

CATHEDRAL: a novel method for assigning domain boundaries and folds to multi-domain proteins

Oliver Redfern

We present CATHEDRAL, an iterative protocol for determining the location of previously observed protein folds in novel multi-domain protein structures. CATHEDRAL builds on the features of a fast secondary-structure-based method to locate known folds (GT) within a multi-domain context and a residue-based, double-dynamic programming (DDP) algorithm, which is employed to align members of the target fold groups against the query protein to identify the closest relative and assign domain boundaries. To increase the fidelity of the assignments, Support Vector Machines are utilised to provide an optimal scoring scheme. Once a domain is verified, it is excised and the search protocol is repeated in an iterative fashion until all recognisable domains have been identified.

We have performed an initial benchmark of CATHEDRAL against other publicly available structure comparison methods using a consensus data set of domains derived from the CATH and SCOP domain classifications. This shows superior performance in fold recognition and alignment accuracy when compared to many equivalent methods.

If a novel multi-domain structure contains a known fold, CATHEDRAL will locate it in 90% of cases, with <1% false positives. For nearly 80% of assigned domains in a manually validated test set, the boundaries were correctly delineated within a tolerance of ± 10 residues. For the remaining cases, previously classified relatives were very remote from the query chain and embellishment to the core of the fold caused significant differences in domain sizes; hence, manual refinement of the boundaries was necessary. Since, on average, 50% of newly determined protein structures contain more than one domain unit and typically 90% or more of these are already classified in CATH, CATHEDRAL will considerably facilitate the automation of protein classification.