53. Amino-Acid Substitutions In Membrane Proteins: Applications To Homology Recognition And Comparative Modelling

Younes MOKRAB and Kenji MIZUGUCHI

Department of Biochemistry, University of Cambridge, United Kingdom.

A comprehensive set of homologous families of highly-curated structure-based alignments of high-resolution structures of -helical transmembrane (TM) protein domains has been constructed. The aim is to derive environment-specific substitutions tables which are used to increase the accuracy of alignments involving TM proteins and improve the performance in TM sequence-structure homology recognition.

Structural genomic initiatives are expected to make 3D experimental data in the near future available for most proteins in nature. Yet, the class of transmembrane (TM) proteins represents a major obstacle to this goal. This is because their physicochemical properties make them extremely difficult to crystallize for X-ray crystallography studies, and hardly tractable for NMR spectroscopy experiments. Consequently, computational methods for predicting 3D structures are highly valuable. Comparative/homology modelling remains the most effective approach to protein structure prediction (Williams et al., 2001). This is because it takes advantage from already available experimental structural templates to build 3D models for a related protein of interest. Therefore the tools that search for structural templates need to be highly accurate. Many algorithms have been developed to increase the sensitivity and specificity of homology recognition for globular proteins, many of which exploit evolutionary and structural information (Mizuguchi, 2004). However, they may not be generally applicable to TM proteins which have different structural features, amino acid composition and substitution rates. Thus, TM-specific algorithms are much needed. Our aim is to develop a sequence-structure homology recognition method that can use environment-specific substitution tables and structure-dependent gap penalties (Shi et al., 2001) to (1) increase the accuracy of alignments involving TM protein sequences and structures and (2) improve the specificity and sensitivity of homology searches for TM proteins.

First, we have generated a highly-curated set of structure-based alignments of TM protein structures from which environment-specific amino acid substitutions are derived. We started by making an exhaustive search for all available high resolution TM structures. These are either alpha-helical and beta barrel folds. We did not include structures from the latter fold since these are very small in number and most likely have different rules of folding. A large number of TM proteins are complexes of multiple domains, with only certain regions spanning the membrane. In order to distinguish those from other globular regions, we developed a geometry optimisation method, PDB2TMD, which searches for the most probable location of the membrane that spans any given TM protein structure.
Next, we used domain definitions and structural-evolutionary classification of protein structures from two databases; SCOP (Andreeva et al., 2004) and HOMSTRAD (Mizuguchi et al., 1998; Stebbings and Mizuguchi, 2004) . We gathered 795 TM domains derived from 226 structures and grouped them into 65 homologous families. To remove redundant structures, we developed MKNEWFAM2. For a given family, this automatic procedure clusters sequences at 90% threshold and select representatives by favouring domains closest to native and having the highest resolution. The number of domains remaining after this filtering process is 129 making 26 multimember families and 39 single-membered families. Structure-based alignments were then generated for each family using COMPARER (Sali and

Blundell, 1990) and the residues structural environment in these alignments were annotated using JOY (Mizuguchi et al., 1998) .
In order to account for the biased distribution of structures among the families, we enriched each family with sequence information. PSI-BLAST (Chen and Rost, 2002) was used to search UNIREF100 protein sequence database (Bairoch et al., 2005) for close homologues (e value cut-off 10-6), and only sequences predicted with TMHMM (Krogh et al., 2001) to contain all expected TM helices were retained. The members of an average family share 20-30% PID.

The patterns of the amino acid substitutions derived from the constructed dataset has lead to a number of findings which are of particular relevance to TM helix packing: (1) lipid-tail-accessible TM residues tend to be more hydrophobic, less conserved and contain different residue types compared to buried residues; (2) charged residues are not always buried and, when accessible to membrane lipid tails, they often interact with phospholipid head-groups or with other residue types and few pair with another charge and (3) residues that are lipid-accessible or located at the interface between different TM chains are more variable than those buried in the cores of individual chains. This suggests that helix–helix interactions within the same chain and those at the interface between different chains may arise differently.

Substitutions tables which take into account residue environments are being calculated and incorporated as scoring matrices for a homology search program we had previously developed, FUGUE (Shi et al., 2001). Benchmarking is carried out in order to examine the quality of alignments and the extent of extra sensitivity and specificity this may offer to homology searches for TM proteins.

References

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32 Database issue, D226-229.
Bairoch, A., Apweiler, R., Wu, CH., Barker, WC., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, MJ., Natale, DA., O'Donovan, C., Redaschi, N., Yeh, LS. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res. 33, D154-159.
Chen, C.P. and Rost, B. (2002) Long membrane helices and short loops predicted less accurately. Protein Sci, 11, 2766-2773.
Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol, 305, 567-580
Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. and Overington, J.P. (1998) JOY: protein sequence-structure representation and analysis. Bioinformatics, 14, 617-623.
Mizuguchi, K. (2004) Fold recognition for drug discovery. Drug Discovery Today, 3, 18-23.
Sali, A. and Blundell, T.L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J Mol Biol, 212, 403-428.
Shi, J., Blundell, T. L., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310, 243-257.
Stebbings, L. A., and Mizuguchi, K. (2004). HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. Nucleic Acids Res 32 Database issue, D203-207.

Williams, M.G., Shirai, H., Shi, J., Nagendra, H.G., Mueller, J., Mizuguchi, K., Miguel, R.N., Lovell, S.C., Innis, C.A., Deane, C.M., Chen, L., Campillo, N., Burke, D.F., Blundell, T.L. and de Bakker, P.I. (2001) Sequence-structure homology recognition by iterative alignment refinement and comparative modeling. Proteins, Suppl 5, 92-97.