

## TEST SET OF STRUCTURAL ALIGNMENTS.

The set of 22 structural alignments used for the tests is derived from the HOMSTRAD database (Mizuguchi, Deane et al. 1998; Stebbings and Mizuguchi 2004). Families were selected based on the following criteria :

- (i) at least 5 experimentally solved structures per family,
- (ii) the average sequence identity in the family should be lower than 25 %.

The HOMSTRAD names of the 22 families of the test set are: ABC transporter (average sequence identity of 25%), acetyltransferase family (24%), alpha amylase, C-terminal domain (21%), alpha amylase, catalytic domain (23%), anticodon binding domain (22%), cytochrome p450 (21%), DEATH domain (20%), fibronectin type III domain (16%), glycosyl hydrolase family 5 (18%), haloperoxidase (25%), histidine kinase, DNA gyrase B and HSP90-like ATPase (25%), integrin I-domain (22%), lipo-calin family (20%), metallo-beta-lactamase superfamily (21%), PH domain (17%), proteasome A-type and B-type (22%), reductases (22%), rhodanese-like domain (23%), RNA recognition motif (23%), short-chain dehydrogenases/reductases (23%), thioredoxin (20%), TPR domain (17%).

In each family, 5 sequences were selected randomly. For each of these sequences, the HMM describing the family was generated with the structural alignment of all members of the family except the selected sequence and all sequences sharing more than 40% sequence identity with it (to avoid any favorable bias). Picking exactly the same number of sequences in each family allows not to over-represent a family.

## REFERENCES

- Mizuguchi, K., C. M. Deane, et al. (1998). "HOMSTRAD: a database of protein structure alignments for homologous families." *Protein Sci* 7(11): 2469-71.
- Stebbins, L. A. and K. Mizuguchi (2004). "HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database." *Nucleic Acids Res* 32(Database issue): D203-7.