Exploiting Protein Structures to Predict Protein Functions

Alison Cuff, Oliver Redfern, Benoit Dessailly, and Christine Orengo

Abstract The exponential growth of experimentally determined protein structures in the Protein Data Bank (PDB) has provided structural data for an ever increasing proportion of genomic sequences. In combination with enhanced functional annotation from sequence, it has become possible to predict protein function from structure. In this chapter we discuss a range of methods which aim to recognise enzyme active sites and predict protein-ligand interactions. We then focus on algorithms developed as part of the CATH database of structural domains, where an evolutionary approach is used to recognise proteins with similar functions. While protein domains that exhibit the same structural fold tend to display related functional activities, there are a several large domain structure superfamilies that show a high degree of functional diversity. In these cases, we have built novel tools (FLORA and GeMMA) which are able to effectively identify sub-families of functionally linked domains, where standard methods of homologue detection (e.g. sequence profile and global structure alignment) fail.

Introduction

Many approaches for assigning protein functions attempt to exploit the 3D structure of the proteins, either to recognise putative active site regions and binding sites (e.g. for known ligands such as ATP), or to identify structural homologues likely to possess similar functions. The prediction of protein function from structure has become increasingly valuable as a significant proportion [1] of structures solved by the structural genomics initiatives (SGI) lack functional annotation [2]. In addition, structure-based approaches are particularly important for predicting binding sites and/or catalytic sites for the purposes of protein engineering and targeting drugs (for reviews see [1, 2]).

C. Orengo (✉)
Department of Structural and Molecular Biology, University College London, London, UK
e-mail: orengo@biochem.ucl.ac.uk

D. Kihara (ed.), *Protein Function Prediction for Omics Era*,
DOI 10.1007/978-94-007-0881-5_6, © Springer Science+Business Media B.V. 2011

Protein structures are more likely to be conserved during evolution than their sequences and structural data has been exploited to classify protein domains into evolutionary superfamilies. Nearly 40 years after the launch of the Protein Databank (PDB), established as a repository of solved 3D structures, the two major structural classifications, SCOP [3] and CATH [4] currently comprise more than 100,000 domain structures from the PDB classified into less than 3000 superfamilies. Furthermore, recent analyses have shown that nearly 70% of domain sequences in completed genomes can be predicted to belong to these families using HMM-HMM and threading protocols [5].

Both SCOP and CATH also further classify homologous structures according to their folds or topologies where structures are assigned to the same fold group if they have equivalent secondary structures, connected in the same way and oriented similarly in 3D space. Domains sharing the same fold are not necessarily evolutionary related and both classifications consider other evidence from sequence similarities or shared functional properties before classifying homologues [6]. Currently less than 1500 folds are recognised in SCOP and CATH. However, the definition of fold is somewhat subjective as no quantitative definitions exist and different protocols, employing manual inspection, are used to capture related folds by the two classifications.

There is no strong tendency for functional conservation across fold groups. Martin and Thornton explored the relationship between fold and function [7] and observed that whilst many small fold groups, comprising single evolutionary superfamilies exhibited only one molecular function, the highly populated fold groups could encompass a wide range of different functions. For example, the TIM barrel fold contains domains with more than 400 GO molecular function terms. However, there is often a tendency for particular surface features to be associated with the domain function. For example, Rossmann folds tend to bind substrates in the cleft created by the chain crossover at the C-terminal ends of the strands in the central β-sheet. Whilst structures adopting TIM barrel folds typically bind substrates in the large pocket at the base of the β-barrel. Russell and co-workers described these common sites as supersites [8]. These supersites may hint at remote homologies but whatever the cause of the similarity, fold recognition can help in identifying residues that are likely to be functionally important.

Whilst grouping protein domains into evolutionary families is important for studying their evolution, it is also valuable for predicting the functions of uncharacterised proteins since many analyses have revealed conservation of functional properties, particularly molecular function, within protein superfamilies [9]. However, it is clear that the degree of functional conservation varies with the domain superfamily as some superfamilies have diverged considerably in their structures and functions during evolution.

In this chapter we review the challenges faced when exploiting protein structures to predict function and describe some of the approaches that have been developed to cope with these challenges. We focus in particular on global methods of structure comparison and methods, developed within our group, which perform structure comparisons across a superfamily to identify specific structural features that are highly conserved within functional subfamilies in the superfamily.

Divergence of Protein Structures and Functions During Evolution

Analyses of structural superfamilies have revealed that many superfamilies are structurally very highly conserved during evolution and that this is accompanied by considerable conservation of function [10, 11]. The CATH classification captures structural divergence by determining the number of structurally distinct subgroups within a superfamily. A structural subgroup clusters together domains whose structures can be superposed with a normalised RMSD of 5Å. This is defined as:

$$\text{Normalised RMSD} = \frac{(\text{max length}) \times \text{RMSD}}{N} \tag{1}$$

where maxlength = number of residues in the largest structure, and N = total number of aligned residues.

Many CATH superfamilies (45%) comprise a single structural subgroup. Appendix A (http://www.biochem.ucl.ac.uk/~cuff/appendixA.html) lists the structurally diverse superfamilies containing more than one structural subgroup and shows the number of distinct GO and EC terms that can be identified for each of these superfamilies.

Figure 1 shows that there is a correlation between the number of structural subgroups and the number of distinct functional categories identified within the superfamily. Previous studies have shown that 75 superfamilies (<4% of CATH superfamilies) have diverged highly in their structures and functions [9]. These superfamilies tend to be highly recurrent in the genomes accounting for nearly 40%

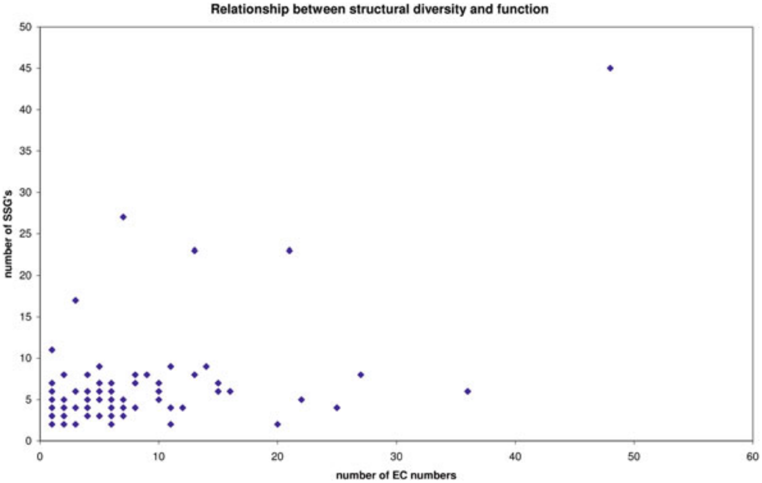

Fig. 1 Graph showing the correlation between number of structural subgroups and number of EC numbers identified within the superfamily

of predicted structural domains in completed genomes. They are also the oldest superfamilies, with the majority found in all three kingdoms of life and therefore probably present in the last common ancestor.

It is likely that the extensive duplication of these superfamilies within genomes and the divergence of structure and function in the duplicated or paralogous domains is accompanied by recruitment of the paralogues to different metabolic pathways or biological processes. Several studies have shown evidence for this in highly duplicated enzyme families [12] where homologues are frequently recruited to different pathways where perhaps they bring a chemical activity characteristic of their superfamily [13, 14]. Other large, diverse, superfamilies display conservation of parts of their ligands [15], possibly as the result of metabolic pathway retrograde evolution where the duplicated copy of an enzyme is recruited to catalyse the previous reaction in the same metabolic pathway [13, 16].

Extensive analyses of structural variation across these superfamilies has characterised the extent to which secondary structures are inserted and/or deleted during evolution. Whilst the secondary structures in the core of the domain tend to be very highly conserved, there can be considerable embellishment of additional secondary structures to this conserved core. Figure 2 illustrates structural divergence across some relatives from the large HAD domain superfamily, showing the conserved core and secondary structure embellishments.

Studies of the 31 most structurally and functionally divergent superfamilies showed that secondary structure insertions are generally distributed along the whole

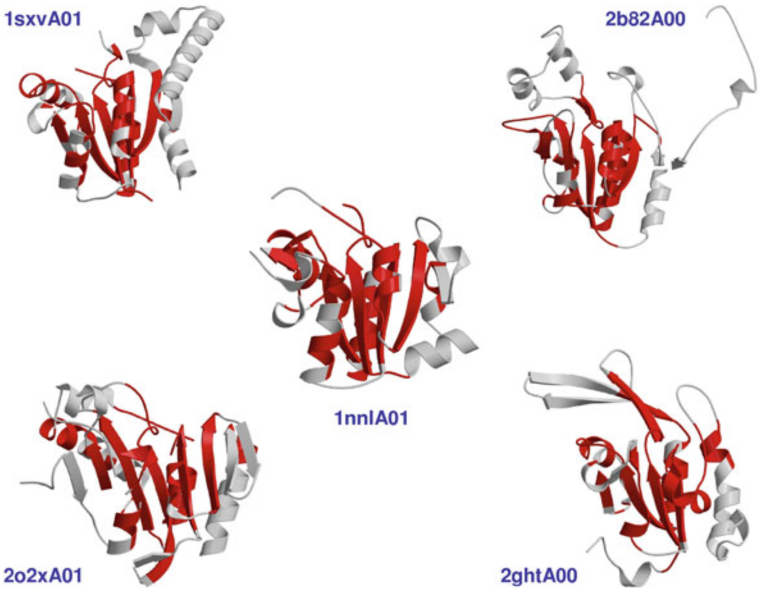

Fig. 2 Structural divergence among members of the mechanistically diverse haloacid dehalogenase (HAD) superfamily. The common structural core is coloured *red* and the structural embellishments are *grey*

length of the polypeptide chain with typically less than 3 being adjacent together in the sequence [11]. However, they accumulate in relatively few locations in 3D to give larger structural features. They were found to be modifying active site geometry or providing alternative protein interaction surfaces in relatives with different embellishments. Superfamilies adopting layered domain architectures such as αβα, αβ and β sandwiches appear more able to accommodate structural embellishments to the domain core [11].

Structural changes in domain relatives can also bring about changes in the domain partners and changes in the protein partners and oligomerisation states which can further modify functional sites or provide additional functional sites. Examples of these phenomena are given in Todd et al. [12], Reeves et al. [11] and Dessailly et al. [17]. Other evolutionary mechanisms causing structural change include circular permutations [18, 19], segment-swapping [18], addition of major structural embellishments to a conserved structural core [11], or more dramatic fold changes [20].

Despite the considerable divergence in structure observed in some superfamilies, some aspect of the function is generally conserved. Early studies by Todd et al. [12] revealed conservation of one or more chemical intermediates along the reaction pathway occurring in many highly diverse superfamilies. Such superfamilies, which are mechanistically diverse but share some common functional feature are being increasingly studied. The SFLD established by Babbitt and her group [21] now describes 6 such superfamilies and sequence diverse relatives within these superfamilies have been deliberately targeted by associated structural genomics initiatives to provide structures for characterising the diverse functional subfamilies. This work has been accompanied by extensive experimental characterisation of relatives within the superfamilies. Similarly the Structural Genomics Consortium (SGC), headed by the Edwards group in Canada, is targeting relatives from large superfamilies, highly expanded in human, to characterise relatives having different ligand specificities. These initiatives, which combine structural characterisation with biochemical studies, will be very useful in expanding the repertoire of diverse structural relatives within superfamilies with known functions which can be used to validate structure function prediction algorithms.

To What Extent Can Function Be Predicted from the Structure of the Domain

Global Structure Comparison

Since most structural domain superfamilies (>70% of superfamilies in CATH) are rather homogeneous in function [13], classifying a new domain in one of these superfamilies generally allows inheritance of function from one of the other experimentally characterised superfamily members [22, 23]. Over the last 20 years a plethora of structure comparison algorithms have been developed which attempt to handle the diverse structural changes that can occur during evolution. For very

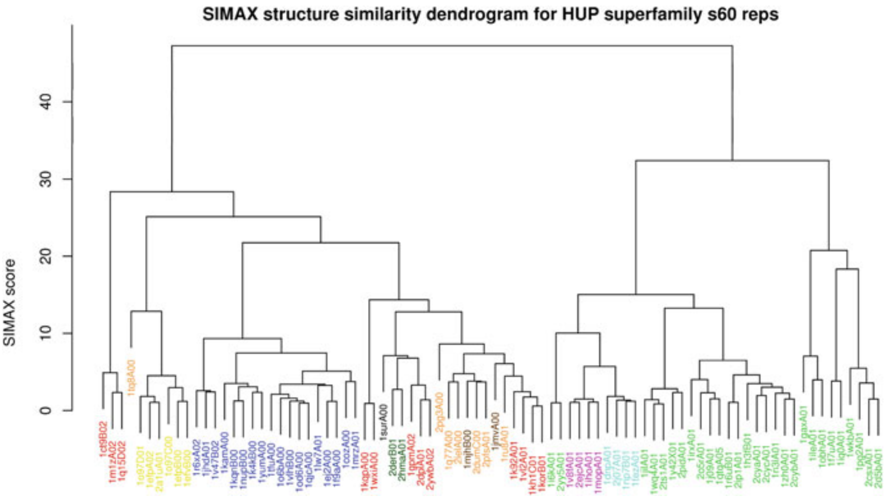

Fig. 3 Dendrogram showing non-redundant relatives of the HUP superfamily clustered by a normalised RMSD score (SIMAX). Domains that share similar functions are highlighted in the same colour

remote homologues in these superfamilies, function can often be assigned using reliable structure comparison methods (e.g. CE [24], DALI [25], CATHEDRAL [26], Structal [27], FatCat [28]; see also [2, 29] for reviews).

Whilst a number of fast structure comparison methods exist [26, 28, 30] most of which compare secondary structures between proteins and can be used to search the PDB for putative fold matches, the most accurate methods compare residue positions between proteins [2, 29]. Some of these algorithms exploit the dynamic programming algorithms or other sophisticated optimisation protocols like simulated annealing to handle residue insertions and deletions. However, whilst global structural similarity is quite a good indication of functional similarity and can be used to cluster together relatives sharing common functions within structural superfamilies (see Fig. 3), rather high thresholds on similarity are required to ensure significant conservation of function (see Fig. 4).

Assigning Functions Based on Local Structural Similarity

Various studies suggest that domains that seem unrelated as a whole may contain evolutionarily-conserved subparts [31, 32] such as their active sites [33].

As structure is more conserved across protein families than sequence [10], structure comparison methods are able to detect far more distant relationships than the most powerful profile methods. However, as discussed already, even domains in the same superfamily can exhibit large amounts of structural variation [11]. This may be due to different protein or domain interactions, or requirements to attach to distinct cellular environments, or might simply be due to random evolutionary drift.

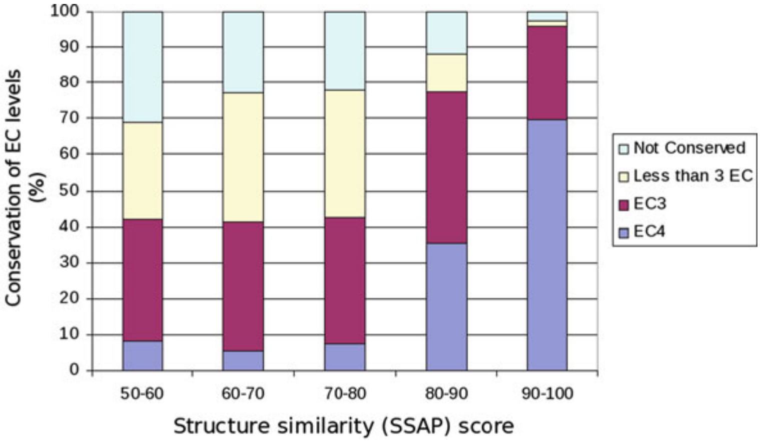

Fig. 4 Plot showing relationship between structural similarity and functional conservation (measured as conservation of EC levels). The SSAP score ranges from 0 to 100 (for identical structures)

Consequently, these structural deviations can mean that even an accurate alignment of two structures can produce a global similarity score that falls below reliable thresholds for transferring a specific function.

In a similar vein to the way PRINTS [34] and PROSITE [35] focus on smaller conserved sequence patterns, there are several approaches to identifying local structure motifs that are associated with specific functions. For example, the Catalytic Site Atlas [36] concentrates on building 3D motifs of residues that are directly involved in ligand binding or the catalytic mechanism in an enzyme. As ab initio prediction of functional residues is a complex problem in itself, the Thornton group at the European Bioinformatic Institute (EBI) have focussed on mining the primary literature to obtain the information on which to build templates. Torrance et al. [37] analysed the performance of this approach for enzymes with more than 2 catalytic residues. They were able to discriminate related proteins from random with 85% accuracy and found that it was important to focus on C-alpha/C-beta residues as their position is better conserved than side chain atoms. However, even by capturing the correct functionally active residues – for example, the catalytic triad in the serine proteases – the flexibility of active sites significantly impacts on the ability of these templates to detect mobile residues in X-ray crystal structures with different bound ligands.

Methods That Search for Patterns of Conservation Without Having Functional Groups or Motifs Defined

In contrast to exploiting information on known functional residues, the DRESPAT method [38] uses graph theory to extract recurring structural patterns across superfamilies in the SCOP database [3]. DRESPAT makes no assumptions about the location or nature of the motif positions, except by excluding hydrophobic residues. A statistical model is built to assess the significance of each recurring pattern and the authors were able to identify different metal binding sites in distantly related proteins. However, as with many methods which seek small structural motifs, distinguishing between genuine similarities and background is hampered by high false positive rates.

The PINTS methods [39] also shows promise for automatically detecting structural motifs in protein families, although is not able to annotate novel proteins with high accuracy. Again, recurring side chain patterns are identified through a pair-wise comparison of diverse members within a protein family. These motifs can then be used to scan against a novel structure.

Instead of detecting 3D templates based on their structural conservation across an enzyme family, Polacco and Babbitt [40] used a genetic algorithm (GASP) to generate a functional template from a given structure based on its ability to identify members of the same enzyme superfamily against a background of unrelated proteins in the SCOP database. An initial PSI-BLAST step builds a multiple sequence alignment for each enzyme structure that is used to create a set of conserved residues, from which a small number ($\sim$10) are selected at random to build a template. The performance of each template is then evaluated by using a geometric matching algorithm, SPASM, to score matches to the functional relatives and the SCOP library. Interestingly, the best template generally contains known functional amino acids, although there are also a few additional residues with no known functional role. This method is a promising development, although each template takes up to 18 h to generate and the performance was only evaluated for five superfamilies.

Methods That Search for Structural Differences Between Defined Functional Groups to Identify Functional Determinants

The FLORA Algorithm

A novel approach was [41], developed recently in our group to provide structural templates for assigning uncharacterised structures to functional subfamilies in the CATH classification, performs global structural comparisons between relatives within a superfamily to identify structural features that are highly conserved within a functional subfamily but less conserved across the complete superfamily.

FLORA does not exploit information on known functional residues such as catalytic residues from the Catalytic Site Atlas (CSA) to characterise functionally important positions in the protein. Functionally relevant positions are identified from structural comparisons within and between the functional subfamilies within a superfamily.

Benchmark Dataset

The method was originally benchmarked by deriving a dataset of functional subfamilies in 29 large, enzyme superfamilies. Only functionally diverse superfamilies were included, with relatives accounting for at least 3 different Enzyme Classification (EC) codes. A non-redundant set of structures were used for each superfamily, generated by clustering relatives sharing 60% or more sequence identity. This threshold was used as it has been shown to be associated with a high likelihood of functional similarity in the EC classification [41]. Subsequently, structures were clustered into functional subfamilies if they shared at least the first 3 EC numbers. A CATH superfamily was then included in the dataset only if it contained at least 3 functional subfamilies, where each subfamily contained at least 4 structures. These criteria were chosen to create a sufficiently diverse data set, which could be effectively assessed using leave-one-out benchmarking. The final dataset contained 82 functional subfamilies from 29 diverse CATH superfamilies (900 domains in total) and constitutes one of the largest datasets available for evaluating structure to function prediction algorithms. Furthermore, although these superfamilies account for <2% of the CATH superfamilies (currently 2600), they are very large comprising nearly 50% of sequences in functionally diverse CATH superfamilies.

Overview of Method

Figure 5 shows a flowchart of the FLORA method. FLORA does not rely on initial seeds of known functional residues but explores the whole structure of the domains in order to find discriminating positions. This information is then captured by generating vectors between these positions which can be compared against query structures to recognise functional homologues.

Structural comparisons within and between functional subfamilies are performed using the CATHEDRAL algorithm, another in-house method [26]. This is a relatively fast comparison method which exploits graph theory and double dynamic

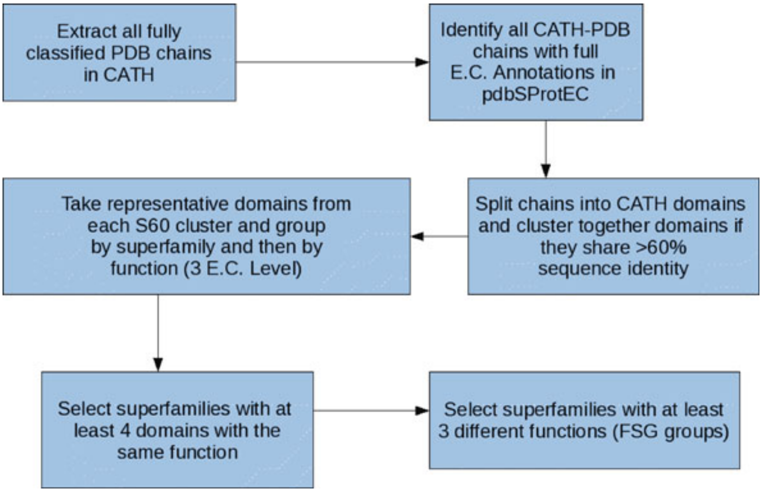
Extract all fully classified PDB chains in CATH
Identify all CATH-PDB chains with full E.C. Annotations in pdbSProtEC
Split chains into CATH domains and cluster together domains if they share >60% sequence identity
Take representative domains from each S60 cluster and group by superfamily and then by function (3 E.C. Level)
Select superfamilies with at least 4 domains with the same function
Select superfamilies with at least 3 different functions (FSG groups)

Fig. 5 Flowchart of the FLORA method

programming and had been shown to perform well against other established methods (e.g. DALI, Structal, CE) and which aligns the largest proportion of equivalent residues with respect to manually curated alignments [26].

Step 1: Identify Structurally Conserved Residues

CATHEDRAL is used to perform pair-wise structural comparisons between all representatives in the given CATH functional subfamily. Subsequently, for each domain, residues are only considered if they can be aligned against residues in at least 75% of other relatives from the subfamily (equivresidues). For each domain, vectors were calculated between the equivresidues.

Vectors were calculated between the C_β atoms of the equivalent residues (A→B) and then multiplied by a co-ordinate frame calculated from the tetrahedral geometry of the bonds of the C_α of residue A as described in [42]. As the C_α geometry of residues A and B are not identical, vectors were calculated in both the A→B and B→A direction. However, we found that taking only one of these vectors forward to the next steps in the algorithm gave the same performance as using both, but increased the speed of FLORA.

Vectors for each domain in the superfamily were then compared against equivalent vectors in all other domain representatives from the superfamily. Equivalent vectors were determined from the structural alignment of the two domains being compared. Vectors were scored using the formula given in Eq. (2) below, where the values for a and b were determined from trials. The optimal values were $a{=}b{=}2$.

$$\text{score} = \frac{a}{|v1 - v2| + b} \tag{2}$$

The next step is to identify those vectors for a given domain that are structurally more conserved between members of the same functional subfamily than compared to members of different functional subfamilies. The aim of this step is to eliminate any vectors that are conserved across the whole superfamily. These vectors are likely to be associated with the core of the domain structure which is common to all members of the superfamily. Any remaining vectors are more likely to be associated with functionally specific regions on the domain structure.

In order to identify these "functionally specific" vectors, two distributions were calculated for each vector considered. One captures the scores obtained by comparing the vector to equivalent vectors in domains in the same functional subfamily and the other, scores for comparisons involving vectors in different subfamilies. The means of these distributions were calculated and the vector was identified as functionally specific if the following condition was met:

$$\text{mean (functional subfamily distribution)} - \text{mean (superfamily distribution)} > 1$$

The set of selected vectors is reduced by jack-knifing the data set and repeating the calculation above. That is, each domain is removed in turn and a vector is only selected as specific if the inequality is always satisfied. At the end of this process, each domain is associated with a *template* set of functionally specific vectors.

Scoring Query Structures Against FLORA Template Sets for Individual Domains

In order to determine whether a query structure can be assigned to a specific functional subfamily within a CATH superfamily, the query is structurally aligned to all representatives in the superfamily, using the CATHEDRAL algorithm again, and a score calculated for each comparison.

When scoring the alignment of the query structure against a given member of functional subfamily, the algorithm only scores the similarity over the set of functionally specific vectors for the subfamily domain. Thus the algorithm is effectively calculating a local score using the correspondences determined by a global structure comparison. Each vector in the template set is scored against the equivalent vector in the query domain using the following formula:

$$\text{florascore} = \frac{\sum_{i=0}^{N} \text{score}(v1, v2)}{N} \tag{3}$$

where N = number of template vectors; $v1$ = template vector; $v2$ = equivalent vector in query domain.

Any vectors that are not aligned (i.e. gapped positions) are given a score of zero. The total similarity of the query domain against enzyme domain (the *florascore*) is simply the sum of these similarities, normalised by the total number of vectors in the template (Eq. (3)).

In order to take account of the different degrees of structural-functional diversity in different superfamilies this score is converted to a Z-score which could be applied regardless of the superfamily being considered.

Assessing the Performance of FLORA

FLORA was benchmarked using the dataset of 29 functionally diverse CATH enzyme superfamilies described above. In order to assess the performance in an unbiased manner we used a standard leave-one-out approach. That is, for a given superfamily being evaluated, one domain member is removed from the set which is then used as a training set for the algorithm. The selected test domain is then scored against FLORA templates for all superfamilies.

We compared the performance of FLORA against global structure comparison algorithms CE [24], CATHEDRAL [26] and against another publicly available structure–function prediction method, Reverse Templates [43]. Unfortunately few structure–function prediction algorithms are available but Reverse Templates is one of the leading methods. We plotted sensitivity (i.e. tp/(tp + fn)) versus

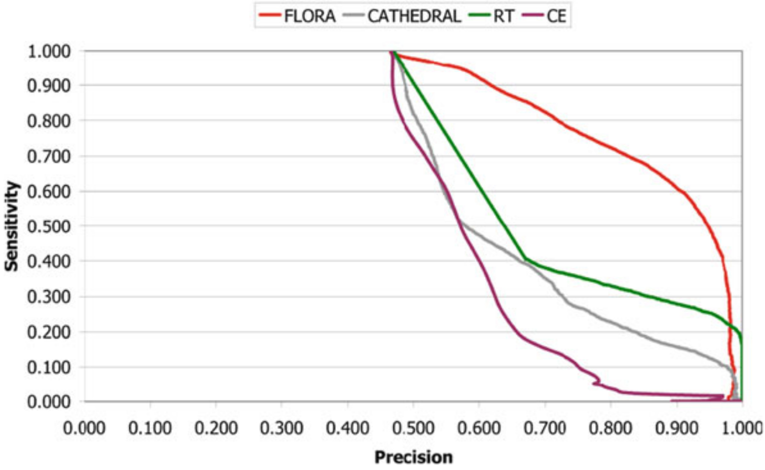

Fig. 6 Graph of sensitivity versus precision to show the performance of CE, CATHEDRAL, RT and FLORA for the prediction of enzyme family

precision (tp/(tp + fp)) and assessed the performance on individual superfamilies by calculating AUC value (area under ROC curve).

It can be seen from Fig. 6 that both global structure comparison methods, CE and CATHEDRAL, are poor at recognising the correct functional subfamily to which a query domain should be assigned. CATHEDRAL outperforms CE, most likely because it is able to align more equivalent positions, as identified in previous studies [26]. However, neither method was specifically designed for recognising functional homologues.

Even at high precision (>95%) FLORA significantly outperforms CE, CATHEDRAL and Reverse Templates. At 90% precision it captures twice the number of functional homologues than Reverse Templates. The sensitivity of the algorithm derives from the fact that although FLORA uses an alignment derived by CATHEDRAL, it only scores positions deemed to be functionally specific (i.e. in the FLORA template set). By exploiting multiple structures from a functional subfamily it can more easily identify these specific positions.

We have also examined the effect on the FLORA performance of using whole protein chains rather than protein domains. There was negligible impact on performance which suggests that there is enough signal in the domain structure to recognise the specific function of the protein containing the domain. This is encouraging if we wish to exploit FLORA as a general function prediction method since the majority of proteins differ between organisms [44] whilst the domain components within them are related and can therefore, from these results, be used to suggest functions for the whole proteins.

Visualisation of Functionally Specific Positions Detected by FLORA

The power of FLORA lies in its ability to identify residues beyond the common structural core of the domain subfamily. Our previous analyses observed that nearly 70% of residue positions identified by FLORA were located close to functional sites [41]. Other FLORA positions were found to be close to interface surfaces involved in protein interactions. To manually assess the ability of FLORA to recognise functionally relevant sites in the domain structures, FLORA positions were mapped onto representative structures from the HUP domain superfamily, which is one of the largest and most structurally and functionally diverse superfamilies in CATH, comprising more than 9 different functional subfamilies.

Domains in this superfamily adopt a Rossmann-like fold with a central parallel β-sheet surrounded on both sides by α-helices. The main active site is always located in the C-terminal half of the central β-sheet and is generally involved in nucleotide-binding.

Figure 7 illustrates residue locations identified by FLORA templates for a subfamily from the HUP domain superfamily. A representative structure for this functional subfamily was chosen as the structure with the highest cumulative structural similarity score to all other non-redundant members (at 100% sequence identity) of the subfamily. Residue positions are highlighted if at least 30% of

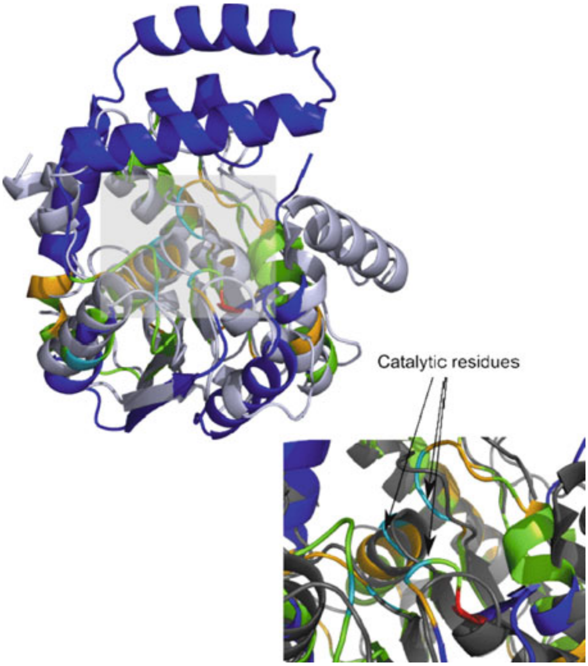

Fig. 7 Superposition of PSI structure 2pbl (*dark grey*) with 1tqh (superfamily 3.40.50.1820, EC 3.1.1.-). Flora residues are coloured *green*, or *gold* if they are conserved across the whole superfamily, and catalytic residues are shown in *light blue*. It can be seen that there is reasonable agreement in the region of the active site

FLORA templates for this subfamily include these positions. Any positions conserved across a majority of the superfamily (i.e. 75% or more of the relatives) are coloured gold.

Incorporating Sequence Based Protocols with FLORA to Identify Functionally Specific Residues

We explored the effects of including sequence matching within the FLORA algorithm. That is including a contribution to the score reflecting identical or similar residues between the query and the template structure. However, this tended to degrade the performance and was not included in the final version of the algorithm.

Instead we have developed a separate sequence based protocol (GeMMA [45]) for identifying residue positions likely to be associated with the function. This allowed us to annotate structural domains within each functional subfamily with residue positions identified as functionally specific from both structural data (FLORA) and sequence data (GeMMA).

More importantly, GeMMA allows to identify functional subgroups amongst all the sequences assigned to a superfamily, even those without known structures. Since the number of sequence relatives can be up to 100-fold greater than the number of structures for some superfamilies, this gives a more accurate representation of functional divergence across the superfamily. Functional subfamilies identified by GeMMA can be used as sets for training the FLORA algorithm, provided they contain three or more non-redundant structures and can therefore be used to identify positions associated with function which are structurally conserved.

GeMMA exploits information from all the predicted sequence domains assigned to a particular CATH structural superfamily and contained within our Gene3D resource (http://gene3d.biochem.ucl.ac.uk/Gene3D). For example in the HUP superfamily mentioned above, there are 85 non-redundant structures (at 60% sequence identity) and 9484 non-redundant sequences stored within CATH-Gene3D. Gene3D contains all the predicted domain sequences for CATH superfamilies identified using HHM models built from the sequences of non-redundant structural domains in CATH [23].

GeMMA initially compares (using BLAST) all the sequences against each other and then progressively merges similar sequences into functional subgroups or subfamilies. This is initially done on the basis of pairwise sequence similarity but as the clusters grow and there are enough sequences to make a sequence profile, profile–profile comparisons are performed between clusters. Clusters are merged provided the E-value returned from the comparison is below a threshold obtained by benchmarking with superfamilies for which there are extensive experimental functional characterisations [45].

Since profile–profile comparisons can be very computationally expensive, we have developed a strategy for reducing the number of comparisons that need to be performed and for running a modified version of the protocol on multiple compute nodes. Alternative sequence based strategies for identifying functional subfamilies within superfamilies tend to exploit tree based approaches that rely on a multiple sequence alignment of all the sequences to build the tree. However, the most functionally diverse superfamilies in CATH, which account for more than half the sequences in the genomes, contain more than 10,000 sequences. This number of sequences is beyond the scope of most multiple sequence alignment methods. Even when non-redundant datasets are generated at 60% sequence identity to ensure functional coherence, there are still large numbers of sequences in these very large superfamilies (i.e. > 5000). Therefore, the iterative clustering protocol used by GeMMA (also described as agglomerative clustering), is the most tractable approach for these very large and functionally diverse superfamilies.

FLORA templates can be derived for GeMMA functional subfamilies which contain 3 or more non-redundant structures (at 30% sequence identity). As mentioned above FLORA analyses can exploit the structural data in these subfamilies to identify structurally conserved positions associated with functional sites (e.g. active sites and protein–protein interaction surfaces). GeMMA identifies >100 functional subfamilies in the diverse HUP superfamily. Figure 8 shows a representative from one of these subfamilies with residue positions highlighted according to whether they are identified as sequence conserved by GeMMA or structurally conserved by FLORA or both sequence and structure conserved. Mapping these conserved residues onto the structure is clearly useful in suggesting the location of functional sites on the protein domain. In the future CATH-Gene3D will be providing information on GeMMA functional subfamilies for selected CATH domain superfamilies being targeted for structural genomics by the protein structure initiative (PSI) in the United States.

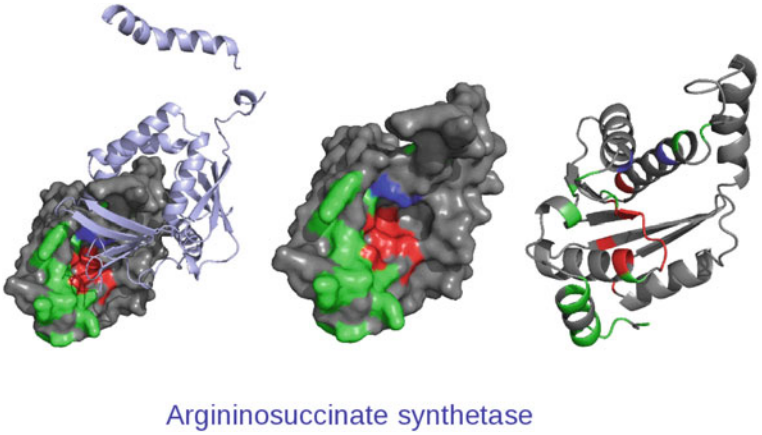

Fig. 8 Representative structure from one of the HUP protein subfamilies. Residues that are conserved by structure are coloured *green*, those conserved by sequence are coloured *blue* and those conserved by both sequence and structure are coloured *red*

References

1. Watson, J.D., Laskowski, R.A., et al. Predicting protein function from sequence and structural data. Curr. Opin. Struct. Biol. **15**(3): 275–284 (2005).
2. Lee, D., Redfern, O., et al. Predicting protein function from sequence and structure. Nat. Rev. Mol. Cell Biol. **8**(12): 995–1005 (2007).
3. Murzin, A.G., Brenner, S.E., et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**(4): 536–540 (1995).
4. Orengo, C.A., Michie, A.D., et al. CATH--a hierarchic classification of protein domain structures. Structure **5**(8): 1093–1108 (1997).
5. Sillitoe, I., Dibley, M., et al. Assessing strategies for improved superfamily recognition. Protein Sci. **14**(7): 1800–1810 (2005).
6. Cuff, A., Redfern, O., et al. Classification of protein structures. *Computational structural biology, methods and applications*. Schwede, T., Peitsch, M.C. (eds.). Singapore: World Scientific, pp. 153–188 (2008).
7. Martin, A.C., Orengo, C.A., et al. Protein folds and function. Structure **15**;6(7): 875–884 (Jul 1998).
8. Russell, R.B., Sasieni, P.D., et al. Supersites within superfolds. Binding site similarity in the absence of homology. J. Mol. Biol. **282**(4): 903–918 (1998).
9. Cuff, A., Redfern, O.C., et al. The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. Structure **17**(8): 1051–1062 (2009).
10. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. **5**(4): 823–826 (1986).
11. Reeves, G.A., Dallman, T.J., et al. Structural diversity of domain superfamilies in the CATH database. J. Mol. Biol. **360**(3): 725–741 (2006).
12. Todd, A.E., Orengo, C.A., et al. Plasticity of enzyme active sites. Trends Biochem. Sci. **27**(8): 419–426 (2002).
13. Todd, A.E., Orengo, C.A., et al. Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. **307**(4): 1113–1143 (2001).
14. Gerlt, J.A. A protein structure (or function?) initiative. Structure **15**(11): 1353–1356 (2007).
15. Chiang, R.A., Sali, A., et al. Evolutionarily conserved substrate substructures for automated annotation of enzyme superfamilies. PLoS Comput. Biol. **4**(8): e1000142 (2008).
16. Rison, S.C., Thornton, J.M. Pathway evolution, structurally speaking. Curr. Opin. Struct. Biol. **12**(3): 374–382 (2002).
17. Dessailly, B.H., Redfern, O.C., Cuff, A.L., Orengo, C.A. Detailed analysis of function divergence in a large and diverse domain superfamily: toward a refined protocol of function classification. Structure. **18**(11): 1522–1535 (2010 Nov 10).
18. Andreeva, A., Murzin, A.G. Evolution of protein fold in the presence of functional constraints. Curr. Opin. Struct. Biol. **16**(3): 399–408 (2006).
19. Taylor, W.R. Evolutionary transitions in protein fold space. Curr. Opin. Struct. Biol. **17**(3): 354–361 (2007).
20. Grishin, N.V. Fold change in evolution of protein structures. J. Struct. Biol. **134**(2–3): 167–185 (2001).
21. Pegg, S.C., Brown, S.D., et al. Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. Biochemistry **45**(8): 2545–2555 (2006).
22. Wilson, D., Madera, M., et al. The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res. **35**(Database issue): D308–D313 (2007).
23. Yeats, C., Lees, J., et al. Gene3D: comprehensive structural and functional annotation of genomes. Nucleic Acids Res. **36**(Database issue): D414–D418 (2008).
24. Shindyalov, I.N., Bourne, P.E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. **11**(9): 739–747 (1998).
25. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. **233**(1): 123–138 (1993).

26. Redfern, O.C., Harrison, A., et al. CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. PLoS Comput. Biol. **3**(11): e232 (2007).
27. Kolodny, R., Koehl, P., et al. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J. Mol. Biol. **346**(4): 1173–1188 (2005).
28. Ye, Y., Godzik, A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics **19**(Suppl 2): ii246–ii255 (2003).
29. Redfern, O.C., Dessailly, B., et al. Exploring the structure and function paradigm. Curr. Opin. Struct. Biol. **18**(3): 394–402 (2008).
30. Zhang, Y., Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. **33**(7): 2302–2309 (2005).
31. Soding, J., Lupas, A.N. More than the sum of their parts: on the evolution of proteins from peptides. Bioessays **25**(9): 837–846 (2003).
32. Manikandan, K., Pal, D., et al. Functionally important segments in proteins dissected using gene ontology and geometric clustering of peptide fragments. Genome Biol. **9**(3): R52 (2008).
33. Xie, L., Bourne, P.E. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. Proc. Natl. Acad. Sci. USA **105**(14): 5441–5446 (2008).
34. Attwood, T.K., Bradley, P., et al. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res. **31**(1): 400–402 (2003).
35. Hulo, N., Bairoch, A., et al. The PROSITE database. Nucleic Acids Res. **34**(Database issue): D227–D230 (2006).
36. Porter, C.T., Bartlett, G.J., et al. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. **32**(Database issue): D129–D133 (2004).
37. Torrance, J.W., Bartlett, G.J., et al. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J. Mol. Biol. **347**(3): 565–581 (2005).
38. Wangikar, P.P., Tendulkar, A.V., et al. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. J. Mol. Biol. **326**(3): 955–978 (2003).
39. Stark, A., Russell, R.B. Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. Nucleic Acids Res. **31**(13): 3341–3344 (2003).
40. Polacco, B.J., Babbitt, P.C. Automated discovery of 3D motifs for protein function annotation. Bioinformatics **22**(6): 723–730 (2006).
41. Redfern, O.C., Dessailly, B.H., et al. FLORA: a novel method to predict protein function from structure in diverse superfamilies. PLoS Comput. Biol. **5**(8): e1000485 (2009).
42. Aravind, L., Anantharaman, V., et al. Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. Proteins **48**(1): 1–14 (2002).
43. Laskowski, R.A., Watson, J.D., et al. Protein function prediction using local 3D templates. J. Mol. Biol. **351**(3): 614–626 (2005).
44. Grant, A., Lee, D., et al. Progress towards mapping the universe of protein folds. Genome Biol. **5**(5): 107 (2004).
45. Lee, D.A., Rentzsch, R., et al. GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. Nucleic Acids Res. **38**(3): 720–737.