Fast Assignment of Protein Structures to Sequences using the Intermediate Sequence Library PDB-ISL

Sarah A. Teichmann* and Cyrus Chothia
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.
George M. Church and Jong Park+
Department of Genetics, Harvard Medical School, Warren Alpert Bldg.,
200 Longwood Ave., Boston, MA, 02115, USA.

*sat@mrc-lmb.cam.ac.uk
+Present address: EBI, Genome Campus, Hinxton, CB10 1SD, UK.

Two homologous sequences, which have diverged beyond the point where their homology can be recognised by a simple direct comparison, can be related through one or more other sequences that are suitably intermediate between the two. A library containing sequences homologous to proteins of known structure (PDB-ISL) is described. Sequences of proteins of unknown structure can be matched to distantly related proteins of known structure by using pairwise sequence comparison methods to find homologues in PDB-ISL. The sequences in the library were collected from a large sequence database using the program PSI-BLAST and the sequences of the domains of proteins of known structure as the query sequences. Searches of PDB-ISL were calibrated. For a given error rate the number of correct matches found is the same as that found with PSI-BLAST on a large sequence database. It therefore appears that a sequence library such as PDB-ISL contains as much information for structural assignment as a set of profiles with the same sequences. The advantage of this library is that , with a pairwise sequence comparison method such as FASTA or GAP-BLAST, it can be searched easily and, in most cases, much more quickly. This makes it particularly useful for large scale searches such as those required in structural assignments to genome sequences.