are likely to come from improvements in the construction of the domain superfamily profiles, and the HMM scanning software itself. DF3 is fast and accurate, and can be run on large-scale sequence databases and with large model libraries. The mechanism of building a weighted graph and finding the heaviest clique is a general solution for creating a minimum set of best representatives given a set of overlapping predictions, and most aspects of the protocol can be adapted as needed. An example use could be optimising the selection of templates for comparative modelling of multi-domain protein structures. Alternatively, it can be used to integrate assignments from multiple resources, by using an appropriate weighting mechanism. For instance, it could be easily adapted to merge Pfam and CATH assignments. As of version v9.0.0 Gene3D provides CATH domain assignments for ∼55% of proteins, while Pfam covers ∼65%. Combining the two provides annotations for ∼75% of known protein sequences.

ACKNOWLEDGEMENTS

We would like to thank all of the CATH team for their advice and providing key datasets for this work, and particularly Jonathan Lees of Gene3D. Gene3D is a member of the BioSapiens and EMBRACE EU Networks of Excellence, the IMPACT consortium and the Midwest Consortium for Structural Genomics. We would also like to thank InterPro for their reviews of the Gene3D predictions.

Funding: European Commission Framework 6 programme [LSHG-CT-2003-503265 to C.Y.]; the National Institutes of Health Structural Genomics [DE-AC02-06CH11357 to O.R.]; and the Higher Education Funding Council for England [to C.O.].

Conflict of Interest: none declared.

REFERENCES

Andreeva,A. *et al*. (2008) Data growth and its impact on the SCOP database: new developments. *Nuicleic Acids Res.,* **36**, D419–D425.

Berman,T.F. *et al*. (2003) Announcing the world-wide Protein Data Bank. *Nat. Struct. Biol.,* **10**, 980.

Cuff,A.L. *et al*. (2009) The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.,* **37**, D310–D314.

Ekman,D. *et al*. (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.,* **348**, 231–243.

Finn,R.D. *et al* (2008) The Pfam protein families database. *Nucleic Acids Res.,* **36**, D281–D288.

Heger,A. and Holm,L. (2003) Exhaustive enumeration of protein domain families. *J. Mol. Biol.,* **328**, 749–767.

Hubbard,T.J. *et al*. (2009) Ensembl 2009. *Nucleic Acids Res*., **37**, D690–D697.

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

Karplus,K. *et al*. (2003) Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction. *Proteins Struct. Funct. Genet. B*, **53**, 491–496.

Liu,J. and Rost,B. (2004) CHOP: parsing proteins into structural domains. *Nucleic Acids Res.,* **32**, W569–W571.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search of similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Ostergard,P.R.J. (2002) A fast algorithm for the maximum clique problem. *Disc. Appl. Math.,* **120**, 197–207.

Pruitt,K.D. *et al*. (2007) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.,* **35**, D61–D65.

Redfern,O. *et al*. (2007) CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multi-domain protein structures. *PLOS Comput. Biol.*, **3**, e232.

Schug,J. *et al*. (2002) Predicting Gene Ontology Functional from ProDom and CDD Protein Domains. *Genome Res.,* **12**, 648–655.

Sillitoe,I. *et al*. (2005) Assessing strategies for improved superfamily recognition. *Protein Sci.,* **7**, 1800–1810.

UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*., **37**, D169–D174.

Wilson,D. *et al* (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.,* **37**, D380–D386.

Yeats,C. *et al*. (2009) Gene3D, Comprehensive structural and functional annotation of genomes. *Nucleic Acids Res.,* **36**, D414–D418.