Available online at www.sciencedirect.com
SCIENCE DIRECT
Computational Biology and Chemistry
Computational Biology and Chemistry 30 (2006) 120–126
www.elsevier.com/locate/compbiolchem

Homology modeling of membrane proteins: A critical assessment

Ch. Surendhar Reddy, K. Vijayasarathy, E. Srinivas, G. Madhavi Sastry *, G. Narahari Sastry *

Molecular Modelling Group, Organic Chemical Sciences, Indian Institute of Chemical Technology, Tarnaka, Hyderabad 500007, India

Received 27 August 2005; received in revised form 10 November 2005; accepted 14 December 2005

Abstract

Evaluation and validation of homology modeling protocols are indispensable for membrane proteins as experimental determination of their three-dimensional structure is an arduous task. The prediction ability of Modeller, MOE, InsightII-Homology and Swiss-PdbViewer (SPV) with different sequence alignments CLUSTALW, BLAST and 3D-JIGSAW have been assessed. The sequence identity of the target and template was chosen to be in the range of 25–35%. Validation protocols to assess the structure, fold and stereochemical quality, are employed by comparing with experimental structures. Two different ranking schemes are suggested to evaluate the performance of each methodology based on the validation scores. While unambiguous preference for any given procedure did not surface, statistically Modeller and the sequence alignment technique, 3D-JIGSAW, gave best results amongst the chosen protocols. The present study helps in selecting the right protocols when modeling membrane proteins, which form a major class of drug targets.

Keywords: Membrane proteins; 3-Dimensional structure; Homology modeling; Low sequence identity; Stereochemical quality; Structural similarity; Fold

1. Introduction

Homology modeling procedures are indispensable tools for conducting research involving structure based drug design when the experimental 3D-structure of the receptor is not available (Sanchez and Sali, 1997). This has resulted in large-scale critical assessments of the performance of various automated structure prediction servers (Bujnicki et al., 2001). Although several procedures exist in the public domain, the choice of the right method is not unambiguous (Wallner and Elofsson, 2005). Therefore, it is extremely important to persistently evaluate and validate these methods especially when applied to membrane proteins due to the difficulty in obtaining their experimental 3D-structures. This is amply evident from a recent analysis of the protein databank (PDB) (Berman et al., 2000), obtained from the PDBbeta site, which contain only 1526 membrane protein structures (about 2% of total) are available as of now (July 2005). However, many of these structures belong to the water soluble domains (Tusnády et al., 2005). The recent increase in the number of crystal structures of membrane proteins has enabled an analysis of the hydrophilic exposed surfaces and the interiors of such molecules on the basis of structure, rather than sequence alone. In spite of the increase in the number of crystal structures many structures having low resolution and high *R* factors exist in the PDB. More than 46% and 10% of the membrane protein structures have resolution ranging between 2–3 and 3–4 Å, respectively, which collectively constitute more than 56%. Therefore, homology modeling techniques are extremely important in obtaining the 3D-structure as well as in refining the existing low accuracy experimental structures (Sanchez et al., 2000). However, various protocols including sequence alignment algorithms, structure building protocols, loop building algorithms, etc., determine the accuracy of the models and choosing the right protocol is imminent to obtain good quality models (Mosimann et al., 1995).

In this work, we would like to assess various procedures when applied to model membrane proteins. We have attempted to assess the methods and servers, which are not restricted to modeling any specific class of proteins. The chosen methods and sequence alignment techniques have varied *modus operandi* of 3D model building adopted by them. The methods that do not produce any breaks in the 3D models built by them and modeled most proteins are chosen for comparison. Thus, we have considered the following methods implemented by: Modeller (Sali and Blundell, 1993), MOE (Kelly, 1999), homology module of InsightII (IH) (Dayringer et al., 1986), InsightII-Homology + Modeller (IHM) and Swiss-PdbViewer (Schwede et al., 2003; Guex and Peitsch, 1997). The central dogma in

* Corresponding authors. Tel.: +91 40 2716 0123x2619; fax: +91 40 2716 0512.
E-mail address: gnsastry@iict.res.in (G.N. Sastry).

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.12.002

homology modeling techniques is the right choice of alignment between sequences. The sequence alignment techniques that differ substantially from each other have been chosen for implementation, which include CLUSTALW (Thompson et al., 1994), 3D-JIGSAW (Bates et al., 2001; Contreras-Moreira and Bates, 2002) and BLAST (Altschul et al., 1990, 1997).

Different classes of membrane proteins are chosen from the PDB for modeling, to ensure diversity. The real challenge is to model structures having low sequence identities with the templates available in the PDB, but which are well within the limits of obtaining reasonably good models. For modeling membrane proteins this is a challenge, since most of these proteins have low sequence identity with the structures available in PDB (Berman et al., 2000). Therefore, template structures whose percentage identity ranges between 25% and 32% with the target sequence whose 3D-structure is known have been chosen to build the models. The 3D models obtained for the target sequence, are compared with the original PDB structure and validated by implementing varied protocols which include root mean square deviation (RMSD) (Carugo and Pongor, 2001), self-compatibility scores (SCS) (Luthy et al., 1992; Eisenberg et al., 1997), template modeling (TM) score (Zhang and Skolnick, 2004) and Ramachandran validation (RC) (http://raven.bioc.cam.ac.uk/rampage.php). Different combinations of the abovementioned methods and sequence alignments lead to 17 methodologies and the built models were subsequently subjected to structural assessment. The whole process was repeated for 20 proteins and the results are generalized to understand the performance of the methodologies. The PDB codes of the considered test and template proteins, together with the structural alignment, number of residues present in loops, self-compatibility scores of the individual proteins are provided in the supporting information. It should be noted that very few structures of membrane proteins are present in the PDB which have a homologous counterpart in the database having high crystallographic resolution and low *R*-factor. The term methodology will be used for the combination of a method and sequence alignment technique throughout the manuscript. The methodologies are finally ranked according to each validation protocol wherein two different ranking schemes are proposed and implemented.

2. Materials and methods

At the outset, crystal structures of membrane proteins belonging to different classes are selected from the PDB so as to ensure diversity. A homologous protein in the PDB having sequence identity between 25% and 32% with the target sequence is identified first. The resolution and *R*-value of the templates selected from PDB was less than 3.0 Å and 0.3, respectively. Care was taken to chose target and template structures which have good structural alignments with low RMSD values. Structural comparison between the target and template protein structures is made to gauge the intricacy of the task (see supporting information). The RMSD values between the target and the template are by and large below 3 Å. Pairs of proteins with variable total loop lengths have been chosen for modeling. Since, the difference in loop lengths plays a major role in protein modeling, we have evaluated the total number of residues present in the loops of the target and the template. Although, the maximum difference in the number of loop residues is 38, the RMSD is quite low (1.2 Å). In contrast, the target–template pair with the highest RMSD has a difference of only 10 loop residues. Thus, the difference in loop lengths do not seem to have any direct correlation with the structural alignment, at least in the chosen data set. In all cases, the SCS values of the target as well as the templates are well within the accuracy limits indicating the reliability of the fold description of the models obtained based on these PDB structures.

The chosen methods adopt varied protocols to fabricate the 3D models of the proteins. Modeller implements comparative protein structure modeling by satisfaction of spatial restraints and performs optimization of various models of protein structure with respect to a flexibly defined objective function (Sali and Blundell, 1993). The method IH, which mostly depends on human interface, builds the model depending on target and template alignment wherein the loops are built using de novo prediction using random conformational searches based on energy criteria. The side chain conformations are optimized using a rotamer library (Dunbrack, 2002). The model created by IH (http://www.accelrys.com/insight/) is further refined by Modeller (IHM) and compared with the initial model. The method implemented by SPV uses constraint space programming (CSP) to model insertions or deletions in the target–template alignment after modeling the core, based on averaging the backbone atom position of the template (Schwede et al., 2003; Guex and Peitsch, 1997). The best loop is built using a force field and interaction energy based scoring scheme and the side chain conformations are selected based on the rotamer library and a scoring function (Dunbrack, 2002). MOE builds the initial partial geometry based on conserved regions and treats insertions and deletions by using a specialized logic (Needleman and Wunsch, 1970). The loops are initially modeled randomly and subsequently, a contact energy function is implemented to identify the possible candidates which are chosen based on Boltzmann weighted averaging (Sippl, 1993a,b).

The chosen sequence alignment techniques differ substantially from each other. BLAST optimizes a measure of local similarity termed the maximal segment pair (MSP) score (Altschul et al., 1990) whereas CLUSTALW implements a progressive multiple sequence alignment method that is improved to align divergent protein sequences (Thompson et al., 1994). On the other hand, 3D-JIGSAW implements PSI-BLAST (Altschul et al., 1997) and uses position specific scoring matrix (PSSM) of the target sequence to predict secondary structure using PSI-Pred (Jones, 1999). The target PSSM files and the template are fed into a dynamic programming algorithm to obtain the best alignment based on a secondary structure based scoring function (Bates et al., 2001).

The models were validated using different protocols. The stereochemical quality of the modeled proteins is assessed from Ramachandran validation score for favored regions and allowed regions (Lovell et al., 2003). In general, a score close to 100% implies good stereochemical quality of the models. The structural similarity between the model and the real structure from