Biochimie 115 (2015) 155–161

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Research paper

An empirical energy function for structural assessment of protein transmembrane domains

Guillaume Postic [a, b, c, d, *], Yassine Ghouzam [a, b, c, d], Jean-Christophe Gelly [a, b, c, d, *]

[a] Inserm U1134, Paris, France
[b] Université Paris Diderot, Sorbonne Paris Cité, UMR_S 1134, Paris, France
[c] Institut National de la Transfusion Sanguine, Paris, France
[d] Laboratory of Excellence GR-Ex, Paris, France

ARTICLE INFO

Article history:
Received 19 January 2015
Accepted 21 May 2015
Available online 2 June 2015

Keywords:
Membrane protein
Model quality assessment
Empirical potential
Kernel density estimation

ABSTRACT

Knowing the structure of a protein is essential to characterize its function and mechanism at the molecular level. Despite major advances in solving structures experimentally, most membrane protein native conformations remain unknown. This lack of available structures, along with the physical constraints imposed by the lipid bilayer environment, constitutes a difficulty for the modeling of membrane protein structures. Assessing the quality of membrane protein models is therefore critical.

Using a non-redundant set of 66 membrane protein structures (41 alpha and 25 beta), we have developed an empirical energy function for the structural assessment of alpha-helical and beta-sheet transmembrane domains. This statistical potential quantifies the interatomic distance between residues located in the lipid bilayer. To minimize the problem of insufficient sampling, we have used kernel density estimations of the distance distributions. Following a leave-one-out cross-validation procedure, we show that our method outperforms current statistical potentials in discriminating correct from incorrect membrane protein models. Furthermore, the comparison of our distance-dependent statistical potential with one optimized on globular proteins provides insights into the rules by which residues interact within the lipid bilayer.

1. Introduction

Advances in computational modeling of protein molecules have greatly facilitated access to tridimensional structure information, by enabling prediction of protein conformations and by accelerating the final steps of their experimental determination. Extending the available structural data to more proteins would improve our understanding of their biological functions and would enable the design of peptides and proteins with medical or industrial applications.

Membrane proteins account for more than 25% of all human proteins and nearly 50% of current and future drug targets [1,2], but structural data exist for only ~400 of these proteins (http://www.blanco.biomol.uci.edu/mpstruc/). Many comparative modeling methods can be applied to membrane proteins, e.g., GPCR structures [3], but they only cover a few of them (~10% of all human membrane proteins), due to the small number of available template structures [4]. Current methods of *de novo* protein structure prediction have been adapted to membrane proteins, whether they be based on energy minimization [5,6], fragment search [7] or coevolution, the latter having recently been used to predict the structures of large proteins with unprecedented precision [8,9].

To be useful, a protein structural model must be close enough to the native conformation, which is considered to be the one that has the lowest Gibbs free energy in the native conditions [10,11]. An accurate free energy potential would therefore enable to distinguish between correct and incorrect structural models. Physics-based potentials, which are developed from molecular mechanics calculations, like CHARMM [12], Amber [13] or GROMOS [14] force fields, can be used to evaluate structural model quality. Besides, there are statistical potentials, which are scoring functions built from the statistical analysis of experimentally determined protein structures and for which the global minimum corresponds to the native conformation [15,16]. These empirical energy functions

* Corresponding authors. INSERM UMR_S 1134, DSIMB, Universite Paris Diderot, Paris 7, INTS, 6, rue Alexandre Cabanel, 75739 Paris Cedex 15, France.
E-mail addresses: guillaume.postic@univ-paris-diderot.fr (G. Postic), jean-christophe.gelly@univ-paris-diderot.fr (J.-C. Gelly).

http://dx.doi.org/10.1016/j.biochi.2015.05.018
0300-9084/

outperform physics-based potentials in assessing protein model quality [17–21], as in many other molecular modeling domains, such as *ab initio* protein folding [22–28], protein–protein docking [29,30] and fold recognition [31–33]. The latest type of scoring functions are composite methods, which combine into a single score several quality measures based on different structural features, such as interatomic distance, solvent accessibility or secondary structure [34–38], several training methods having been used to optimally weigh the individual contributions to the total score, including advanced machine learning methods, such as neural networks or support vector machines.

Membrane protein structures could be directly assessed using statistical potentials developed so far. However, these potentials are optimized on water-soluble proteins and describe the effect of a homogeneous 'implicit' solvent on atomic interactions. Therefore, since membrane proteins are shared between two environments, aqueous and organic, currently available statistical potentials would likely fail to correctly assess membrane protein structures. They would also fail because, given their optimization on globular proteins, these statistical potentials do not take into account the specific distribution of amino acids in membrane proteins resulting from the physical constraints imposed by the lipid bilayer environment. Building a statistical potential from a set of membrane proteins would probably improve the quality assessment of membrane protein structures. Nevertheless, given the few native structures available, this would likely require to solve a problem of insufficient sampling. While several single- and multiple-component scoring functions have been developed specifically for membrane proteins [39–44], only one statistical potential has been dedicated to the model quality assessment of alpha-helical and beta-sheet membrane protein models (MEMEMBED [45]). This method depends on the membrane depth of residues and uses two separate pseudo-energy matrices for alpha and beta protein structures.

In this article, we present the first empirical energy function optimized on both alpha and beta membrane proteins structures (MAIDEN, Model quality Assessment for Intramembrane Domains using an ENergy criterion). Our statistical potential quantifies the interatomic distance between all 20 standard residue types and focuses on intramembrane residues. To overcome the problem of undersampling, a smoothing of each interatomic distance distribution has been performed using kernel density estimations. The efficiency of MAIDEN has been evaluated on i) 700 predicted alpha-helical and beta-sheet protein structures built by comparative modeling and representing 76 unique membrane protein targets, and ii) 15340 models generated by (and published along with) the *de novo* prediction method for membrane protein structures EVfold_membrane [8]. We also have evaluated the efficiencies of two other methods: the above-mentioned membrane-specific MEMEMBED and DOPE [46], one of the most cited statistical potential, integrated to the widely-used MODELLER program [47,48]. Thus, compared with these two quality assessment programs, MAIDEN is more efficient in discriminating correct from incorrect models of membrane proteins, for both alpha-helical and beta-sheet intramembrane domains. Finally, the similarity between the formalisms of DOPE and MAIDEN, two distance-dependent statistical potentials, opens the door for physico-chemical interpretation of their performance, by comparing their pseudo-energy profiles.

2. Methods

2.1. Set of native membrane protein structures

The set of structures used for the calculation of MAIDEN contains 66 representative structures (41 alpha and 25 beta) from the PDBTM database [49] determined by crystallography at $\leq$2.5 Å resolution and with an *R*-factor $\leq$0.3 (Table S2). These representative structures share no more than 30% sequence identity with each other. The list was culled by entries from 1709 PDBTM structure files, using the PISCES server [50,51].

It has recently been shown that fold conservation in transmembrane regions requires less sequence identity than for water-soluble proteins [52]. Indeed, the hydrophobic residues in the membrane can change without altering the conformation, resulting in similar membrane protein structures that have much lower sequence identities than for globular proteins. Therefore, after having filtered the sequence redundancy with PISCES, we further reduced the redundancy of the optimization dataset in terms of structural similarity, in order to avoid overfitting and consequent overestimation of the performance. Thus, guided by the classification of the Orientations of Proteins in Membranes (OPM) database [53], we discarded from our training set proteins with striking structural similarities.

The assignments of intramembrane residues were obtained using the TMDET web server [54].

2.2. Calculating MAIDEN

Our statistical potential of mean force quantifies pairwise interatomic distances. Like most of the well-established statistical potentials, MAIDEN is calculated by applying the inverse Boltzmann law to discrete distance distributions derived from a sample of native structures [16]. Thus, the interaction potential of two atom types i and j is estimated as:

$$\overline{u}_{i,j}(d) = -k_B T \ln\left[\frac{f_{i,j}^{OBS}(d)}{f^{REF}(d)}\right]$$

where k_B is the Boltzmann constant and T is the Kelvin temperature. $f_{i,j}^{OBS}(d)$ is the observed frequency of finding two atom types i and j within a distance bin $[d, d + \Delta d]$ in native protein conformations. $f^{REF}(d)$, called the reference state, is the expected frequency of finding two atom types i and j within the distance bin in random protein conformations without specific interactions between the amino acids. The reference state aims at eliminating the pairwise correlations of atoms not due to physical interactions. Proximate amino acids in primary sequence have geometrically constrained interatomic distances, due to the covalent peptide bonds between adjacent residues. Therefore, $\overline{u}_{i,j}(d)$ is calculated between atoms more than 4 residues apart, in order to reduce the geometrical bias introduced by sequence proximity. $f_{i,j}^{OBS}(d)$ and $f^{REF}(d)$ are calculated by $N_{i,j}(d)/\Sigma_d N_{i,j}(d)$ and by $\Sigma_{i,j}N_{i,j}(d)/\Sigma_d\Sigma_{i,j}N_{i,j}(d)$, respectively, where $N_{i,j}(d)$ is the number of atom type pairs (i, j) at a distance d within $[d, d + \Delta d]$. For the optimization of MAIDEN, while the reference state was calculated for all atom types, $f_{i,j}^{OBS}(d)$ was calculated only for Cα. Thus, 20 atom types were considered for $f_{i,j}^{OBS}(d)$ and 167 atom types were considered for $f^{REF}(d)$ (e.g., 11 atom types for the arginine: R-N, R-CA, R-CB, R-CG, R-CD, R-NE, R-CZ, R-NH1, R-NH2, R-C and R-O). Calculations were simplified by considering $k_B T$ equal to 1. A maximum value of $\overline{u}$ had to be defined, due to the fact that no atomic pair is observed for short distance bins, which would result in divisions by 0 when calculating the corresponding pseudo-energies. Thus, we set the upper limit of $\overline{u}$ at 10, an arbitrary value in the same order of magnitude as the variation amplitude of all pairwise potentials. For membrane proteins, a number of experimental and theoretical structures available are backbones or 'Cα-only'. MAIDEN only uses Cα to assess a model, which makes our method compatible with low resolution structures.