MS12-O2

CATH funfams for domain boundary recogntion and template selection

Christine Orengo[1], Ian Sillitoe[1], Sayoni Das[1], Jon Lees[1]

1. University College London, London, United Kingdom

email: c.orengo@ucl.ac.uk

The CATH domain structure database (www.cathdb.info) classifies domains of known structure from the PDB and predicted domains from sequences in UniProt. The major focus is on identifying evolutionary relatives, which are classified into superfamilies. Superfamilies are subsequently annotated using a hierarchical model, according to their fold group, architecture and protein class (Dawson et al. 2017). Sequence profiles are derived from representatives in each superfamily and Hidden Markov Models (HMMs) generated using HMMer3 (Lees et al. 2017). CATH currently comprises, ~450,000 domain structures and ~90 million domain sequences, classified into ~6000 evolutionary superfamilies. The latest version (CATH 4.2) captures more than 90% of domains identified in the PDB. However more than 50% of the domains are classified in fewer than 200 'mega' superfamilies, which have diverged considerably in the structures and functions of the relatives – although the structural core remains very highly conserved. Detailed analyses have revealed considerable structural diversity outside this core, often in functional regions of the protein eg around active site pockets, where the changes may be associated with modifications in function, particularly for paralogous proteins.

In order to explore the divergence of structure and function within superfamilies, we have developed a sub-classification protocol which groups relatives having highly similar structures and functions into functional families (FunFams). This is achieved using an agglomerative clustering protocol, with functional groupings recognised from distinct differences in patterns of sequence conservation, likely to be associated with functional determinants (Das et al. 2015). CATH-FunFams tend to be highly structurally coherent – relatives typically superpose within 2A RMSD. We have developed protocols that use the FunFams to provide improved domain boundary assignment for query sequences and to select suitable templates for homology modelling or structure determination. FunFams were used by the PSI MCSG Structural Genomics Consortium to improve construct generation. FunFams were also used to identify targets likely to have novel functions, for structure determination. They can also be used to facilitate the interpretation of genetic variants linked to disease.

References:

[1] CATH: an expanded resource to predict protein function through structure and sequence. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I. Nucleic Acids Res. 2017 Jan 4;45(D1):D289-D295

[2] An overview of comparative modelling and resources dedicated to large-scale modelling of genome sequences. Lam SD, Das S, Sillitoe I, Orengo C. Acta Crystallogr D Struct Biol. 2017 Aug 1;73(Pt 8):628-640.

[3] Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Das S, Lee D, Sillitoe I, Dawson NL, Lees JG, Orengo CA. Bioinformatics. 2016 Sep 15;32(18):2889.

Keywords: structural bioinformatics, domain boundary identification, functional family classification