Assessing functional novelty of PSI structures via structure-function analysis of large and diverse superfamilies

Benoît H Dessailly*, Oliver C Redfern, Christine A Orengo
Dept. of Biochemistry, University College London, Gower Street, London WC1E 6BT, UK
*To whom correspondence should be addressed: benoit@biochem.ucl.ac.uk

1. INTRODUCTION

The structural genomics initiatives have had as one of their aims to improve our understanding of protein function by providing representative structures for many structurally uncharacterised protein families. As suggested by the recent assessment of the Protein Structure Initiative (Structural Genomics Initiative, funded by the NIH), doubts have arisen as to whether Structural Genomics as initially planned were really beneficial to our understanding of biological issues, and in particular of protein function.

A few protein domain superfamilies have been shown to account for unexpectedly large numbers of proteins encoded in fully sequenced genomes [1]. These large superfamilies are generally very diverse, spanning a wide range of functions, both in terms of molecular activities and biological processes. Some of these superfamilies, such as the Rossmann-fold P-loop nucleotide hydrolases or the TIM-barrel glycosidases, have been the subject of extensive structural studies which in turn have shed light on how evolution of the sequence and structure properties produce functional diversity amongst homologues [2]. Recently, the Structure-Function Linkage Database (SFLD) has been setup with the aim of helping the study of structure-function correlations in such superfamilies [3]. Since the evolutionary success of these large superfamilies suggests biological importance, several Structural Genomics Centers have focused on providing full structural coverage for representatives of all sequence families in these superfamilies.

In this work we evaluate structure/function diversity in a set of these large superfamilies and attempt to assess the quality and quantity of biological information gained from Structural Genomics.

2. RESULTS

The CATH database [4] was filtered to select superfamilies with *(a)* more than 20 sequence-diverse relatives (sequence identity lower than 35%), *(b)* at least 3 different EC numbers (down to the 3rd level), and *(c)* structures solved as part of the Protein Structure Initiative and in particular the Midwest Center for Structural Genomics (MCSG) with which we collaborate directly. We then manually checked the set of superfamilies thus selected, to further restrict our analysis to 3 superfamilies with a wide range of molecular functions: the HUP-domain superfamily (CATH 3.40.50.620) that was chosen because it is very ancient and presents a particularly striking diversity of entirely unrelated functions; the HAD superfamily (CATH 3.40.50.1000) that was chosen as one of the superfamilies considered in the above-mentioned SFLD; and the PBP-like superfamily (CATH 3.40.190.10) that we selected to study function diversity in a superfamily that includes many non-enzymatic functions.

We first define functional sub-groups of proteins in each superfamily (*i.e.* proteins that share a common molecular function different from that of their other homologues) by applying a protocol that combines manual curation of the relevant literature, annotations from diverse databases (GO, EC, KEGG, FUNCAT, SFLD), and structural information (*i.e.* multiple domain architectures, active sites, ligands, secondary structure motifs, global structure comparisons). Using all these data, we illustrate the considerable functional diversity of the superfamilies under study, and show that the different functional sub-groups also tend to differ significantly from one another in terms of their 3D structure.

Next, PSI (MCSG) structures in each superfamily are systematically compared with members from each functional sub-groups so as to determine whether these structures represent functions without previously existing structural representatives or not. In total, 47 (8) PSI (MCSG) structures were considered for these 3 superfamilies. Cases where PSI (MCSG) structures shed light on apparently 'novel' functions are discussed and functional predictions for these proteins are suggested on the basis of comparisons with other functionally characterised members of the superfamily.

3. CONCLUSION

Preliminary steps of an automated protocol to group homologous proteins in functional sub-groups are presented. This protocol is applied to three large and diverse well-characterised superfamilies, and the correlated evolution of function and structure therein is analysed. Finally, the extent to which Structural Genomics iniatitives, in particular those from the MCSG center (PSI), have helped understanding functional diversity and evolution in large superfamilies is discussed.

ACKNOWLEDGMENTS

This work was supported by the NIH and the BioSapiens NOE.

REFERENCES

1. Marsden RL et al. (2006). Exploiting protein structure data to explore the evolution of protein function and biological complexity. *Philos Trans R Soc Lond B Biol Sci* **361**, 425-440.
2. Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA (2006). Structural diversity of domain superfamilies in the CATH database. *J Mol Biol* **360**, 725-741.
3. Pegg SC et al. (2006). Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry* **45**, 2545-2555.
4. Greene LH et al. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucl Acids Res* **35**, D291-297.