BRANEart: identify stability strength and weakness regions in membrane proteins

Sankar Basu\textsuperscript{1,2}, Simon S. Assaf\textsuperscript{1}, Fabian Teheux,\textsuperscript{1}
Marianne Rooman\textsuperscript{1,3,†} and Fabrizio Pucci\textsuperscript{1,3,†,∗}

\textsuperscript{1}Computational Biology and Bioinformatics, Université Libre de Bruxelles, Brussels, Belgium
\textsuperscript{2}Department of Microbiology, Austosh College, under University of Calcutta, Kolkata, India
\textsuperscript{3}Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium

† These authors have contributed equally to this work.

August 22, 2021

Abstract

Understanding the role of stability strengths and weaknesses in proteins is a key objective for rationalizing their dynamical and functional properties such as conformational changes, catalytic activity, and protein-protein and protein-ligand interactions. We present BRANEart, a new, fast and accurate method to evaluate the per-residue contributions to the overall stability of membrane proteins. It is based on an extended set of recently introduced statistical potentials derived from membrane protein structures, which better describe the stability properties of this class of proteins than standard potentials derived from globular proteins. We defined a per-residue membrane propensity index from combinations of these potentials, which can be used to identify residues which strongly contribute to the stability of the transmembrane region or which would, on the contrary, be more stable in extramembrane regions, or vice versa. Large-scale application to membrane and globular proteins sets and application to tests cases show excellent agreement with experimental data. BRANEart thus appears as a useful instrument to analyze in detail the overall stability properties of a target membrane protein, to position it relative to the lipid bilayer, and to rationally modify its biophysical characteristics and function. BRANEart can be freely accessed from http://babylone.3bio.ulb.ac.be/BRANEart.

1 Introduction

The broad family of integral membrane proteins provides indispensable components of living cells. Being embedded in biological lipid membranes, these proteins include important scaffolds and functional sites, which bind targeted molecules floating around in the cytosol or in the extracellular medium. They therefore serve as attractive drug targets [1, 2].

Stability and physico-chemical characteristics greatly vary between transmembrane (TM) and extramembrane (EM) regions of an integral membrane protein due to the difference in their surrounding chemical

*Correspondence at Fabrizio.Pucci@ulb.be or Marianne.Rooman@ulb.be
environment. EM domains in the cytosolic or extracellular medium resemble globular proteins as their surfaces are exposed to water, while residues embedded within the membrane are exposed to lipids and are thus characterized by an elevated hydrophobicity [3, 4]. Because of this mixed environment, the study of folding and stability of membrane proteins is very challenging and only few tools have been dedicated to investigate them [5, 6, 7, 8]. Another reason why membrane proteins are less studied than globular proteins is their lower number of experimental three-dimensional (3D) structures [9].

Just as for globular proteins, the native structure of a membrane protein corresponds to the global minimum of the free energy landscape in physiological conditions. In general, however, some protein residues or regions, taken individually, are not in their global free energy minimum and correspond to what we call stability weaknesses [10, 11, 12]. Indeed, not all residues can simultaneously adopt their lowest free energy conformations, because of the polypeptide chain which constraints the relative motions of the residues. Native conformations can be viewed as the best compromise between conflicting interactions. We define strong and weak residues as residues that are, or are not, optimized for stability properties, respectively. For example, stability weaknesses occur when a contact between two residues in the native structure does not have a favorable free energy contribution and can thus easily break, change conformation, and/or remain flexible. They often play a key role in protein function, interactions, and conformational changes. The concept of stability weaknesses is close to the notion of frustration except that the latter also includes kinetic constraints on fast folding [13].

To further advance these issues, we implemented and expanded a series of new statistical mean-force potentials designed to specifically describe the stability properties of membrane proteins, which were first defined in [3]. Based on these potentials, here we derived a per-residue membrane propensity index which predicts whether residues situated in the membrane have a stabilizing contribution or would prefer to be in EM regions, and similarly for residues outside the membrane. This index thus identifies strong and weak residues in EM and TM regions, and can also be used to predict how a protein is, or is not, inserted in the membrane.

It has to be noted that we are studying in this paper relative, EM/TM, strengths and weaknesses, defined from the difference in folding free energy according to whether a residue is in one or the other region. They are different from the strengths and weaknesses defined in globular proteins as residues of which the folding free energy contribution is either highly optimal or not optimal at all, as for example computed by the SWOTein webservice [12].

We made BRANEart freely available in the form of a web server called BRANEart, designed to help the scientific community to explore stability strengths and weaknesses in membrane proteins, which is a key element in the study of their stability and function.
2 MATERIALS AND METHODS

2.1 Protein structure data sets

The membrane protein data set $D_{\text{mem}}$ comes directly from a recent study [3]. It consists of 163 X-ray structures of integral membrane proteins that have been collected from the Protein Data Bank (PDB) [14]. The selected structures have a resolution of at most 2.5 Å and a pairwise sequence identity of at most 30%, computed using the PISCES protein sequence culling server [15].

The set $D_{\text{mem}}$ contains 107 $\alpha$-helical and 52 $\beta$-barrel integral proteins, and 4 $\alpha$-helical bitopic integral proteins that do not span the lipid bilayer completely. Each of these proteins was annotated using OPM ( Orientations of Proteins in Membranes) [16], a curated web resource that positions the biological lipid bilayer on experimentally resolved structures of integral membrane proteins and membrane-bound peptides. Using these annotations, the TM and EM portions of each protein were further segregated in two subdata sets $D_{\text{TM mem}}$ and $D_{\text{EM mem}}$.

All selected protein chains in $D_{\text{mem}}$ were considered in the context of their biological assembly (or biounit), which corresponds to their functional quaternary conformation. This ensures a more realistic representation of the surrounding protein environment experienced by the protein. The biological units were taken to be those defined by the authors of the X-ray structures or, in absence of annotations from the authors, as those predicted by the Protein Interfaces, Surfaces and Assemblies (PISA) tool [17].

Finally, we also set up a second data set $D_{\text{glob}}$ of 4,860 monomeric globular protein structures from the PDB, to be used as an independent set for validating our method. The proteins from this data set have a monomeric biological unit, a good quality X-ray structure with a maximum resolution of 2.5 Å and a pairwise sequence identity of at most 25%.

The list of all proteins belonging to the data sets $D_{\text{mem}}$ and $D_{\text{glob}}$ are given in the GitHub repository github.com/3BioCompBio/BRANEart.

2.2 Statistical potentials

Statistical potentials are coarse-grained mean-force energy functions derived from frequencies of associations of sequence and structure motifs in a data set of known protein structures. These frequencies are transformed into free energies using the inverse Boltzmann law [18, 19, 20]. These potentials depend on the characteristics of the data set from which they are derived. For example, temperature-dependent statistical potentials are obtained from protein data sets of different melting temperatures [21, 22], and solubility-dependent potentials from proteins of different solubility [23].

Here we derived a series of membrane protein potentials from the sets $D_{\text{TM mem}}$ and $D_{\text{EM mem}}$, extending the work of [3]. More precisely, we considered the following first and second order statistical potential terms [20]:

$$
\Delta W_{xy} = -k_B T \ln \frac{F(x,y)}{F(x) F(y)}
$$

$$
\Delta W_{xyz} = -k_B T \ln \frac{F(x,y,z)}{F(x,y) F(x,z) F(y,z)}
$$

where $k_B$ is the Boltzmann constant, $T$, the absolute temperature conventionally taken to be room temperature, and $F$, the relative frequencies computed from a given data set of protein structures. The variables $x, y, z$ stand for any of the four elementary structure or sequence descriptors $s, d, t$ and $a$: $s$ is an amino acid
type, \( d \), the spatial distance between the average side chain geometric centers of two residues separated by at least one residue along the polypeptide chain [19], \( t \), a \((\phi, \psi, \omega)\) backbone torsion angle domain [24], and \( a \), a solvent accessibility bin where the solvent accessibility is defined as the ratio (in %) between the solvent accessible surface area of a residue in the structure and in an extended Gly-X-Gly conformation [19].

We constructed two versions of each of the potentials defined by Eqs (1)-(2). In the first, all frequencies \( F \) were computed from the structure set \( D_{\text{mem}}^{\text{TM}} \), i.e. considering only protein regions embedded in the lipid membrane. In the second, all frequencies \( F \) were computed from \( D_{\text{mem}}^{\text{EM}} \), thus considering only extramembrane protein regions. The potentials extracted from TM regions, \( \Delta W_\chi^\text{TM} \) with \( \chi = xy \) or \( xyz \), describe the stability properties of membrane proteins inside the lipid membrane, while the potentials extracted from EM regions, \( \Delta W_\chi^\text{EM} \), describe the stability properties outside the lipid bilayer. Note that the only membrane protein potentials that we constructed and analyzed earlier are the inter-residue distance potentials \( \Delta W_{sd}^\mu \) and \( \Delta W_{sds}^\mu \) [3], where \( \mu \) is either EM or TM.

The full list of 19 membrane statistical potentials derived here and their characteristics are given in Section 1 of Supplementary Materials.

### 2.3 Per-residue folding free energies

The contribution of each residue \( i \) in a protein to its overall folding free energy computed with a given potential \( \Delta W_\chi^\mu \), noted \( \Delta G_{i,\chi}^\mu \), was computed in such a way that an equal amount was allocated to each of the residues that carry the structural descriptors \( a \), \( t \) and \( d \) included in \( \chi \). According to the number of residues that the structural descriptors encompass, we used the following rules [11, 12]:

- if the structure descriptors involved in \( \chi \) are localized on a single residue \( i \) (which is the case when e.g., \( \chi = st, sa, sst, ssa \)), the total contribution is assigned to that residue;
- if the structure descriptor is localized on two residues \( i \) and \( j \) (e.g., \( \chi = sd, tt, aa, sds, saa, stt \)) half of the energy contribution is assigned to each of these residues;

For the potential \( \chi = ss \) that does not contain any structure descriptor, an equal amount was allocated to each residue carrying a sequence descriptor \( s \).

For example, let us consider the three potentials \( \Delta W_{st}^\mu \), \( \Delta W_{saa}^\mu \) and \( \Delta W_{sds}^\mu \). The per-residue folding free energy contributions are:

\[
\Delta G_{i,\chi}^\mu = \sum_{j=1}^{L} \Delta W_{s_{i,j}}^\mu
\]

\[
\Delta G_{i,\chi}^{s,j,\mu} = \frac{1}{2} \sum_{j,k=1}^{L} \Delta W_{s_{k,j,a_j,a_i}}^\mu
\]

\[
\Delta G_{i,j}^{s,ds} = \frac{1}{2} \sum_{j=1}^{L} \Delta W_{s_{i,d_j,s_j}}^\mu
\]

where \( i, j \) and \( k \) denote positions along the amino acid sequence and \( L \) the sequence length. For further details, we refer the reader to previous studies [11, 3, 12].

In a last step, the folding free energy values so obtained for each residue \( i \) were smoothed by taking a
weighted average over the 5-residue sequence window $[i-2,i+2]$ [3]:

$$\Delta G_{\chi}^{i,\mu} = \frac{1}{1 + 2\gamma + 2\beta} (\gamma \Delta G_{\chi}^{i-2,\mu} + \beta \Delta G_{\chi}^{i-1,\mu} + \Delta G_{\chi}^{i,\mu} + \beta \Delta G_{\chi}^{i+1,\mu} + \gamma \Delta G_{\chi}^{i+2,\mu})$$  \hspace{1cm} (6)

The weighting parameters $\gamma$ and $\beta$ were chosen to minimize the level of weaknesses in the transmembrane regions (see Section 2.5). For the N- and C-terminal residues, the smoothing was done on the residues present in the $[i-2,i+2]$ sequence interval.

In this way, each residue $i$ was tagged with two folding free energy values $\Delta G_{\chi}^{i,EM}$ and $\Delta G_{\chi}^{i,TM}$ for each statistical potential $\chi$, irrespective of its location, in either EM or TM regions.

2.4 Membrane propensity Index

We introduced a per-residue membrane propensity index MPr$^i$ to predict to what extent a residue $i$ in a folded protein corresponds to a stability strength or to a weakness when placed in a given, lipid or aqueous, environment. From a physico-chemical perspective, MPr$^i$ estimates whether a residue shows a preference for the EM or TM environments, and can be interpreted as an index of stability of a residue within its structural context. It is defined as a linear combination of all folding free energy terms derived in the previous subsection:

$$\text{MPr}^i = \sum_{\mu} \sum_{\chi} \alpha_{\chi}^{i,\mu} \Delta G_{\chi}^{i,\mu} + \alpha_L \log L + \alpha_N$$ \hspace{1cm} (7)

where $\alpha_{\chi}^{i,\mu}$ are real-valued parameters that need to be optimized (see Section 2.5). The MPr$^i$ score is built in such a way that it returns a value close to one for residues predicted to be stable in lipid environments, and a value close to zero for residues predicted to be stable in aqueous environments.

Residues that strongly contribute to the stability of the region to which they belong are considered as stability strengths, and residues that would be more stable elsewhere in the protein are called stability weaknesses.

2.5 Training and optimization

To estimate the membrane propensity index MPr$^i$ introduced in Eq. (7), we have to identify the 40 free parameters introduced in Eqs (6-7). We performed this parameter identification by minimizing the overall amount of structural weaknesses in the proteins from the $D_{\text{mem}}$ set. The general idea behind this procedure comes from the minimal frustration principle [13] stating that proteins have evolved, and still evolve, to optimize the folding energy landscapes.

To perform the parameter identification on the $D_{\text{mem}}$ dataset, we need to know if each of the residues in the proteins of this set is in an EM or TM region. For this, we searched the residue annotations in the OPM database [16]. These annotations were then assigned to the vector $O$, with the convention that zero corresponds to an EM region and one to a TM region. $O$ served as target function in the model training. Note that the OPM annotations are actually predictions and may also suffer from inaccuracies.

We can now define the cost function $C$ as the square of the difference between the OPM annotations and the predicted membrane propensity MPr defined in Eq. (7):

$$C = \sum_i (\text{MPr}^i - O^i)^2$$ \hspace{1cm} (8)
where the sum is over all residues in the training set \( D_{\text{mem}} \). The parameter identification was done with Python (v.3) using the regression subroutine LinearRegression. Note that logistic regressions were also tested in the training process, but appeared to yield less good results.

The performance of the predictor was evaluated on the training data set \( D_{\text{mem}} \) using a leave-one-out cross validation procedure, in which \( D_{\text{mem}} \) but one protein was used as a training set and the removed protein as a test set. The rounds of training and testing were repeated as often as there are proteins in \( D_{\text{mem}} \) (163). As a final result of the training, the BRANEart model outputted a continuous membrane propensity score \( \text{MPr} \) varying approximately between \{-0.5, 1.5\}. Values that are close to one identify residues that are stable in the phospholipid bilayer, while values close to zero represent residues that are stable in the aqueous solvent.

Note that all the cross validations that we performed in this paper are strict: the target protein is excluded from all steps of our computations, from the derivation of the statistical potentials to the computation of \( \text{MPr} \) and the identification of \( \phi_0 \), and the sequence identity between learning and test set proteins is at most 30%.

### 2.6 EM/TM prediction

The membrane propensity score \( \text{MPr} \) was also used to set up a binary classifier that predicts whether residues in a membrane protein belong to TM or to EM regions. For that purpose, we transformed the continuous \( \text{MPr} \) scores into a discrete binary function, where 0 means EM and 1 TM, using an appropriate cutoff value \( \phi_0 \) such that a residue \( i \) is predicted to be in the TM region if \( \text{MPr}_i \geq \phi_0 \) and in the EM region otherwise.

The value of \( \phi_0 \) was identified to minimize the difference with OPM assignments.

To evaluate the performance of this classifier, we used the balanced accuracy (BACC) defined as:

\[
BACC = \frac{1}{2} \left( \frac{TP}{P} + \frac{TN}{N} \right)
\]

where \( TP \) and \( TN \) mean true positive and true negatives, respectively, and \( P \) an \( N \) positives and negatives.

### 3 Results and Discussion

#### 3.1 Setting up BRANEart

We set up the BRANEart predictor, which predicts for each residue in a protein structure a membrane protein index \( \text{MPr} \), defined as a linear combination of several types of statistical potentials, derived from either EM or TM regions of a set of membrane protein structures (\( D_{\text{EM} \text{mem}} \) and \( D_{\text{TM} \text{mem}} \)), as defined in Eqs (1)-(7). The coefficients of the linear combination were identified to follow OPM annotations [16], through the minimization of Eq. (8), with the definition that an \( \text{MPr} \) close to one means a preference to be located in a TM region and an \( \text{MPr} \) close to zero, to be in an EM region. This identification was performed in strict cross validation (Section 2.5). In what follows, when computing \( \text{MPr} \) on the sets \( D_{\text{EM} \text{mem}} \) and \( D_{\text{TM} \text{mem}} \), we used the cross-validated values.

The \( \text{MPr} \) index yields a quantitative measure of the stabilizing or destabilizing contribution that a residue has in a specific environment, in other words, whether it acts as a weakness or a strength in that environment. Note that we compared here aqueous and lipid environments, but that this approach can be generalized to the comparison between other environments.
| Data set      | Protein type            | ⟨MPr⟩ | σ  | \(N\) |
|--------------|-------------------------|-------|----|-------|
| \(\mathcal{D}_{\text{glob}}\) | Globular proteins | 0.13  | 0.13 | \(1.3 \times 10^6\) |
| \(\mathcal{D}_{\text{mem}}\) | Membrane proteins | 0.40  | 0.37 | \(5.7 \times 10^5\) |
| \(\mathcal{D}_{\text{TM}}\) | Trans-membrane regions | 0.74  | 0.27 | \(2.3 \times 10^5\) |
| \(\mathcal{D}_{\text{EM}}\) | Extra-membrane regions | 0.18  | 0.22 | \(3.4 \times 10^5\) |

Table 1: Mean per-residue index \(\langle \text{MPr} \rangle\) and standard deviation \(\sigma\) of the distributions computed for the four protein structure sets considered in this paper; \(N\) is the number of residues in the data set.

3.2 Large-scale application of BRANEart

We applied BRANEart to the membrane protein data set \(\mathcal{D}_{\text{mem}}\) and to the globular monomeric protein set \(\mathcal{D}_{\text{glob}}\) defined in Section 2.1, and computed the per-residue MPr score for the 56,715 and 1,258,648 amino acid residues contained in these two sets. Note that the huge difference in size between these two sets, of more than an order of magnitude, is due to the paucity of experimental structures available for membrane proteins.

As seen in Table 1, the mean \(\langle \text{MPr} \rangle\) value is close to zero in globular proteins, \(\langle \text{MPr} \rangle = 0.13\), with a low standard deviation of \(\sigma = 0.13\). This means that the large majority of the residues are stabler in an aqueous than in a lipid environment, which is obviously the case. Residues in membrane proteins have intermediate preferences \((\langle \text{MPr} \rangle = 0.40)\) with a much larger \(\sigma\) of 0.37. However, splitting membrane proteins into TM and EM regions clarifies these findings: residues in EM regions clearly prefer to be in an aqueous environment \((\langle \text{MPr} \rangle = 0.18)\), while residues inside the membrane prefer a lipid environment \((\langle \text{MPr} \rangle = 0.74)\).

These excellent results are a first validation of BRANEart.

The MPr distributions are plotted in Fig. 1 for each of the four data sets. They all show a unimodal bell-shaped distribution, except the one computed from \(\mathcal{D}_{\text{mem}}\) which is more spread out with two peaks, a narrow peak and a flatter one. This distribution is the union of the two unimodal distributions computed from the sets \(\mathcal{D}_{\text{TM}}\) and \(\mathcal{D}_{\text{EM}}\), which are shifted relative to each other and show a marked preference for lipid and aqueous environments, respectively. The distribution from \(\mathcal{D}_{\text{EM}}\) is roughly centered around the same MPr value as the one obtained from \(\mathcal{D}_{\text{glob}}\), but it is much less peaked although the proteins from both sets are in an aqueous environment. This suggests that there are more stability weaknesses in membrane proteins than in globular proteins or more precisely, that there are more residues in the EM regions of membrane proteins than in globular proteins which would prefer to be in a lipid environment.

To graphically visualize here and in the BRANEart webserver the level of stability of a residue, we defined stability classes in terms of the mean and standard deviations of the MPr distribution and we colored the residues accordingly, as explained in Table 2. Residues that are strong in aqueous environments but weak in lipid environments are colored in blue. Residues that are strong in the phospholipid bilayer and weak in aqueous solvent are colored in green. Different color graduations were defined to further differentiate between highly stable, stable, moderately stable and mildly stable residues in EM regions, and equivalently in TM regions.

Interestingly, the amount of strengths and weaknesses in TM and EM regions are almost identical, as computed from the MPr distributions: 7% of the residues are weak in both regions, 75% are strong and the remaining ones are neutral. Without surprise, most residues thus contribute to the stabilization of the regions to which they belong; note that the functional residues are usually among the weak residues.

We also examined how the MPr values change as a function of the solvent accessibility in globular proteins. Therefore, we divided all the residues in \(\mathcal{D}_{\text{glob}}\) into three groups: core residues with accessibility \(\leq 20\%\),
Table 2: Classes of MPr scores defined in terms of the mean and standard deviations of the MPr distribution. 

### Color Index range Type

| Color | MPr \[^i\] ≤ \(\mu - 2\sigma_{\text{EM}}\) | Highly stable in water |
|-------|------------------------------------------|------------------------|
| Dark blue | \(\mu - 2\sigma_{\text{EM}} < \text{MPr} \[^i\] ≤ \(\mu - 3/2\sigma_{\text{EM}}\) | Stable in water |
| Medium blue | \(\mu - 3/2\sigma_{\text{EM}} < \text{MPr} \[^i\] ≤ \(\mu - \sigma_{\text{EM}}\) | Moderately stable in water |
| Light blue | \(\mu - \sigma_{\text{EM}} < \text{MPr} \[^i\] ≤ \(\mu - 1/2\sigma_{\text{EM}}\) | Mildly stable in water |
| Olive | \(\mu - 1/2\sigma_{\text{EM}} < \text{MPr} \[^i\] ≤ \(\mu + 1/2\sigma_{\text{TM}}\) | Neutral |
| Green | \(\mu + 1/2\sigma_{\text{TM}} < \text{MPr} \[^i\] ≤ \(\mu + \sigma_{\text{TM}}\) | Mildly stable in lipids |
| Light green | \(\mu + \sigma_{\text{TM}} < \text{MPr} \[^i\] ≤ \(\mu + 3/2\sigma_{\text{TM}}\) | Moderately stable in lipids |
| Green | \(\mu + 3/2\sigma_{\text{TM}} < \text{MPr} \[^i\] ≤ \(\mu + 2\sigma_{\text{TM}}\) | Stable in lipids |
| Dark green | \(\mu + 2\sigma_{\text{TM}} ≤ \text{MPr} \[^i\] | Highly stable in lipids |

\(\mu\) is the mean of the MPr distribution computed on \(D_{\text{mem}}\) and \(\sigma_{\text{TM}}\) and \(\sigma_{\text{EM}}\) the standard deviations of the MPr distributions computed from \(D_{\text{TM}}\) and \(D_{\text{EM}}\), respectively. The colors used in the visualization frameworks of the BRANEart webserver are here defined.

Intermediate residues with 20% < accessibility ≤ 50%, and surface residues with accessibility > 50%. The average MPr score in these groups gradually drops from core to surface: 0.15 (\(\sigma=0.15\)), 0.13 (\(\sigma=0.10\)), 0.08 (\(\sigma=0.10\)). As seen in Fig. 2, the whole MPr distribution is shifted towards lower values, and becomes more peaked. This reflects the fact that residues in the core of globular proteins locally feel a hydrophobic environment and have a slightly higher MPr, while surface residues are in direct contact with water and a lower MPr.

### 3.3 Application to α-helical and β-barrel membrane proteins

Membrane proteins of which the TM region has an α-helical or a β-barrel conformation exhibit different folding and stability properties [3, 4]. Indeed, the former type of proteins are essentially localized in the cytoplasmic membranes of eukaryotic and prokaryotic cells and quite rarely in outer membranes, whereas the latter proteins are found in outer membranes of gram-negative bacteria, mitochondria or chloroplasts.

We observe from the distributions depicted in Fig. 3.a-b that residues pertaining to α-helical folds have a bigger preference for lipid environments than those pertaining to β-barrel folds. Indeed, average \(\langle \text{MPr} \rangle\) values are equal to 0.44 (\(\sigma=0.40\)) and 0.34 (\(\sigma=0.31\)) for α and β proteins, respectively. If we focus on the TM regions, this tendency is even more marked (Fig. 3.c-d): \(\langle \text{MPr} \rangle = 0.78\) (\(\sigma=0.28\)) and 0.63 (\(\sigma=0.23\)).

These results are in line with a series of facts. First, α-helices are coiled structures stabilized by regularly spaced hydrogen bonds between residues at positions \(i\) and \(i + 4\) along the polypeptide chain. In contrast, β-sheets are stabilized by hydrogen bonds between residues from different β-strands, which are usually not close along the chain [25]. They thus harbour greater geometric variability than α-helices and display greater deviations from ideal backbone bond angles [26, 27]. Secondly, β-fold membrane proteins mainly have channel or porin conformations, through which molecules can cross the membrane. The internal faces of these
Figure 1: Distribution of MPPr scores computed from different protein data sets: (a) membrane proteins $D_{\text{mem}}$, (b) globular proteins $D_{\text{glob}}$, (c) TM regions $D_{\text{TM}}$ and (d) EM regions $D_{\text{EM}}$. The color code is defined in Table 2.

β-barrels are therefore more hydrophilic even though they are in TM regions. They thus often correspond to weaknesses when computed in lipid environments. Finally, the asymmetrical bilayer of phospholipids in which β-barrel proteins are usually inserted have different characteristics than standard phospholipid bilayers. This implies that our statistical potentials generated from TM regions of the whole set of membrane proteins could be less accurate for this kind of proteins (see [3] for further details).

3.4 Protein embedding in the membrane

The MPPr score can be used not only to identify weak and strong regions but also to predict the protein embedding in the lipid bilayer membrane and more specifically, whether residues are inside or outside the membrane. For that purpose, we transformed the per-residue MPPr scores into a binary function, where 0 and 1 mean EM and TM, respectively, using an appropriate threshold value as defined in the Methods section 2.6. We computed the BACC score (Eq. (9)) between the so predicted membrane assignments and OPM.
Figure 2: Distribution of MPr scores computed from the globular protein set $D_{\text{glob}}$ for residues that are (a) situated in the core, (b) half buried and (c) solvent exposed. The color code is defined in Table 2.

Figure 3: MPr distribution computed from the membrane proteins in $D_{\text{mem}}$ that have an $\alpha$-helical or $\beta$-barrel structure. (a) $\alpha$-folds (TM and EM), (b) $\beta$-folds (TM and EM), (c) TM region of $\alpha$-folds and (d) TM region of $\beta$-folds. The color code is defined in Table 2.
score is very high and we did not expect better as weak residues in TM and EM domains are by definition
wrongly predicted, even though they are essential for membrane protein functioning as we will see in the next
two subsections.

Note that we used annotations of OPM [16] to train this predictor, and that these annotations are actually
also computational predictions. Some discrepancies between the results of our prediction and OPM could
therefore be due to errors of OPM rather than of our predictor. We could thus use our predictor to yield
new EM/TM assignments, which would differ from OPM assignments in some places. Alternatively, the
MPr score could be used as a feature, in combination with other features, to improve the accuracy of the
membrane protein embedding in the membrane [16, 28, 29].

3.5 Application to Leu transporter

In general, membrane proteins undergo conformational changes to accomplish their biological functions
[30, 31]. Residues that are at the basis of these large-scale movements are often stability weaknesses [12] or
in other words, frustrated [13]. Indeed, functional constraints prevent them to adopt conformations with high
stability contributions. Here we show how we can use the MPr score to gain insight into the functional role
of these residues. Note that, to our knowledge, no other method employs dedicated energy functions to study
frustration in membrane proteins.

As an example, we analyzed leucine transporter (LeuT), a bacterial homodimeric protein containing twelve
transmembrane helices, which uses the electrochemical potential of sodium ions to transport leucine from
outside to inside the cell [32]. Both the substrate-free outward-open LeuT structure and the inward-open apo
conformation have been resolved via X-ray crystallography [33] (PDB codes 3tt1 and 3tt3, respectively).

We analyzed the difference in membrane propensity index (ΔMpr) between the two conformations in Fig
4.a. We can easily identify the two regions with highest ΔMPr in absolute value: the transmembrane helix 1
(TM1) and the extracellular helix 4 (EL4), which are key elements that allow the opening and closing of the
intracellular and extracellular gates, respectively [33]. We thus focus on these two regions and analyze their
MPr values in both outward and inward conformations (Fig. 4.b-e).

The transmembrane segment TM1 (residues 11-35) is a kinked helix formed by two helical regions, TM1a
(residues 11-21) and TM1b (residues 25-35), separated by a hinge fragment (residues 22-24). TM1a undergoes
a large movement upon opening the intracellular gate (Fig. 4.b-c). Despite the fact that the whole TM1 is
inside the lipid membrane, the hinge residues have a very low MPr score in both conformations. They are
thus stability weaknesses, which is related to their functional role in facilitating the opening and closing of the
gate, as confirmed by experiments [33]. The MPr of TM1a changes significantly during the conformational
change: in the inward-open conformation, when the intracellular gate is open, TM1 is in only weakly stable
(Fig. 4.c), whereas in the outward-open conformation, when the intracellular gate is closed, TM1a establishes
stabilizing contacts with TM5 and TM7 which is reflected by high MPr values (Fig. 4.b).

In a similar way, we identified the strong and weak residues in the outer membrane environment related
to the conformational movement that EL4 undergoes to open and close the extracellular gate. EL4 (residues
309-321) is a helix of which the last residues (319-321) are unwound and called hinge; its position appears
displaced in the outward-open and inward-open conformations. In the outward-open conformation, when the
extracellular gate is open, the N-terminal EL4 residues 309-311 are stable in the aqueous environment but
the hinge residues are weak (Fig. 4.d). In closing the gate, EL4 changes stability profile, with its N-terminus
that becomes weak and its C-terminus and hinge that get more stable (Fig. 4.c) due to extensive contacts
including hydrophobic interactions and a hydrogen bond between Ala 319 and Asp 401 in TM10.
This example illustrates how we can use the BRANEart webserver to identify in a simple way residues in membrane proteins that are not optimized for their stability but that have functional roles.

Figure 4: Leu transporter: MPr score in the outward-open (PDB code 3tt1) and inward-open (PDB code 3tt3) conformations (a) Difference in MPr score ($\Delta$Mpr) between the two conformations as a function of the sequence position; the positions of the helices TM1 and EL4 are indicated. (b-e) MPr of the TM1 and EL4 helices as a function of the sequence position, and TM1 and EL4 helices embedded in the 3D structure, colored according to the code defined in Table 2; (b) TM1, outward-open conformation; (c) TM1, inward-open conformation; (d) EL4, outward-open conformation; (d) EL4, inward-open conformation.

3.6 Application to human phospholamban

As second test case, we studied the stability of human phospholamban (PLB), a protein anchored into the cardiac sarcoplasmic reticulum membrane, which is essential to myocardial contractility [34]. Let us start with the analysis of the MPr index as a function of the sequence, computed from the standard pentameric form of PLB (PDB code 1ZLL). As shown in Fig 5.a, the two domains of the protein are easy to identify: the EM part (residues 1-22) has low MPr values and thus the propensity to be stable in water, and the TM part (residues 23-52) has overall much higher MPr values indicating its stability in lipids. A closer look shows that the TM region must be divided in two: the C-terminal part (residues 33-52) is apolar, contains the well known motif LxxIxxx [35] and has high MPr values indicating its stability in lipids, whereas the N-terminus
(residues 23-32) is a TM weakness.

Note that the latter part connects the EM and TM regions and is very close to the water-membrane interface. It thus probably changes dynamically from the lipid TM to the aqueous EM environment and vice versa. This prediction is in perfect agreement with both NMR data [36], molecular dynamics simulations [37, 35] and mutagenesis data [38, 39], which identified this region as highly dynamical, unimportant for the stabilization of the structure, but crucial for modulating functional interactions of PLB.

We also analyzed the MPr index of the different types of PLB structures. More specifically, we compared the stability of the wild-type monomeric and pentameric forms (PDB code 1ZLL) [36], of a highly stable designed pentameric variant (PL5, PDB code 6MQU) [35] and of a water-soluble tetrameric variant (WSPLB, PDB code 1YOD) [40]. The MPr values of the TM region (residues 28-50) in all these structures are reported in Fig. 5.b. The TM region of the monomeric form shows a clear preference for the lipid environment with all residues having an MPr value bigger than $\sim 0.5$. The pentamer form is predicted to be more stable than the monomer form since inter-chain interactions between the TM fragments strongly stabilize this structural assembly. Note that both the monomer and pentamer forms do exist in phospholipid bilayers even though the latter is dominant. Such oligomer equilibrium is dynamic and can change due to mutations or phosphorylation [41].

The PL5 variant has the largest MPr values and is the most stable form in the lipid environment. This is again in perfect agreement with experiments; in redesigning this protein fragment, the polar residues have been substituted by apolar residues to get stronger interchain interactions driven by apolar sidechain packing and an increased stability of the overall fold [35]. Finally, the redesigned, truncated, water-soluble variant has extremely low propensities to be in a lipid environment, as can easily be seen from our predictions.

![Figure 5: Phospholamban case study. (a) MPr values computed for the pentameric structure of the full-length phospholamban (PDB code 1ZLL); (b) amino acid sequences and MPr values for the TM fragment (residues 28-50) of four different PLB structures: PLB monomer (light green, PDB code 1ZLL), PLB pentamer (green, PDB code 1ZLL), redesigned pentamer PL5 (dark green, PDB code 6MQU) and water soluble PLB tetramer (blue, PDB code 1YOD).]
3.7 BRANEart webserver: features and functionalities

We implemented our structure-based prediction method into the easy-to-use BRANEart webserver. To run a query, the user first chooses the structure of a membrane protein by either providing its 4-letter code which is then automatically retrieved if available in the PDB, or by uploading a protein structure file in PDB format. The user is then asked to select the relevant chain(s) on which BRANEart has to perform the predictions, taking into account the other chains present in the structure file. Finally, the computation starts.

The main BRANEart output is the MPr score for each amino acid residue in the selected chain(s) of the target protein structure. In addition to returning these values as a downloadable text file, the webserver also displays a table with the MPr of each residue of the targeted protein chain(s), colored according to the code defined in Table 2. In addition, BRANEart provides a multi-featured browser tool to visualize the protein 3D structure, where each residues is colored according to its MPr score. Additionally, there are several advanced visualization functionalities:

- Possibility to hide/show specific chain(s) in the displayed structure, and zoom in and out by left dragging the mouse.
- Possibility to switch to a full screen visualization mode and take a ”.png” snapshot of the visualized structures.
- Possibility to select a residue of interest by double clicking on it. The neighboring residues are then displayed in ”ball and sticks”.
- Possibility to download a PyMol session file (.pml) to switch to the PyMol representation of the 3D structure.

For further technical details and information about the webserver, we refer to the BRANEart help page.

4 CONCLUSION

We presented BRANEart, a computational method to identify the stability strengths and weaknesses in membrane proteins. Extending and combining the newly developed membrane statistical potentials introduced in [3], we defined a MPr score that quantifies whether a residue is stable in a lipid or aqueous environment. Large-scale predictions and applications to test cases show BRANEart’s ability to correctly identify regions in their respective environment that strongly contribute to the stabilization of their host protein, and residues that have instead low impact on stabilization but have functional roles.

Note that our approach can be extended to identify strengths and weaknesses in other environments than the membrane, for example in hot and cold environments using temperature-dependent potentials [21].

We additionally provided a user friendly webserver that, on the basis of the 3D structure of the target membrane protein, computes the MPr score for each residue in the input structure. Visualization tools are provided which simplify the understanding and interpretation of BRANEart results.

To our knowledge, BRANEart is the first accessible, fast and accurate tool that use dedicated membrane protein potentials to identify stability strengths and weaknesses in membrane proteins. The use of such mean force potentials drastically simplifies the analysis of membrane protein stability, since the effect of the lipid environment is implicitly considered. BRANEart can be used in a wide series of applications that range from the analysis of the conformational changes in membrane proteins, the localization of proteins with respect to
the lipid membrane, to the identification of residues to target for the rational modification of biophysical characteristics of membrane proteins.

Acknowledgments

SB wants to convey his sincere gratitude to Dr. Abhirup Bandyopadhyay, Institute De Neurosciences Des Systems, Aix-Marseille University, France for his constant motivation and helpful discussions during the work.

Conflict of interest statement.

None declared.

Funding

S.B, S.A and F.P. are or were Postdoctoral researcher at the Belgian F.R.S.-FNRS Fund for Scientific Research. M.R. is Research Director at the same fund.

Data Availability Statement

The list of all the proteins as well as the membrane dependent statistical potentials used in this study can be found in our GitHub repository github.com/3BioCompBio/BRANEart.

References

[1] Andreas Engel and Hermann E Gaub. Structure and mechanics of membrane proteins. *Annu. Rev. Biochem.*, 77:127–148, 2008.

[2] Zoe Cournia, Toby W Allen, Ioan Andricioaei, Bruno Antonny, Daniel Baum, Grace Brannigan, Nicolaeviorel Buchete, Jason T Deckman, Lucie Delemotte, Coral Del Val, et al. Membrane protein structure, function, and dynamics: a perspective from experiments and theory. *The Journal of membrane biology*, 248(4):611–640, 2015.

[3] Mame Ndew Mbaye, Qingzhen Hou, Sankar Basu, Fabian Teheux, Fabrizio Pucci, and Marianne Rooman. A comprehensive computational study of amino acid interactions in membrane proteins. *Scientific reports*, 9(1):1–14, 2019.

[4] Julia Koehler Leman, Richard Bonneau, and Martin B Ulmschneider. Statistically derived asymmetric membrane potentials from α-helical and β-barrel membrane proteins. *Scientific reports*, 8(1):1–14, 2018.

[5] Wei Lu, Nicholas P Schafer, and Peter G Wolynes. Energy landscape underlying spontaneous insertion and folding of an alpha-helical transmembrane protein into a bilayer. *Nature communications*, 9(1):1–10, 2018.

[6] Guillaume Postic, Yassine Ghouzam, and Jean-Christophe Gelly. An empirical energy function for structural assessment of protein transmembrane domains. *Biochimie*, 115:155–161, 2015.

[7] Rebecca F Alford, Julia Koehler Leman, Brian D Weitzner, Amanda M Duran, Drew C Tilley, Assaf Elazar, and Jeffrey J Gray. An integrated framework advancing membrane protein modeling and design. *PLoS Comput Biol*, 11(9):e1004398, 2015.
[8] Mikhail A Lomize, Irina D Pogozheva, Hyeon Joo, Henry I Mosberg, and Andrei L Lomize. Opm database and ppm web server: resources for positioning of proteins in membranes. *Nucleic acids research*, 40(D1):D370–D376, 2012.

[9] Kentaro Shimizu, Wei Cao, Gull Saad, Michiru Shoji, and Tohru Terada. Comparative analysis of membrane protein structure databases. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1860(5):1077–1091, 2018.

[10] J. M. Kwasigroch and M. Rooman. Prelude and fugue, predicting local protein structure, early folding regions and structural weaknesses. *Bioinformatics*, 22(14):1800–1802, 2006.

[11] M. De Laet, D. Gilis, and M. Rooman. Stability strengths and weaknesses in protein structures detected by statistical potentials: Application to bovine seminal ribonuclease. *Proteins*, 84(1):143–158, 2016.

[12] Qingzhen Hou, Fabrizio Pucci, François Ancien, Jean Marc Kwasigroch, Raphaël Bourgeas, and Marianne Rooman. Swotein: A structure-based approach to predict stability strengths and weaknesses of proteins. *Bioinformatics*, btab034(Epub ahead of print), 2021.

[13] Diego U Ferreiro, Elizabeth A Komives, and Peter G Wolynes. Frustration in biomolecules. *Quarterly reviews of biophysics*, 47(4):285–363, 2014.

[14] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[15] Guoli Wang and Roland L Dunbrack Jr. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

[16] Mikhail A Lomize, Irina D Pogozheva, Hyeon Joo, Henry I Mosberg, and Andrei L Lomize. Opm database and ppm web server: resources for positioning of proteins in membranes. *Nucleic acids research*, 40(D1):D370–D376, 2012.

[17] Evgeny Krissinel and Kim Henrick. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372(3):774–797, 2007.

[18] Manfred J Sippl. Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*, 213(4):859–883, 1990.

[19] Jean-Pierre A Kocher, Marianne J Rooman, and Shoshana J Wodak. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *Journal of Molecular Biology*, 235(5):1598–1613, 1994.

[20] Yves Dehouck, Dimitri Gilis, and Marianne Rooman. A new generation of statistical potentials for proteins. *Biophysical journal*, 90(11):4010–4017, 2006.

[21] B. Folch, Y. Dehouck, and M. Rooman. Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials. *Biophysical Journal*, 98(4):667–677, 2010.
[22] Fabrizio Pucci, Malik Dhanani, Yves Dehouck, and Marianne Rooman. Protein thermostability prediction within homologous families using temperature-dependent statistical potentials. *PloS one*, 9(3):e91659, 2014.

[23] Q. Hou, R. Bourgeas, F. Pucci, and M. Rooman. Computational analysis of the amino acid interactions that promote or decrease protein solubility. *Scientific reports*, 8(1):14661, 2018.

[24] Marianne J Rooman, Jean-Pierre A Kocher, and Shoshana J Wodak. Prediction of protein backbone conformation based on seven structure assignments: influence of local interactions. *Journal of Molecular Biology*, 221(3):961–979, 1991.

[25] J.M. Berg, J.L. Tymoczko, and L Stryer. *Biochemistry*, chapter Secondary Structure: Polypeptide Chains Can Fold Into Regular Structures Such as the Alpha Helix, the Beta Sheet, and Turns and Loops. W.H. Freeman, 5th edition, 2002.

[26] Wouter G Touw and Gert Vriend. On the complexity of engh and huber refinement restraints: the angle $\tau$ as example. *Acta Crystallographica Section D: Biological Crystallography*, 66(12):1341–1350, 2010.

[27] Sankar Basu, Dhananjay Bhattacharyya, and Rahul Banerjee. Applications of complementarity plot in error detection and structure validation of proteins. *Indian Journal of Biochemistry and Biophysics*, 51(June):188–200, 2014.

[28] Guillaume Postic, Yassine Ghouzam, and Jean-Christophe Gelly. Orempro web server: orientation and assessment of atomistic and coarse-grained structures of membrane proteins. *Bioinformatics*, 32(16):2548–2550, 2016.

[29] Andrei L Lomize and Irina D Pogozheva. Solvation models and computational prediction of orientations of peptides and proteins in membranes. In *Membrane Proteins*, pages 125–142. Springer, 2013.

[30] Lucas MP Chataigner, Nadia Leloup, and Bert JC Janssen. Structural perspectives on extracellular recognition and conformational changes of several type-i transmembrane receptors. *Frontiers in Molecular Biosciences*, 7, 2020.

[31] Naomi R Latorraca, AJ Venkatakishnan, and Ron O Dror. Gpcr dynamics: structures in motion. *Chemical reviews*, 117(1):139–155, 2017.

[32] Aravind Pennatsa and Eric Gouaux. How leht shapes our understanding of the mechanisms of sodium-coupled neurotransmitter transporters. *The Journal of physiology*, 592(5):863–869, 2014.

[33] Harini Krishnamurthy and Eric Gouaux. X-ray structures of leut in substrate-free outward-open and apo inward-open states. *Nature*, 481(7382):469–474, 2012.

[34] Kimberly L Koss and Evangelia G Kranias. Phospholamban: a prominent regulator of myocardial contractility. *Circulation research*, 79(6):1059–1063, 1996.

[35] Marco Mravic, Jessica L Thomaston, Maxwell Tucker, Paige E Solomon, Lijun Liu, and William F DeGrado. Packing of apolar side chains enables accurate design of highly stable membrane proteins. *Science*, 363(6434):1418–1423, 2019.

[36] Kirill Oxenoid and James J Chou. The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *Proceedings of the National Academy of Sciences*, 102(31):10870–10875, 2005.
[37] Taehoon Kim, Jinhyuk Lee, and Wonpil Im. Molecular dynamics studies on structure and dynamics of phospholamban monomer and pentamer in membranes. *Proteins: Structure, Function, and Bioinformatics*, 76(1):86–98, 2009.

[38] J Fujii, K Maruyama, M Tada, and DH MacLennan. Expression and site-specific mutagenesis of phospholamban: studies of residues involved in phosphorylation and pentamer formation. *Journal of Biological Chemistry*, 264(22):12950–12955, 1989.

[39] Heather KB Simmerman, Yvonne M Kobayashi, Joseph M Autry, and Larry R Jones. A leucine zipper stabilizes the pentameric membrane domain of phospholamban and forms a coiled-coil pore structure. *Journal of Biological Chemistry*, 271(10):5941–5946, 1996.

[40] Avram M Slovic, Steven E Stayrook, Benjamin North, and William F DeGrado. X-ray structure of a water-soluble analog of the membrane protein phospholamban: sequence determinants defining the topology of tetrameric and pentameric coiled coils. *Journal of molecular biology*, 348(3):777–787, 2005.

[41] Razvan L Cornea, Larry R Jones, Joseph M Autry, and David D Thomas. Mutation and phosphorylation change the oligomeric structure of phospholamban in lipid bilayers. *Biochemistry*, 36(10):2960–2967, 1997.