

Automated Protein Structure Prediction Using Templates from the CATH Protein Family Database

Adrian Shepherd^{1,*}, Christine Orengo¹, Nigel Martin², Roger Johnson²

¹ Department of Biochemistry & Molecular Biology, University College London, Gower Street, London WC1E 6BT

² Department of Computer Science, Birkbeck College, Malet Street, London WC1E 7HX

* Correspondence email: a.shepherd@biochem.ucl.ac.uk

Recent developments in the CATH database include the generation of multiple structural alignments for each CATH homologous family with two or more non-identical structures (~400 families), using the program CORA (Orengo, 1999). This enables equivalent residues to be identified for each relative within the family so that structural and functional characteristics can be compared and consensus properties identified. To improve functional annotation within each family, a Dictionary of Homologous Superfamilies has also been set up. This includes any relevant functional information which can be electronically extracted from public databases (e.g. EC Classification numbers and functional keywords from the SWISS-PROT database). The DHS also contains information about conserved protein-ligand interactions, some of which correspond to consensus sequence motifs identified by PROSITE patterns. The CATH schema design reflects the need to model this multiple alignment data, equivalent residue positions and functional relationships for each protein family.

Protocols have also been established for identifying sequence relatives for the CATH homologous superfamilies in the sequence databases. With more than 27 complete genomes, there are now more than 500,000 sequences in the translated Genbank (Genpept) database. Using a 1D-profile based approach PSI-BLAST (Altschul *et al.* 1997), we have developed a reliable method for identifying sequence relatives to all the non-identical domain structures in CATH and for integrating these sequences into the database. This has resulted in nearly 200,000 domain sequences being added to the database. This additional sequence data is essential for expanding the population of the families and thereby allowing us to develop highly sensitive search algorithms for recognising more distant relationships in the sequence databases and to partial sequences in the EST data. A number of techniques are currently being explored for doing this (e.g. Hidden Markov Models).

A conceptual model has been developed in UML for the data and metadata in the CATH database and related data sources. Evaluation of both object (O2) and object-relational (Oracle8) DBMSs has taken place through the implementation of prototype databases, and as a result Oracle8i has been adopted for on-going schema and database development. The schema design supports future reclassification within CATH families, with the preservation of previous versions. Work is now under way on the population of the database from existing data sources and the development of database validation routines to maintain the integrity of related data. Once this has been completed, the next step is the investigation of database searching in the presence of uncertain information, such as the position of domain boundaries, which the schema is designed to capture.

References:

Orengo CA (1999). CORA - Topological fingerprints for protein structural families. *Protein Science*, 8, 699-715

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389-3402