Amino Acids (2015) 47:2303–2322
DOI 10.1007/s00726-015-2010-5

ORIGINAL ARTICLE

Sequence–structure relationship study in all-α transmembrane proteins using an unsupervised learning approach

Jérémy Esque[1,2,3,4,5,6,7,8] · Aurélie Urbain[9] · Catherine Etchebest[1,2,3,4] · Alexandre G. de Brevern[1,2,3,4]

Received: 31 October 2014 / Accepted: 15 May 2015 / Published online: 5 June 2015

Abstract Transmembrane proteins (TMPs) are major drug targets, but the knowledge of their precise topology structure remains highly limited compared with globular proteins. In spite of the difficulties in obtaining their structures, an important effort has been made these last years to increase their number from an experimental and computational point of view. In view of this emerging challenge, the development of computational methods to extract knowledge from these data is crucial for the better understanding of their functions and in improving the quality of structural models. Here, we revisit an efficient unsupervised learning procedure, called Hybrid Protein Model (HPM), which is applied to the analysis of transmembrane proteins belonging to the all-α structural class. HPM method is an original classification procedure that efficiently combines sequence and structure learning. The procedure was initially applied to the analysis of globular proteins. In the present case, HPM classifies a set of overlapping protein fragments, extracted from a non-redundant databank of TMP 3D structure. After fine-tuning of the learning parameters, the optimal classification results in 65 clusters. They represent at best similar relationships between sequence and local structure properties of TMPs. Interestingly, HPM distinguishes among the resulting clusters two helical regions with distinct hydrophobic patterns. This underlines the complexity of the topology of these proteins. The HPM classification enlightens unusual relationship between amino acids in TMP fragments, which can be useful to elaborate new amino acids substitution matrices. Finally, two challenging applications are described: the first one aims at annotating protein functions (channel or not), the second one intends to assess the quality of the structures (X-ray or models) via a new scoring function deduced from the HPM classification.

Handling Editor: L. Taher.

J. Esque and A. Urbain the first two authors should be regarded as joint first authors.

C. Etchebest and A. G. de Brevern the last two authors should be regarded as joint last authors.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-015-2010-5) contains supplementary material, which is available to authorized users.

✉ Alexandre G. de Brevern
alexandre.debrevern@univ-paris-diderot.fr

[1] INSERM, U 1134, DSIMB, 75739 Paris, France

[2] Univ. Paris Diderot, Sorbonne Paris Cité UMR-S 1134, 75739 Paris, France

[3] Institut National de la Transfusion Sanguine (INTS), 75739 Paris, France

[4] Laboratoire d'Excellence GR-Ex, 75739 Paris, France

[5] Laboratoire d'Ingénierie des Fonctions Moléculaire (IFM), ISIS, UMR 7006, 67000 Strasbourg, France

[6] Department of Integrative Structural Biology, INSERM U964, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), 67404 Illkirch, France

[7] UMR7104, Centre National de la Recherche Scientifique (CNRS), 67404 Illkirch, France

[8] Université de Strasbourg, 67404 Illkirch, France

[9] Institut Jean-Pierre Bourgin, INRA, UMR 1318, 78026 Versailles, France

Keywords Transmembrane protein · Learning approach · Sequence–structure relationship · Protein structure · Artificial neural network · Hybrid protein model · Structural alphabet · Classification

Introduction

Membrane proteins (MP) are estimated to cover 25–30 % of the whole set of proteins (Liu et al. 2002; Nugent and Jones 2009; Wallin and von Heijne 1998; Sutormin et al. 2003). Thus, they play a crucial role in the cells in performing many complex physiological functions, such as signalling, transport through the membrane with ions (Gamper and Shapiro 2007), metabolites (Visser et al. 2007), peptides or RNA (Cohen 2005), energy generation, regulating intracellular vesicular transport, controlling membrane lipid composition (Szalontai 2009) and maintaining of structural architecture of cells (Burgess et al. 1994).

Their role in many crucial physiological functions also makes them important drug targets, accounting about 60–70 % of the studied drug targets (Arinaminpathy et al. 2009; Yildirim et al. 2007; Giacomini et al. 2010). Therefore, studying the membrane protein structures and their function stays a crucial topic in chemistry, biology and computational sciences (Arinaminpathy et al. 2009; Elofsson and von Heijne 2007; Nam et al. 2009; Nugent and Jones 2012; von Heijne 2011).

Expression, stabilization in a near-native environment, or crystallization of transmembrane proteins (TMP) (Pieper et al. 2013) are technical issues that limit the number of solved structures (~1–2 %) by comparison with globular proteins (White 2009). Hence, computational approaches can be valuable tools for predicting membrane protein structures and understanding their function (Nam et al. 2009; Nugent and Jones 2012). Most methods dedicated to TMP structure prediction start with the detection of transmembrane segments. The state-of-the-art methods are based on Hidden Markov models (HMM), neural networks (NN), support vector machine (SVM) and recently weighted-random forests. The prediction rate reaches on average ~70 %. From this starting point, methods have been developed to predict the protein topology. This prediction can be a useful prelude to identify a 3D structure using fold recognition approaches. Indeed, the popular homology techniques are limited by the low number of templates and the difficulty to obtain accurate sequence alignments between the template and targets, as exemplified in different studies. Beside the alignment algorithm itself, a key element is the amino acid substitution matrix (SM). The two most popular SM, PAM (Dayhoff and Schwartz 1978) and BLOSUM (Henikoff and Henikoff 1992), were initially deduced from comparison of sequences of soluble proteins. Then, specific matrices were developed for membrane proteins, e.g. JTT transmembrane (Jones et al. 1994), Persson–Argos (Persson and Argos 1994), PHAT 75/73 (Ng et al. 2000) and very recently Membrane Fugue (Hill et al. 2011) to take better account for the different environments experienced by a transmembrane protein, from highly hydrophilic for the extramembranous regions to highly hydrophobic for the transmembrane domains. Hence, some authors combined in a bipartite scheme, different substitution matrices adapted to each environment (Forrest et al. 2006; Pirovano et al. 2008; Shafrir and Guy 2004; Sutormin et al. 2003). In most cases, some improvements were achieved in sequence alignments. Recently, Deane and co-workers (Kelm et al. 2010) went a step further by proposing different substitution matrices specific for different locations along the transmembrane segments. Significant differences with soluble proteins were emphasized that were mainly attributed to changes in hydrophobicity and also in secondary structures. To summarize, all these studies underline the importance of the secondary structures (Stamm et al. 2013) and the specific usage of amino acids in membrane proteins compared to soluble proteins.

From a structural point of view, the folds of TM proteins are generally separated into two classes, the β-barrel and the all-α class, the latest one being the most abundant one and the most studied. In the all-α class, the proteins fold as α-helices bundles crossing the membrane (Fuchs et al. 2009; Lo et al. 2009; Nugent and Jones 2010; Wang et al. 2011), the structural stability being governed by tight and specific interactions between residues (Marsico et al. 2010a, b; Walters and DeGrado 2006; Nagarathnam et al. 2011). Besides secondary structures, local structures (e.g. structural motifs) have also been identified. For example, re-entrant regions, which are non-helical segments located at least within one leaflet of the membrane, have been shown to play important structural or functional roles (Viklund et al. 2006). They are found for instance in the pore region of some channels. Kinks are other important local distortions motifs that impact the interactions with neighbouring helices or lipids (Hall et al. 2009; Langelaan et al. 2010; Meruelo et al. 2011; Yohannan et al. 2004a). To summarize, the 3D structures available nowadays show a larger local structural diversity than was initially suspected.

Hence, in the present article, we have performed a specific and accurate insight of these proteins, with the aim to better understand them and develop tools for predicting or analysing (Hill et al. 2011). We focused on the all-α class to have sufficient data for deciphering relevant rules. We have analysed the sequence–structure relationship and examined the two aspects mentioned above, i.e. the local structures and the amino acids use. To do it, we have taken advantage of a combined sequence–structure learning provided by an original and reliable unsupervised learning

approach, called Hybrid Protein Model (HPM) (de Brevern and Hazout 2000, 2003; Benros et al. 2006; Bornot et al. 2009). This method has shown its efficiency and usefulness in globular proteins, for elaborating a structural alphabet i.e. a library of peptide fragments (de Brevern et al. 2000, 2002; Joseph et al. 2011), finding similar folds, structural alignments (de Brevern and Hazout 2001) or the prediction of flexibility (Bornot et al. 2011; de Brevern et al. 2012). Its originality relies on the capacity to combine and compact physico-chemical properties and structural information.

First, we built a set of fragments obtained by dividing a non-redundant databank of TM structures into overlapping pieces of 5-residues length. The fragments were described by structural and physico-chemical properties. The learning process resulted in 65 clusters of peptide fragments, representing at best the combination of local sequence and structure information of TMPs. Interestingly, the method identified distinct helical patterns, and also local motifs related to the positioning with respect to the membrane. Importantly, the method has brought new amino acid associations, which enables to define a new substitution matrix dedicated to membrane proteins comparison. Finally, the interest of the method is exemplified by two challenging applications: (1) its capacity to distinguish channel from non-channel proteins; (2) its potentiality to evaluate the quality of structures (X-ray or models). Promising results were obtained, which opens the way to help for modelling membrane proteins using general modelling software [I-TASSER (Roy et al. 2010), Modeler (Sali and Blundell 1993; Eswar et al. 2006)] or dedicated software [Medeller (Kelm et al. 2010)].

Materials and methods

Data set

The non-redundant databank of all-α TMP structures was obtained using various dedicated databases, e.g. OMP (Lomize et al. 2006), TMPDB (Ikeda et al. 2003), PDBTM (Tusnady et al. 2004, 2005), and Nugent's data set (Nugent et al. 2011). Protein structures with X-ray resolution better than 2.5 Å were selected. Sequence redundancy was eliminated using firstly Cd-hit (Li and Godzik 2006) and then CLUSTALW (Thompson et al. 1994). In the final dataset, the sequences share less than 50 % sequence identity. Representative proteins were selected based on (1) best resolution, (2) no missing residues in the transmembrane regions, and (3) no mutations in the transmembrane regions. In case of alternate positions, atoms with occupancy values upper than 0.5 were selected. For occupancy factors equal to 0.5, the first atom was selected. Selenium atoms from selenocysteine (Sec) and selenomethionine (Sem) were replaced by sulphur atoms and the residue types were modified accordingly, i.e. cysteine (Cys) and methionine (Met), respectively. The final databank contained 52 PDB files (Berman et al. 2000) corresponding to 65 protein chains and 15,992 residues (see table S1). We also checked the influence of the databank by considering two other datasets, one with fewer proteins and a new one containing more recent structures. We calculated the Pearson's Correlation Coefficient between the *Z*-scores amino acid and protein blocks obtained with our databank and the two other datasets. The high PCC obtained in both cases, i.e. 0.9 and 0.96 with the smaller databank and the newer one, respectively, demonstrates that the results described in the following are solid.

Protein blocks

Protein blocks (PBs) correspond to a library of small fragments able to approximate the local protein backbone (de Brevern et al. 2000). PBs are widely used for the study of globular proteins (Joseph et al. 2011). This structural alphabet comprises 16 small structural fragments of 5-residue long labelled by letters from *a*–*p*. These fragments are encoded in (φ, ψ) vectors. PBs *a*–*f* describe mainly β-sheets, with *d* corresponding to the more regular central part. PBs *d* are flanked on N-caps by PBs *a*–*c* and C-caps by PBs *e* and *f*. PBs *k*–*p* mainly correspond to α-helix, with PB *m* associated with the central part of a right helix. PBs *g*–*j* are associated with coil structures (de Brevern 2005).

Definition of the Hybrid Protein Model (HPM)

Hybrid Protein Model is an unsupervised learning approach able to compact simultaneously protein sequence and structure information (de Brevern and Hazout 2000, 2003). The principle is similar to Self-Organizing Maps (Kohonen 2013) and consists in optimally learning a set of fragments from a protein structure databank. In this study, 14,649 fragments (noted *f*) of 5-residue length were learnt using HPM. Each fragment *f* is described by a vector *v* with 31 components, i.e. 15 values coding sequence properties (3 sequence properties × 5 amino acids) and 16 values coding structure properties. Sequence properties considered are polarity, volume and hydrophobicity (see Table S2). The associated values are obtained from dedicated scales (Grantham 1974; Zamyatnin 1984; Eisenberg et al. 1984) and were normalized between −1 and 1. The 16 structural properties correspond to cosine and sine functions of the 8 (φ, ψ) dihedral angles. It must be noticed that, for a given protein, fragments are overlapping, and consequently, two successive fragments *f* have 4 residues in common, e.g. residues 2–5 for fragment f_i and residues 1–4 for fragment f_{i+1}.

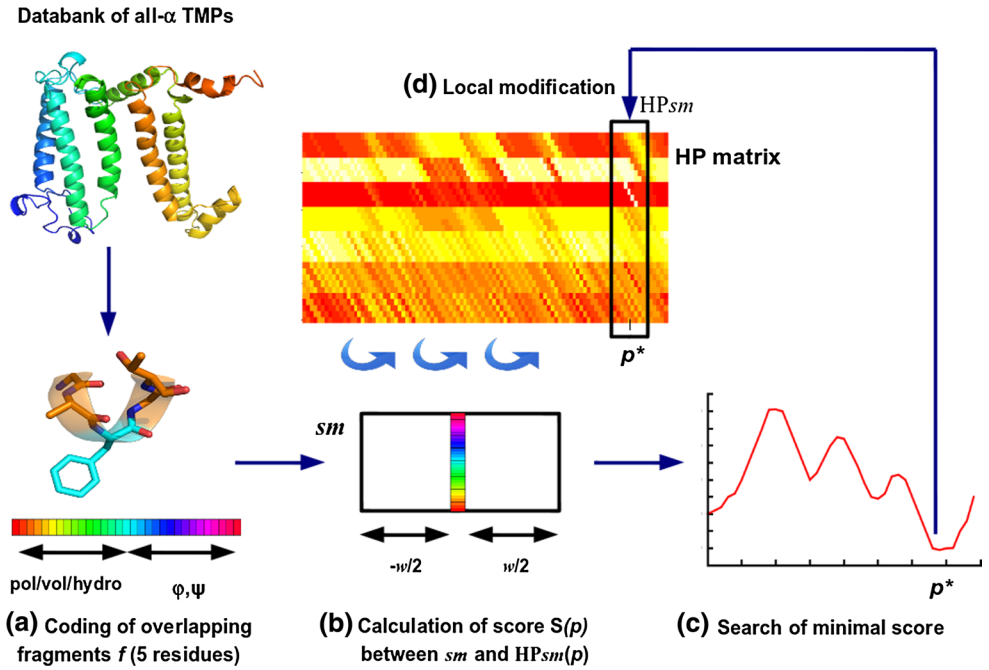


Fig. 1 Learning approach HPM to study the sequence–structure relationship in transmembrane proteins. **a** From the non-redundant databank, overlapping fragments of five-residue length are generated. These fragments are coded in a vector of 31 components (5 for polarity, 5 for volume, 5 for hydrophobicity, 8 for cosine and sine of φ angles, 8 for cosine and sine of ψ angles). **b** A matrix named Hybrid Protein (HP) is built from fragments chosen randomly. Then, all fragments are learnt along with their residue environment (window size, $W = 13$). For each sub-matrix (*sm*), a Euclidean distance score is computed at all positions of the matrix. **c** The position p with the minimal score is identified, namely p^*. **d** The position p^* and its local surroundings are modified to learn the presented example, i.e. it reinforces the agreement with *sm*

In the present study, Hybrid Protein (HP) matrix is a matrix of dimensions $L \times m$, where L is the number of classes and m correspond to 31 characteristics described previously. Each class contains a family of fragments *f*. The learning step of all fragments f is similar to a Kohonen's Self-Organizing Map or SOM (Kohonen 2013), except that the diffusion is implicitly included as the fragments f are overlapping and their environment $[f - w; f + w]$ is taken into account during the learning (here, $w = 6$).

(i) Initialization; a HP matrix of dimension $L \times m$ is randomly initialized in choosing L fragments f coded by their vectors v in the databank.

HPM relies on 2 main steps (see Fig. 1):

(ii) Learning; the process is iterative: (1) one fragment f with its environment ($\pm w$) is randomly selected from the databank. It is associated with a sub-matrix *sm* of dimension $W \times v$, where W equals to $(w \times 2) + 1$, and v corresponds to vectors of 31 components (see Fig. 1a). (2) For every position p of the HP, the Euclidean distance $S(p)$ (for Score p), is calculated between *sm* and a sub-matrix HP*sm* of the same dimension, the central position corresponding to p (see Fig. 1b). A score profile is then established along HP. (3) The minimal score *Smin* is associated with the maximal similarity between both sub-matrices. The corresponding position in HP is noted p^* (see Fig. 1c). (4) To improve agreement with *sm*, the sub-matrix HP*sm*(p^*) is then slightly modified according to Eq. 1:

$$\mathrm{HP}sm(p^*) = \mathrm{HP}sm(p^*) + \left(sm - \mathrm{HP}sm(p^*)\right).\ \alpha(n) \quad (1)$$

$$\alpha(n) = \frac{\alpha_0}{1 + \frac{n}{N}}, \quad (2)$$

where n is the number of sub-matrices *sm* already seen by the HP, N is the total number of sub-matrices in databank and α_0 is the initial learning coefficient (see Fig. 1d). Convergence of HPM is ensured by progressively decreasing the learning coefficient $\alpha(n)$ during the training (see Eq. 2). This process is iterated C times for all fragments *f*. Note that HP is circular, e.g. the last position L is contiguous with the first position.

Calibration for finding the optimal HPM

As detailed above, the HPM mainly depends on four major parameters, which are coupled. First, the learning process efficiency requires a good balance between the learning coefficient α_0 and the number of cycles C. A too small value for α_0 would prevent from bypassing a poor initialization, whereas a high value for α_0 would require high C value to reach convergence.

Second, the size of the learning window W and the number of classes L will also largely influence the output of the procedure. These two parameters mainly depend on the nature of data and need to be calibrated accordingly. In the present work, the choice of W was fixed and based on recommendations drawn from previous studies (de Brevern and Hazout 2000, 2003; de Brevern et al. 2000) and values used in the literature, namely W equals 13. This window size, which corresponds to 17 residues, allows taking into account the most hydrophobic part of transmembrane helices, which may range from 16 to 42 residues as observed by (Papaloukas et al. 2008) and our study on OMP (Lomize et al. 2006). The number of L classes is also crucial and will strongly impact the final description of learnt patterns. It results from a fine balance between a detailed description, i.e. a large value for L, and enough data to avoid bias, i.e. a minimal number of fragments for each class.

Last, overlapping is a main feature of HPM learning approach; it is so important to check that consecutive fragments are often found in successive classes. Hence, we used two additional measures, defined by Eqs. 3 and 4, respectively, to control the efficiency of the classification. Given the transition $T_{i,k}$ between positions i and k and the continuity CO_i in position i, which is defined as to $T_{i,j+1}$

$$T_{i,k} = \frac{N_{i,k}}{N_i} \tag{3}$$

$$\mathrm{CO}_i = T_{i,j+1} = \frac{N_{i,j+1}}{N_i} \tag{4}$$

with $N_{i,k}$ the number of fragments f_i found at the ith position in the HP with following fragment f_{i+1} found at the kth position in HP, while N_i is the total number of fragment f_i found at the ith position. Note that the overlapping of clusters also hampers the initialization step of HPM.

For defining the optimal HP, we tested several values for the different parameters mentioned above, i.e. L (from 50 to 100), α_0 (from 0.01 to 0.5) and C (from 20 to 100) values (data not shown). 100 independent simulations were done for each length L with different parameter values. The most representative HP was obtained in two steps: (1) from the 100 independent HP simulated (100 random initializations), the distance between all pairs of HP matrices was computed and the HP associated with the smallest distance to all the others was selected, i.e. the centroid. The optimal learning parameters for the first step correspond to $\alpha_0 = 0.35$ with $C = 30$ cycles. (2) This HP was used as a new initialization matrix for 100 new trainings with lower learning coefficient α_0 (0.05), and C equals to 25 cycles. As previously, the HP with the smallest distance to the others amongst the 100 simulations was finally selected as the most representative one.

Implementation and statistical analyses

HPM program was entirely coded in C language based on previous program (de Brevern and Hazout 2000). PDB analysis was done with separate software also written in C (de Brevern 2005). Helix geometry analyses were done with HELANAL software (Bansal et al. 2000). R software, version 2.10.1 (http://cran.r-project.org/) was used for some statistical analyses and figures. Hierarchical clustering was performed using Ward algorithm implemented in R software (*hclust*).

Sequence and structure specificities were obtained by computing Z-scores that quantify over- and under-representation of amino acids or PBs (de Brevern et al. 2000):

$$Z - \mathrm{Score}(i,j) = \frac{n_{ij}^{\mathrm{obs}} - n_{ij}^{\mathrm{theo}}}{\sqrt{n_{ij}^{\mathrm{theo}}}} \tag{5}$$

$$n_{ij}^{\mathrm{theo}} = N_i.f_j, \tag{6}$$

where n_{ij}^{obs} is the observed number of jth amino acid ($j = 20$) or PB ($j = 16$) in the position i. N_i is the total number of amino acids (respectively, PBs) in position i and f_j is the frequency of amino acid j (respectively, PB) in the databank.

Kullback–Leibler asymmetric divergence measure *(KLd)* relative entropy gives a precise estimation of the informativity of a given position (Kullback and Leibler 1951; de Brevern et al. 2000):

$$KLd_i(p,q) = \sum p_i \ln\left(\frac{p_i}{q_i}\right). \tag{7}$$

This value quantifies the contrast between the observed amino acid (or PB) frequencies **p**: $\{p_i\}$ 1,…, 20; or PB frequencies **p**: $\{p_i\}$1,…, 16 and a reference probabilistic distribution **q**$\{q_i\}$ (de Brevern et al. 2000).

Amino acid equivalences and substitution matrices

Sequence specificities evidenced for each HPM position were used to identify equivalent residues.

As HP* matrix represents the optimal clustering of fragments encoded in terms of sequence and structural

properties, it is possible to identify for each HP* position the corresponding amino acid distribution. This distribution was defined as a *Z*-score (*j*, *i*) for amino acid *i* in position *j* [see Eqs. (5) and (6)]. So, the characteristics of each amino acid *i* are stored in a vector v_i of 65 length, whose each component *j* contains *Z*-score information for each position *j* of HP*. The Euclidian distance between two amino acids *i* and *k*, is defined as:

$$D_{i,k}=\sqrt{\sum_{j=1,65}(Z-\text{score}(j,i)-Z-\text{score}(j,k))^2}. \quad (8)$$

Groups were based on this Euclidean distance $D_{i,k}$. The grouping was computed using the pvclust (Suzuki and Shimodaira 2006) routine available in R package that consists in a hierarchical clustering using the complete method. The procedure allows assessing the reliability of the grouping by computing an Approximately Unbiased (AU) *p* value, resulting from multiscale bootstrap resampling. The method was also applied for different substitutions matrices (SMs) widely used for sequence alignment or mining purposes. In this case, the distance between amino acid *i* and *j*, $D_{i,j}$, is the Euclidean distance between two vectors of 20 length and containing substitution values. Different substitution matrices were considered, specific of transmembrane sequences, i.e. PHDhtm 80 (Ng et al. 2000), Persson–Argos 80 (Ng et al. 2000) and PHAT 75/73 (Ng et al. 2000) (noted TM-SM) or not specific as BLOSUM 62 matrix (Henikoff and Henikoff 1992). Indeed, BLOSUM was shown to be as efficient as PHAT matrix for TM pair-wise sequences alignment (Forrest et al. 2006) and to achieve good multiple sequence alignments for membrane proteins (MP) when used in a bipartite scheme with PHAT (Pirovano et al. 2008). Statistical analyses were performed to compare these matrices with HP*-SM, by computing correlation coefficients with Spearman or Pearson methods, noted SCC and PCC, respectively.

HPM-substitution matrix (HPM-SM) for TM sequence alignment

The HPM-substitution matrix had been used to perform TM sequences alignment. Please notice, that HP*-SM values from *Z*-scores were shifted to be in the same range as Blosum62 (HPM-SM is given in Table S4 in SI).

For an ease comparison of HPM matrix to Blosum62 matrix on alignment, we perform a scaling on both variance and average of values our matrix to scale those of Blosum62. For this purpose, we computed *Z*-scores that were further translated and scaled to reach a similar mean and standard deviation of Blosum62 scores. Finally, the matrix was symmetrised such as $S_{ij}^* = (S_{ij} + S_{ji})/2$, where S_{ij} represents the substitution cost of amino acid *i* by *j*.

Besides HPM-SM values, gap opening (named *gapopen*) and gap extension (named *gapextend*) parameters are required to perform sequence alignments. An extensive test was conducted to optimize these two parameters. Gapopen and gapextend were varied in the range (4.0–25.0) and (0.1–8.0), respectively, by steps of 0.5. All combinations of pairs were tested. Muscle 3.6 (Edgar 2004a) was chosen for performing the multiple sequence alignments as it was shown efficient for aligning membrane sequences (Stamm et al. 2014) and offers the opportunity to input a substitution matrix.

Two datasets were used to assess the performance of HPM-SM: (1) sequence alignments available in reference 7 of Balibase dataset (denoted as Bali in the following) (Thompson et al. 2005); (2) Homep2 dataset (homep2) developed by Forrest's group and detailed in (Stamm et al. 2013). For Bali, we calculated multiple sequence alignments for each of the 8 families: 7tm, acr, dtd, ion, msl, Na, photo and ptga. For Homep2, as it corresponds to sequences with 3D known structures, we first performed pairwise structural alignment with TM-align software (Zhang and Skolnick 2005) and compared the resulting alignment of pairs of sequences to those obtained with HPM-SM. The pairwise alignment analyses led to 177 alignments based on 81 sequences clustered in 22 groups [see SI of ref. (Stamm et al. 2013)]. The quality of the alignments was compared to their reference using Qscore tool (Edgar 2004b) that provides Q-score (identical to Bali score), TC score, Cline score and Modeler score (Sauder et al. 2000; Tress et al. 2003; Cline et al. 2002; Thompson et al. 1999). HPM-SM performance was finally evaluated in comparison to that of Blosum62 matrix used by default in Muscle 3.6. For each pair of gap values, we calculated the ranking for each type of scores, and they were compared with the results with Blosum62 matrix.

Results

The original unsupervised learning procedure used in the present work allowed classifying simultaneously structure and sequence properties of fragments in similar clusters, while taking into account overlapping between these fragments. Besides the learning parameters, which were checked carefully, the number of classes (or positions in HPM) was the key feature that strongly influenced the identification of relevant patterns in α-helical transmembrane proteins.

The result of the procedure is a matrix in which the number of rows represents the number of features considered (sequence–structure), whereas the columns correspond to clusters or classes. The labels of the clusters are important due to the existence of transition properties. Hence, to

underline this relationship we will use preferentially the term "positions" for clusters in the following.

We detail in the following sections the main characteristics of the positions obtained with the optimal HP*. Then, we propose some alternative groupings and we examine the substitution properties of amino acids deduced from the present classification. Finally, we suggest two useful applications of HP*: (1) the annotation of protein functions by discriminating the channels and others; (2) the access to the quality of structural models of TMPs via global and local scoring functions. This last point is very important, as no dedicated tool still exists for ranking TMPs structural models.

Characteristics of the selected HP matrix (HP*)

The HP* matrix was defined as the most representative HP of all simulations. It was obtained after the two-step learning process described in "Materials and methods". A number of 65 positions ($L = 65$) in the HP* matrix allowed obtaining a good balance between the number of fragments, significant sequence signatures and fine structural characteristics in the clustering. Different lengths of HPM were tested, with L varying between 50 and 100 positions as done previously (de Brevern and Hazout 2003). The shortest HPM led to characterize only one kind of helix, whereas the longest HPMs led to positions with a too small number of occurrences to be relevant. For instance for $L = 75$, the positions with the minimal occurrence was of 15 and the percentage of continuity reached only 65 %. Hence, HP*, illustrated in Fig. 2a, is a matrix of dimension 65 × 31. For each position p ($p = 1, \ldots, 65$), the vector v contains the following properties from learnt fragments (5 residues): polarity (zone ranging from 1 to 5), volume (6–10), hydrophobicity (11–15), cosine/sine φ (16–23), and cosine/sine ψ (24–31). The small diagonal lines reflect shared properties between successive positions, mainly due to the overlapping before mentioned. However, the vector v in position p is not a simple shift of previous position ($p - 1$). For instance, comparing the positions 4 and 5 (x-axis), the value of 8th row (position 4) is −0.203 in the range [−0.4:−0.2]; whereas the 7th row (position 5 equals to −0.199). However, all differences in diagonals are smooth looking at the value scale, i.e. the values are continuous.

Analysis of this matrix showed in most positions, well-defined patterns representative of sequence and/or structural characteristics. For instance, positions 31–36 are associated with high hydrophobic properties (zone [11–15] along the y-axis) while positions 7–11 and 21–48 are associated with peculiar structural properties (zone 16–31 along the y-axis). In this latter, the cosine and sine ranged between [0.2:0.6] and [−1.0:−0.8] for φ angles and [0.6:0.8] and [−0.8:−0.4] for ψ angles, respectively. These values mostly correspond to helical conformation.

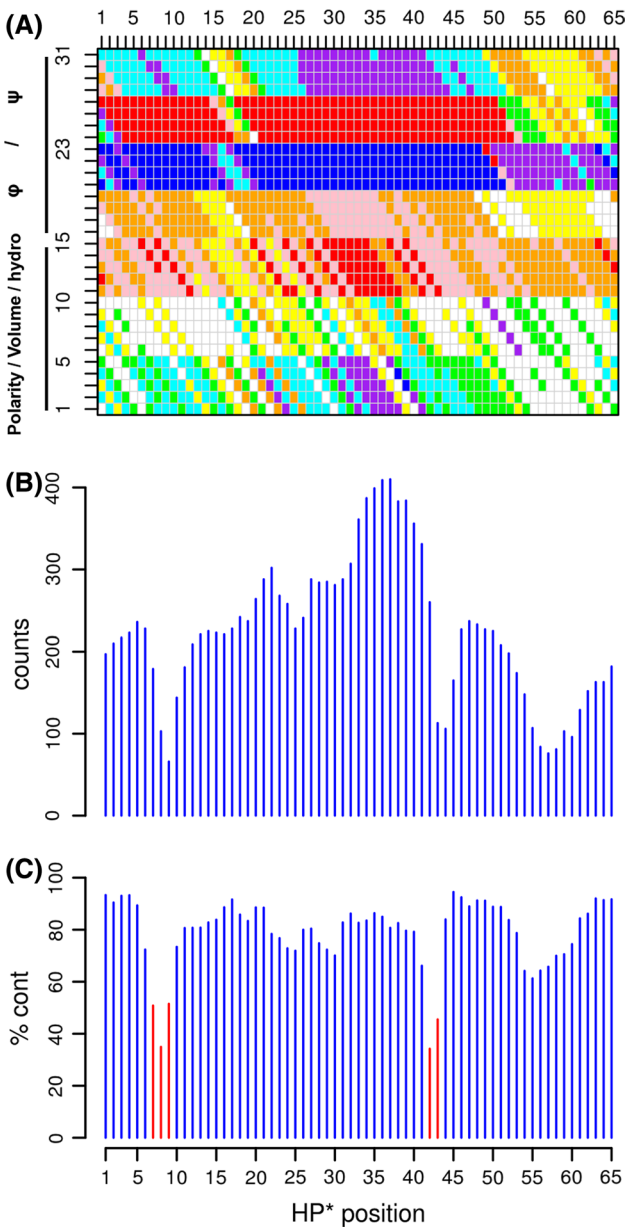


Fig. 2 Characteristics of the HP* matrix. **a** The HP* is composed of 65 classes (x-axis), characterized by vectors of length 31. y-axis (for fragments of 5 residues): polarity [zones 1–5], volume [zones 6–10], hydrophobicity [zones 11–15] and cosine/sine of dihedral angles φ [16–23] and ψ [24–31]. The colour scale ranges from −1 (*blue*) to 1 (*red*). **b** The histogram shows the distribution of protein fragments at each position of HP*. **c** The percentage of continuity (% cont) between fragments is plotted for each class of HP*. For a position p and given fragments f, % cont corresponds to percentage of fragments ($f + 1$) found in the position ($p + 1$). The only exception is the last position ($p = 65$), where the % cont is calculated using the first position (due to the periodicity of the matrix) (colour figure online)

As already mentioned, the number of fragments (see Fig. 2b) and the percentage of continuity CO_i (see Fig. 2c) are also major parameters for assessing HP learning, e.g. they were used to select the final length of HP*. During the

training, 14,649 fragments were learnt. Position 37 corresponds to the most populated cluster (410 fragments) while position 9 is the least populated one with only 66 fragments (see Fig. 2b). The average number of fragments per position equals 225 with a standard deviation of 87. The continuity value CO_i equals 79 % on average. This value, which can reach 94.5 % in position 45, demonstrates that the learning procedure has efficiently trapped the overlapping between successive fragments f and $f + 1$ in a given protein and mainly distributed these fragments in consecutive positions i and $i + 1$ in HP* (see Fig. 2c). For most positions, alternative transitions are only weakly populated with less than 5 % on average (see Figure S1). Only three zones located in positions 5–10, 42–43 and 54–60 in HP* exhibit more fuzzy patterns with a continuity value dropping to 34 % for positions 8 and 42. Nonetheless, the limited number of fragments in the corresponding positions prevented from identifying significant additional preferential transition (see Figure S1).

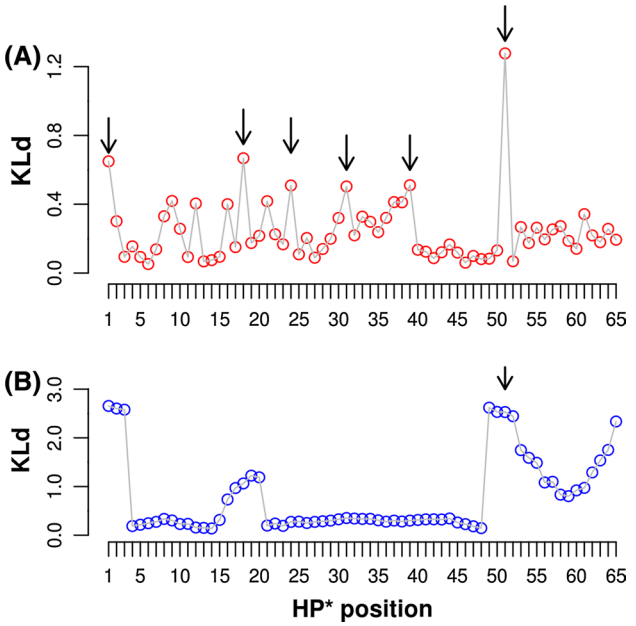


Fig. 3 Entropy and significance of positions in HP*. **a** The Kullback–Leibler asymmetric divergence measure (*KLd*) was calculated for each position. This measure quantifies the difference between the observed distribution of amino acids occurred at each position and the one occurred in the databank. The *blacks arrows* indicate the highest values (high significance): position 1 = 0.65, pos 18 = 0.67, pos 24 = 0.51, pos 31 = 0.50, pos 39 = 0.51 and pos 51 = 1.28. **b** The *KLd* value for each position was calculated in the same way as (**a**), but on PBs. The *black arrow* indicates the position 51 which has a very high value (2.53) in both analyses

Analysis of the learning: distribution of amino acids (AA) and protein blocks (PBs)

The sequence–structure relationship learned by HP* was analysed using protein blocks description and amino acid over and under representations. First, the information content of each position was examined by computing the Kullback–Leibler asymmetric divergence measure *(KLd)* relative to entropy for AA and PBs (see Fig. 3a, b, respectively). The two profiles are different with a large flat zone ranging from 20 to 45 for PBs. Yet, the most informative positions are found in similar locations 1, 18, 24, 31, 39 and 51 of HP* for amino acids and in 18–20, 49–54 for PBs. In both cases, the position 51 is the most informative position (see arrows on Fig. 3), i.e. associated with high *KLd* values.

Second, *Z*-scores were computed to underline in details the different positions (see Fig. 4). Considering structural data, PB *m,* mainly associated with α-helical conformation, is over-represented in two regions located between positions 21–48 (Type 1 helix) and 4–13 (Type 2 helix) (see blue lines of Fig. 4a). These two helical regions are well segregated in HP*, indicating that each of them comprises distinct features (see below). From a structural point of view, Type 1 helix corresponds to a straight helix, whereas Type 2 helix is characterized by large kink values (see Fig S4). Two non-helical regions are found in positions 16–20 and positions 52–3 (notice that position 1 is contiguous to position 65 in HP model). The fragments clustered in these non-helical regions correspond to loops connecting helices or elongated fragments in β-strands (e.g. PBs *c*, *d* or *e*).

Z-score values of amino acids (see Fig. 4b) allowed completing HP* features description: a strongly hydrophobic region (positions 28–36) with an over-representation of isoleucine (I), valine (V), leucine (L), phenylalanine (F) is located in the longest helical region of HP*, i.e. Type 1 helix. It corresponds to the hydrophobic core of the transmembrane helices. Polar and charged amino acids (N, Q, D, E, R and K) were not segregated in distinct zones, but are over-represented in positions 12, 16, 21 or 24. These positions are associated with helix extremities. Note that Type 2 helix does not show any amino acid preferences except position 12 characterized by an over-representation of charged residues and under-representation of hydrophobic residues.

HP* also highlights known features about proline and glycine residues. Proline is over-represented in non-helical regions or at ends of helical regions, i.e. positions 1, 2, 19, 52–54, 62 and 64. This result relates to its role of distortion inducer ("helix breaker") in soluble and in membrane helices as discussed in different studies (Cordes et al. 2002; Sansom and Weinstein 2000; Yohannan et al. 2004b). Interestingly, glycine is over-represented at the most informative positions 18 and 51 and also located inside a helical region (positions 30, 37 and 38). In contrast to what was observed in soluble proteins, glycine does not seem to introduce significant distortion in helical zones. The presence of glycine

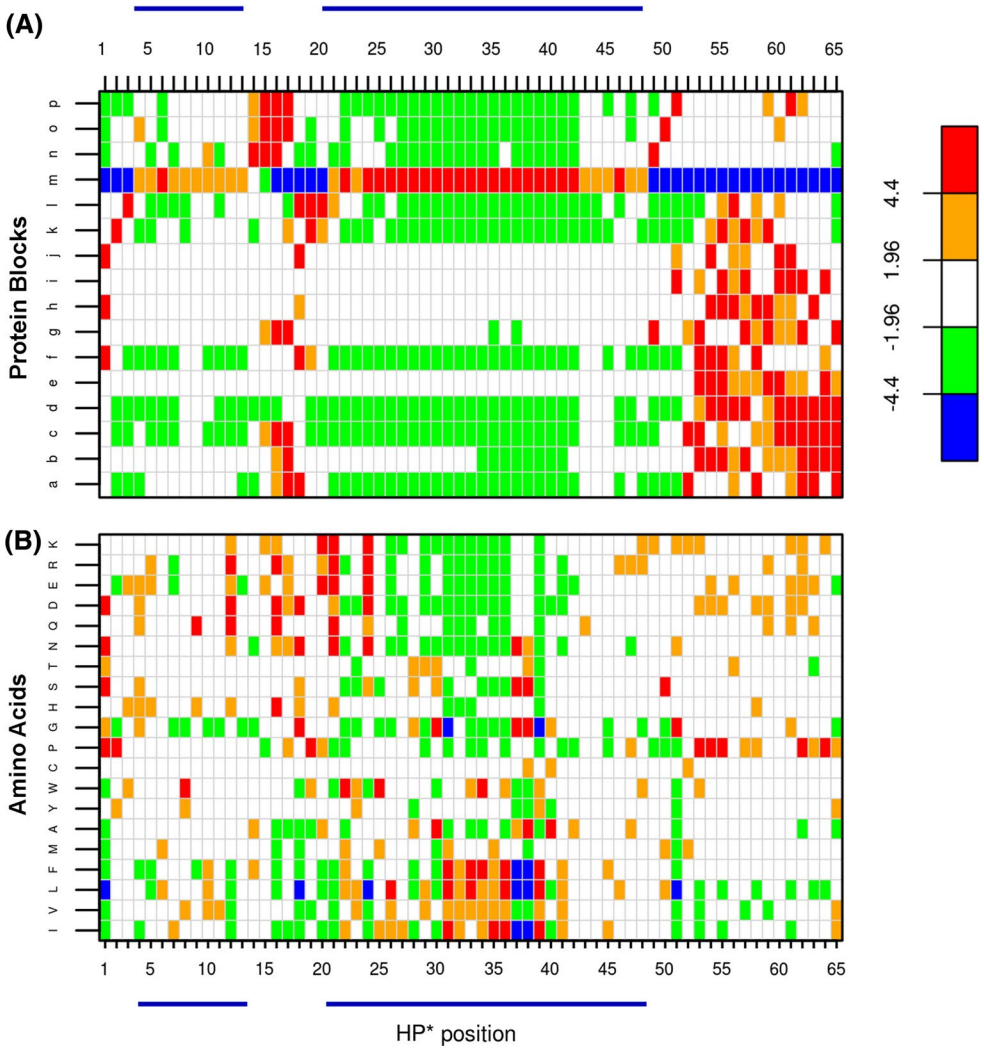


Fig. 4 Distribution of PBs and AA for each position of HP* matrix. The over/under-representation of PBs (**a**) and Amino Acids (**b**) was measured by the calculation of *Z*-score values. The over-representation is shown by the *red colour*, whereas the *blue colour* indicates an under-representation. The *blue lines* correspond to the delimitations of helical zones (over-representation of PB *m*). The amino acids are ranked by hydrophobic properties (**b**). The *Z*-score values follow the χ^2 law. Thus a *Z*-score value of 1.96 corresponds to $p = 0.05$ and a value of 4.4 corresponds to $p = 0.01$. The *colour gradient* is shown on the *right side* of the figure (colour figure online)

in helical regions is likely related to some pre-eminence of glycine in helix–helix interfaces in membrane proteins (Walters and DeGrado 2006). These last positions are also associated with an under-representation of hydrophobic residues I, V, F and L.

A simplified view of HP*

We then performed a hierarchical clustering of the 65 vectors of HP* to better highlight similarities between HP* positions. We focus here on the results obtained with 8 independent clusters that distinguish sequentially the groups along the HP* and emphasize the stacking learning of HPM (see Fig. 5a). Groups 1 (G1) and 8 (G8) are mostly related to loops between TM helices, groups 2 (G2) and 3 (G3) to the N-termini of helices, while group 7 (G7) characterizes the C-termini and make junctions between helical regions and connecting loops (see Fig. 5b). Interestingly, the two distinct helical regions (positions 5–13 and 22–49) are associated with group 6 (G6) but also to group 4 (G4) and group 5 (G5), which intercalate between two G6 for the longest helical stretch (position 22–49). This analysis demonstrates the existence of sequence–structure specificity in helical motif(s) in membrane proteins, which was trapped by the HPM method. These distinctions are reflected in (1) the amino acids distributions (see Figs. 4b, S2) and (2) the location of the fragments with respect to the membrane (Figure S3). Indeed, an over-representation of charged and polar residues is found in groups G1, G7 and G8 in relation with fragments mostly found outside the membrane (see Figure S3). In contrast, G4, G5 and G6 fragments are associated with helices embedded in the membrane, deeply for G4 and G6 and towards a slightly more interfacial location for G5 (see Figure S3). The differences between G6 and G4, G5 mainly originate from peculiar sequence properties, in particular over-representation of glycine, and small polar residues or charged residues (see Figs. 4b, S2), which impacts the geometry of the corresponding fragments (see Figure S4). However, note that subtle distinctions can be found depending on the positions along HP*. For example,

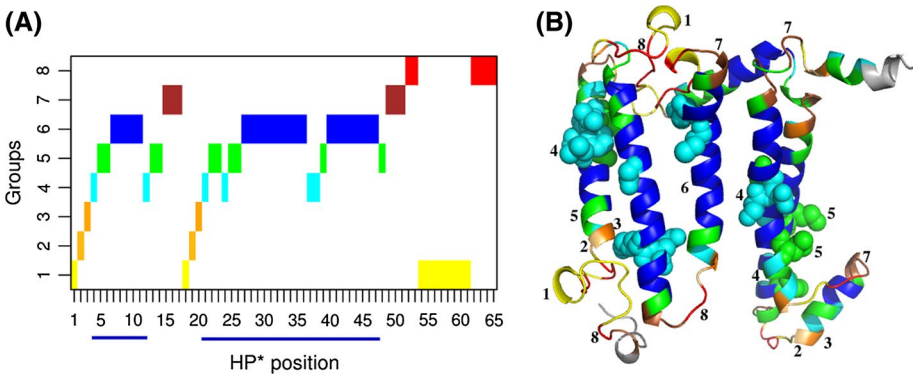


Fig. 5 Clustering analysis of the HP* matrix. **a** A hierarchical clustering analysis was performed on the vectors *v* of HP* matrix, and the result was clustered using complete method into 8 groups (on *y*-axis) defined as : group 1 (positions 1, 18 and 54–61); group 2 (pos 2, 19); group 3 (pos 3, 20); group 4 (pos 4, 12, 21, 24, 37–38); group 5 (5–6, 13–14, 22–23, 25–26, 39 and 48); group 6 (7–11, 27–36 and 40–47); group 7 (pos 15–17 and 49–51); group 8 (pos 52–53 and 62–65). On *x*-axis, the positions p of HP* are shown and the blue lines delimit the helical zones. **b** Visualization of clustering groups on the chain L of photosynthetic reaction centre from *Rhodopseudomonas viridis*. Each fragment of the protein is coloured according to the corresponding group colour in **a**. The labels refer to group number. The residues shown in Van der Waals spheres correspond to fragments clustered in positions 37, 38 and 39 of HP* (see also **a**)

G4 and G5 fragments in positions 4–5, located mainly outside the membrane, are significantly more distorted than G4, G5 fragments in positions 37–39 located inside the membrane (see Figures S4 and S5). Therefore, helical distortions depend on the membrane environment, the helices being more deformed at the extremities or in the periphery of the membrane (see positions 4–6, 12–14, 21–23, 48 in Figure S4) compared to helices inside the membrane (see positions 37–39). This result is related to the distribution of *Z*-scores PB and AA discussed above, about the role of glycine inside the transmembrane helices. Thus, this simplified view offers a first insight of the structural organization of membrane proteins, as an indicator of the localisation in the protein structure (helical ends, connecting loops, distorted helices) but also regarding the membrane location (inside, outside, periphery).

Comparison of amino acid relation in HPM with substitution matrices

As described above, HPM procedure enlightens membrane protein sequence specificities. In the present section, we examine the distances $D_{i,k}$ between HPM *Z*-scores of amino acid *i* and amino acid *k* (see "Materials and methods" section). Although they were calculated from sequence and structure property classification, we chose to compare these distances $D_{i,k}$ with the amino acid substitution values used in classical substitution matrices (SM) and established from direct sequences comparison. Indeed, we hypothesized that the shorter the distance between amino acids, the more favoured the substitution between them.

The resulting 20 × 20 matrix, named HP*-SM, was compared to different substitution matrices, specific of membrane proteins (TM-SMs) or not. We calculated Spearman correlation coefficients, which relate on ranking of pairings. Pearson correlation coefficients showed similar tendencies but were smaller. We also included in the comparison the environment-specific substitution matrices developed by Hill et al., which considered different substitution properties depending on the location of the amino acid relative to the membrane, its secondary structure and its degree of exposure to its environment, i.e. TH(A/a) specific of residues in contact with the tails of lipids (T) in a helical configuration (H) and Accessible (A) or not (a), IHA for Interfacial (I) helices (H) which are Accessible (A) and span the hydrophilic and hydrophobic parts of the membrane; and PHA for helices (H) lining a Pore (P) which is Accessible (A) (Hill et al. 2011).

As a main result, we observed that Spearman Correlation Coefficient (SCC) values between the different TM-SMs (PHAT 75/73, PHDhtm 80 and Persson–Argos 80) are quite large (~0.8 on average) as well with the Hill's environment-specific matrices corresponding to the helical segments located in the most hydrophobic regions of the membrane, at the interface or within a pore region, i.e. TH*, IHA, PHA (see Table 1). In comparison, SCC values between TM-SMs (including TH*, IHA PHA) and BLOSUM 62 are significantly smaller, which fall from 0.8 to 0.5. SCC between HP*-SM and TM-SMs are even lower, ranging from 0.32 to 0.47 while being slightly higher with BLOSUM 62 (0.53).

Considering the strong difference in the way the matrices were established, this result is not unexpected. HP*-SM was constructed from a set of non-redundant sequences, contrary to what was done for establishing TM-SMs or BLOSUM matrix. This leads to a main difference in the

Table 1 Spearman correlation on amino acid relationship, between HPM and standard substitution matrices

	Blosum62	THa	THA	IHA	PHA	PHAT75-73	Persson_argos_80	PHDhtm80	HPM
THa	0.49	1	0.79	0.69	**0.81**	**0.89**	**0.85**	**0.87**	0.37
THA	0.68	0.79	1	**0.83**	**0.85**	0.77	**0.92**	**0.85**	0.54
ICA	**0.85**	0.53	0.73	**0.83**	0.79	0.52	0.75	0.59	0.53
IEA	0.76	0.37	0.52	0.64	0.71	0.36	0.58	0.44	0.52
IHA	0.72	0.69	**0.83**	1	**0.82**	0.72	**0.84**	0.79	0.38
PCA	**0.88**	0.43	0.57	0.69	0.73	0.44	0.67	0.46	0.39
PEA	0.69	0.33	0.43	0.58	0.67	0.32	0.53	0.38	0.38
PHA	0.80	**0.81**	**0.85**	**0.82**	1	0.77	**0.9**	0.82	0.49
PHAT75-73	0.50	**0.89**	0.77	0.72	0.77	1	**0.87**	**0.94**	0.32
Persson_argos_80	0.73	**0.85**	**0.92**	**0.84**	0.9	**0.87**	1	**0.92**	0.47
PHDhtm80	0.56	**0.87**	**0.85**	0.79	**0.82**	**0.94**	**0.92**	1	0.4
HPM	0.53	0.37	0.54	0.38	0.49	0.32	0.47	0.4	1

Values larger than 0.8 are in bold

values of the diagonal that are all null in HP*-SM and different and non-null in other matrices. Clearly, when the local structures are accounted, the relations between amino acids differ from those trapped by evolutionary approaches.

An alternative view of the similarities and striking differences between the matrices can be yielded by a hierarchical clustering of the matrices (see Fig. 6), which quantifies the closeness or co-association of amino acids in the different TM-SMs, BLOSUM and HP*-SM.

The resulting dendrogram shows at the top of the hierarchy two branches separating aromatic and hydrophobic residues (I, L, V, M, F, Y and W) from polar and charged residues (D, E, R, K, N, D, H and Q), with one exception, PHDhtm 80, where R and K are separated from other charged residues. Consequently at this level, we found similar associations as observed for "classical" SM. However, as progressing along the hierarchy, significant differences appear and the situation becomes more contrasted between the different matrices. Small residues S, A and T are grouped together but can be located into different subbranches, closer to charged residues in HP*-SM and BLOSUM in contrast to dedicated TM-SMs in which they are closer to hydrophobic residues. The location of C, P and G is more fluctuating. The reliability of the clustering is significantly larger for dedicated TM-SM and BLOSUM compared to HP*-SM, as assessed by the bootstrap p values and whatever the number of resampling cycles.

Considering frequencies of amino acids in transmembrane and non-transmembrane regions separately for HPM does not strongly modify the previous conclusions (data not shown). Only P and G behave differently between non-transmembrane and transmembrane regions. Interestingly, the p values for HP*-SM are improved when the distribution of amino acids in transmembrane and non-transmembrane regions are considered separately.

The differences mainly occur because the 3D structures are taken into account. As a consequence, using HPM amino acid associations will help to preserve both the physico-chemical properties but also the local 3D structures. It will be useful for example to design new sequences respecting a given fold.

HPM-substitution matrix (HPM-SM) for TM sequence alignment

As HP*-SM was shown to be rather different from other substitution matrices, we chose to compare its performance to the most widely used substitution matrix, namely Blosum62 matrix. We examined two different situations that consist in evaluating HPM-SM (see Table S4 in SI) in the case of (1) multiple sequence alignments (BaliBase dataset or Bali) and (2) pairwise sequence alignments deduced from 3D structure superimposition (Homep2). This second dataset is supposed to be more appropriate for evaluating the HPM-SM performance insofar as HPM-SM was built from a combination of sequence and structural information.

The first step required determining the optimal couple of parameters for gaps (gapopen and gapextend). The optimization was performed independently with a fine grid search on the two datasets. For the evaluation of the quality of the alignments, different scores frequently used in similar studies were considered (see "Materials and methods" section). The results were significantly different according to the scores used. Hence, we chose to select the couple of parameters that were the least sensitive to the scoring scheme, while giving better results than Blosum62 in more than 60 % of the cases, which is an arbitrary threshold. For the Bali dataset, the optimal parameters would be those that led to the greatest number of families (Bali NF) with HPM-SM scores larger than Blosum62 scores, 8 being the goal target.

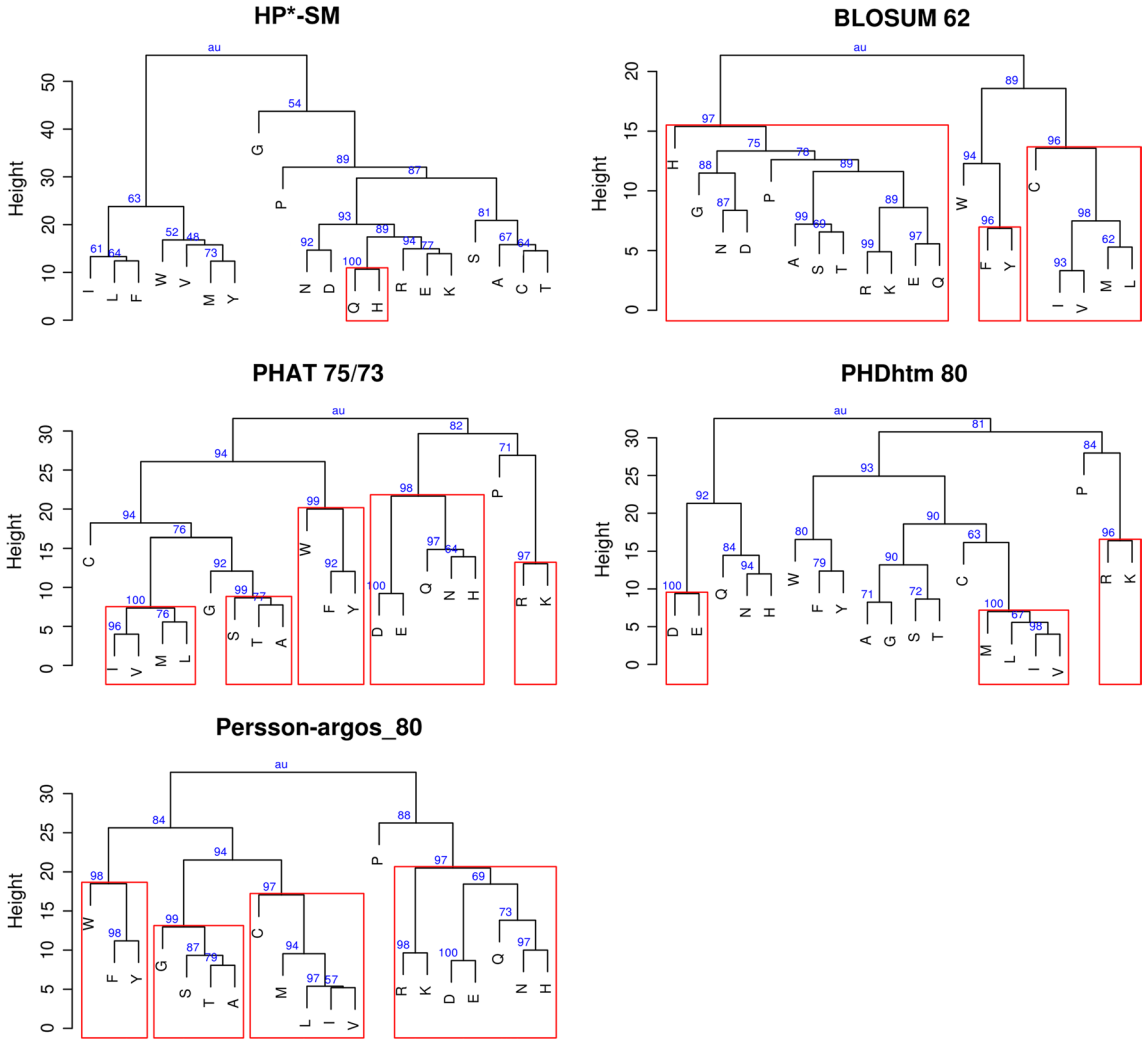


Fig. 6 Comparison of residue relationship in HPM with substitution matrices. The dendrograms were built from the following matrices: BLOSUM 62 (Henikoff and Henikoff 1992), TM-SMs [PHAT 75/73 (Ng et al. 2000), PHDhtm 80, (Ng et al. 2000), Persson-Argos 80 (Persson and Argos 1994)] and amino acid *Z*-score matrix for HPM (see Table S3). For each matrix, a Euclidean Distance Matrix, based on pair of amino acid vectors, was computed. Then, these distance matrices were used to perform hierarchical clustering using complete method (clustering based on similarity), bootstrap analysis (1000 replications) for the estimate of uncertainty, *p* values (au) and building of dendrograms. *Red boxes* were drawn using pvrect with a threshold of 0.95, i.e. a red box is drawn when an edge has a *p* value greater than or equal to 95 %. Finally, all these analyses were performed by using Pvclust package of R (Suzuki and Shimodaira 2006) (colour figure online)

Homep2 dataset

For this dataset, we ended with a set of 9 couples of gapopen/gapextend parameters that led to a better ranking of HPM-SM alignments in more than 60 % of the cases, i.e. 106 among 177, whatever the type of scores considered. The results are detailed in Table 2. In most cases, the best alignments were obtained with gapopen values larger than 15 and with gapextend smaller than to 3.1 (see an example in Figure S6 of SI).

BaliBase Ref7

The situation for this dataset was much more challenging. Indeed, the Bali reference multiple alignments of each membrane protein family include a large number of gaps. The scores Q, TC and Cline, mainly assume that the widths of the alignments are similar. The HPM-SM multiple alignments were in general less wide than those established with Blosum62, even when small gapopen values were considered. It is the reason why HPM-SM did not perform better that Blosum62 with Q, TC and Cline scores in most cases. The only score which accounts for the size of the test alignment is Modeler. We report in Table 2 a subset of parameters with the corresponding number of families having larger Modeler scores with HPM-SM compared to Blosum62. The largest number of families was systematically obtained with a gapextend equals to 0.1, while the gapopen can cover a larger range.

Table 2 Gap values providing HPM-SM alignments with a better ranking than Blosum62, for the four scores

Gapopen	Gapextend	Homep2 Q[a]	Homep2 TC[a]	Homep2 Cline[a]	Homep2 Modeler[a]	Homep2 Q[b]	Homep2 TC[b]	Homep2 Cline[b]	Homep2 Modeler[b]
16.5	1.1	115	115	107	109	65.0	65.0	60.5	61.6
17	1.1	115	115	107	109	65.0	65.0	60.5	61.6
17.5	1.1	117	117	109	112	66.1	66.1	61.6	63.3
18	1.1	116	116	108	111	65.5	65.5	61.0	62.7
18.5	0.6	111	111	106	113	62.7	62.7	59.9	63.8
19	0.6	112	112	107	114	63.3	63.3	60.5	64.4
19	1.1	114	114	106	110	64.4	64.4	59.9	62.1
19.5	0.6	113	113	109	115	63.8	63.8	61.6	65.0
19.5	1.1	116	116	109	112	65.5	65.5	61.6	63.3

Gapopen	Gapextend	NF Bali[c]	Homep2 Modeler[a]	Homep2 Modeler[b]
16.0	0.1	8	111	62.7
16.5	0.1	8	108	61
13.5	0.1	8	106	59.9
14.0	0.1	8	106	59.9
13.0	0.1	8	102	57.6
15.5	0.1	7	112	63.2
19.5	0.1	7	111	62.7
18.0	0.1	7	110	62.1
18.5	0.1	7	110	62.1
19.0	0.1	7	110	62.1
17.0	0.1	7	109	61.6
17.5	0.1	7	108	61.1
20.0	0.1	7	106	59.9
15.0	0.1	7	105	59.3
20.5	0.1	7	104	58.8
22.0	0.1	7	104	58.8
22.5	0.1	7	102	57.6
23.0	0.1	7	102	57.6
24.5	0.1	7	102	57.6
23.5	0.1	7	101	57.1
24.0	0.1	7	101	57.1
25.0	0.1	7	100	56.5
17.5	**0.6**	**6**	**106**	**59.9**
14.5	0.1	6	105	59.3
21.0	0.1	6	103	58.2
21.5	0.1	6	103	58.2

The gapopen and gapextend values recommended for HPM-SM are in bold

[a] The numbers of pairwise alignments with a HPM-SM score better than the one of BLOSUM 62

[b] The % of pairwise alignments with a HPM-SM score better than the one of BLOSUM 62, based on 177 alignments in total

[c] NF Bali represents the number of Bali families with a HPM-SM score better than the one of BLOSUM 62, based on 8 families

We tested these parameters on the Homep2 dataset in the aim to cross-validate our results (Table 2). With a gapopen larger than 16.0, the results for Bali can reach the gold target and comply with the threshold we chose for Modeler score in Homep2, i.e. >60 %. However, the other scores were below the threshold, mainly due to the gapextend value. Reciprocally, we calculated Bali NF with the optimal couples established with Homep2. In this case, the largest NF equals 4 and was obtained with gapopen equals to 16.5 or 18.5.

Interestingly, HPM-SM outperforms Blosum62 for 6 families among 8 with the couple of gapopen/gapextend equals to 17.5/0.6. For this couple of values, the percentage of pairs from the Homep2 dataset with a better ranking with HPM-SM compared Blosum62, are 59, 59, 56 and 60 % for Q, TC, Cline and Modeler scores, respectively.

These preliminary but encouraging results give substantial support to the usefulness of HPM-SM in the membrane protein sequence comparison, in particular when searching for optimal pairwise alignment for homology modelling.

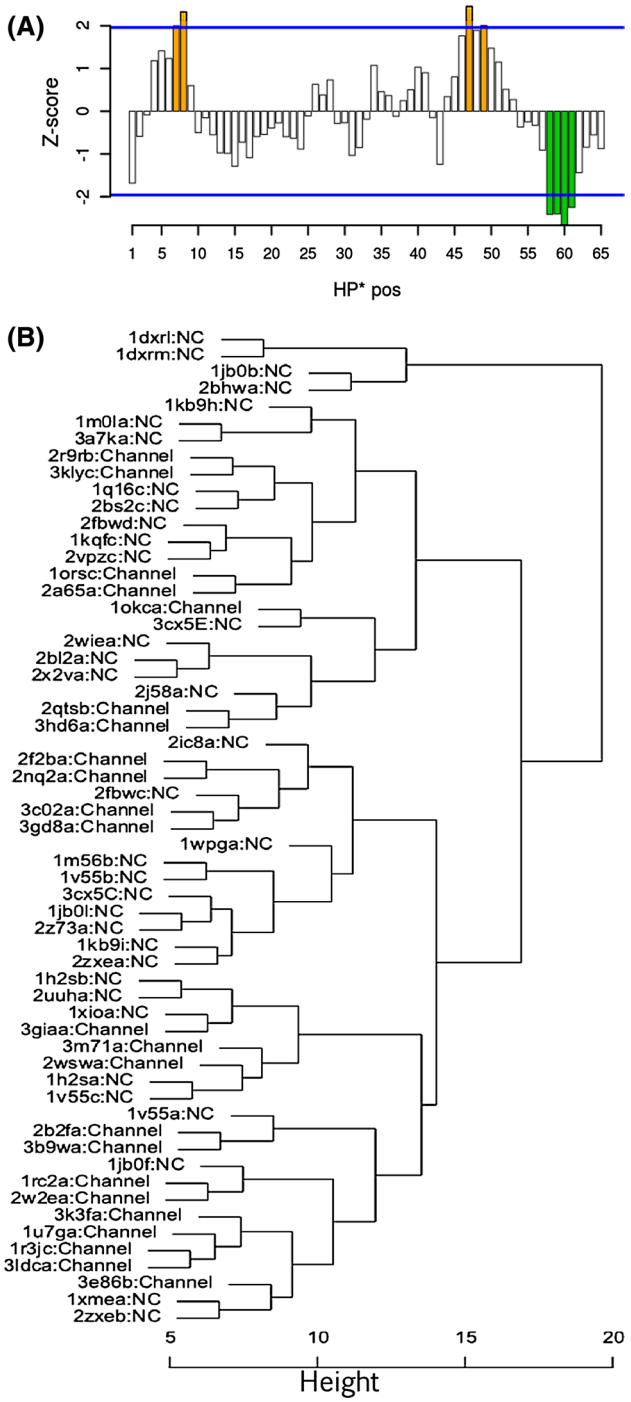
Discussion

Besides interesting features that were detailed above, useful applications of HPM methodology are proposed and discussed in the following section. The first one relates on determining protein function from sequence and structure properties. Few tools are nowadays available (Zdobnov and Apweiler 2001; Watson et al. 2005; Gabdoulline et al. 2003, 2006; Laskowski et al. 2005a, b) while the needs are tremendous. Most current methodologies consist in using information from similar (e.g. co-evolved) proteins, functional sites or motifs in proteins, accurately annotated. In the present case, we examined whether HPM would enable to highlight some specific signature(s) that might be characteristic of channels/transporters and would help in deciphering function of transport. In this aim, (1) we selected a subset of proteins that we identified as transporters or channels in the whole dataset (see "Materials and methods" section). Each protein was divided in fragments of 5 residue length; (2) each fragment was encoded according to the 31 properties. This target vector was compared to the 65 vectors of HP* matrix; (3) the HP* position having the smallest distance with the target vector was representative of the given fragment. The distribution of fragments for the subset was then compared to the distribution of fragments in the whole dataset, by calculating *Z*-scores in each position (see Fig. 7a). While positions 58–61 were underrepresented (*Z*-score values of −2.4, −2.4, −2.7 and −2.3, respectively), four positions (7, 8, 47 and 49) were significantly over-represented (*Z*-score values >1.96), and might be considered as markers of the transport/channel function. However, this observation entails thorough analyses to be confirmed. Nevertheless, some hints were brought when we considered the HPM signature of each protein in the dataset. We ran the steps 1–3 of the procedure described above, except that the analysis was performed on individual protein chains. For each protein chain, we compared the distribution of fragments in each HP* position with the distribution of fragments in the whole dataset. The HPM signature of a protein is defined by the corresponding *Z*-score values in each position. The resulting vectors were then clustered using a hierarchical clustering using complete method and Euclidean distance as metric. The leaves of the dendrogram correspond to the pairing of proteins based on the distances between the *Z*-score profiles computed for each protein. By focusing on the leaves, we observe only two mispairings (1okcA:C-3cx5E:NC, 1xioA:NC-3giaA:C), among a total of 22 pairs. This value is rather low, considering the number of possible combinations (36 Non-Channels × 23 C), i.e. 828 maximum of possible mispairings. Thus, apart these two cases, we found channel proteins paired with channel proteins and non-channel proteins paired with non-channel proteins. Insofar as "Channel/Non channel" was not an input classification feature, the clustering shown here suggests that the HP* profile might be helpful for annotating. Further investigations with a larger validation dataset are necessary to confirm these preliminary observations.

Hence, HPM enables to classify functional properties using local 3D information and physico-chemical properties. This preliminary but encouraging result opens the way to use HPM to annotate new structures and discover new functions, which could be tested further.

In a second application, we examined whether the sequence–structure relation highlighted in HPM is appropriate to evaluate the quality of 3D structures. As an example, we first studied Msba protein, an ABC transporter, for which correct and incorrect structures are available (3B60 and 1Z2R, respectively) (Matthews 2007; Reyes and Chang 2005; Ward et al. 2007). Note that neither the correct structure nor the wrong one of MsbA was included in the dataset.

We calculated the Smin score (called HPM Score) of each fragment for each position of HPM (see Eq. 1) and compared to the scores along the sequence for the two structures (see Fig. 8a). The HPM score shows a significant difference in the region ranging from residues 45 to 75, reaching a maximum ~66 at residue 57. This difference is considerably larger than the maximal difference observed along a molecular dynamics simulation of a high-resolution X-ray structure (see Figure S7). This difference is due to the loop (47–69 residue) observed in the incorrect structure (1Z2R), which is replaced by the ends of two transmembrane helices connected by a short loop in the correct structure. Hence, in this region, the HPM sequence–structure relationship disfavours loop of 1Z2R. A visual inspection of the 3D structures shows that this zone strongly impacts the relative position of the two sub-regions and the final topology (see Fig. 8b). This explains why large RMSD was observed (~15.5 Å) between the two whole structures. We also tested two additional MsbA X-ray structures (PDB codes 3B60 and 4Q4A). The sequences share less than 20 % of identity. The folds are similar but the 3D structures correspond to different conformational states and are difficult to superimpose (see Figure S8A). Nevertheless, the HPM profiles are similar as exemplified in Figure S8B,



◀**Fig. 7** Specificities of channels/transporters in the databank. **a** *Z*_ scores of channel/transporter fragments in each position of the HP* position. A sub-databank was built from channel/transporter proteins (labelled "Channel" in the dendrogram). *Z*-score values of the distribution of sub-databank fragments were computed for each position of HP* matrix. The *green* positions correspond to an under-representation of sub-databank fragments compared with the whole one, whereas the orange ones correspond to an over-estimation. The gradient colour is the same than Fig. 4. **b** Dendrogram of HPM signature of each protein. Each protein is characterized by a vector *v* of length 65, containing the *Z*-score values of corresponding fragments along the HP* matrix. The resulting dendrogram is obtained by *h-clust* function (R package) with complete method and Euclidean distance. The label of leaves correspond to PDB codes for each protein and its chain (*lower case letter*). The labels NC correspond to non-channel/transporter proteins, whereas channel labels merge channel and transporter proteins (colour figure online)

although the sequences and the 3D significantly differ. Therefore, HPM profile was able to retrieve sequence and structure adequacy in both cases.

Consequently, HPM would have helped to identify the incorrect zone and to suggest how to correct it. In comparison, when we calculated DOPE score, a measure based on 3D contacts compatibility and frequently used to assess model quality, we did not observe significant difference between the two structures in the corresponding zone (see Fig. 8a).

In the last example, we tested whether HPM score may be used to evaluate models or to design proteins. To do so, we generated five 3D models for the sequence of chain L of photosynthetic reaction centre from a thermophilic bacterium, *Thermochromatium tepidum* (PDB code: 1EYS (Nogi et al. 2000)) using I-TASSER webserver (Zhang 2008). We excluded the correct structure and all the structures with a sequence identity upper than 10 %. After superimposition of five models with the reference X-ray structure, all the models exhibited a similar fold, see Fig. 9a. The main differences were located in the connecting loops, N- and C-termini (see Fig. 9a, b).

Then, we ranked these models using different scoring measures. Several scoring functions were benchmarked, and divided into two groups (see Table 3). The first group (noted P-G for predicted group) corresponds to scores that are used to predict the quality of models [I-Tasser Score (CS) (Zhang 2008), HPM Score and ProQM Score (Ray et al. 2010)]. These scores are valuable in real situations for which no experimental X-ray structures are available, i.e. for blind tests. The second group (noted E-G for evaluation group) includes classical measures of the quality of model. They are based on the comparison with a known reference 3D structure [GDT_TS, GDT_HA, TM-align (Zhang and Skolnick 2005), TM Score (Zhang and Skolnick 2004), RMSD, Maxsub (Siew et al. 2000)] and are frequently used in CASP competition rounds to rank models. We examined local and global structural alignments as well.

Because scoring values are quite different as well as their ranges, we mainly discuss the rank of the models (see Table 3). As a first remark, the ranks of models are significantly different as well within P-G, between P-G and E-G but also more surprisingly within E-G. For example, ProQM, specifically designed for the prediction of the quality of membrane protein models, gave model 5 in rank 1, while it was ranked 3 or 4 with most of E-G tools. Similarly, I-Tasser CS score considered model 1 as the best one, whereas it was frequently badly ranked by E-G measures.

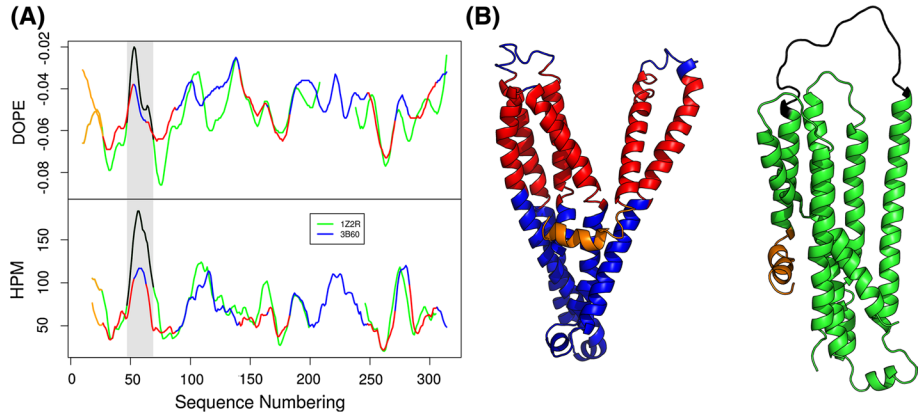


Fig. 8 Detection of incorrect topology by using HPM Score. **a** DOPE/HPM Scores along the sequence numbering, for two structures (1Z2R and 3B60). The *orange* regions correspond to the N-terminus (residues 10–26). The *red* regions along the *blue curve* correspond to the transmembrane regions determined by OPM (Lomize et al. 2006) for the protein 3B60. The grey zone and the black region of the green curve correspond to the residues 47–69 (incorrect loop in 1Z2R protein). **b** Cartoon representations of 3B60 on the left and 1Z2R on the right. The RMSD on Cα in the region 10–314, without taking into account the gap between 208 and 238 residues in superposition, is about 15.5 Å as large as for non-related protein

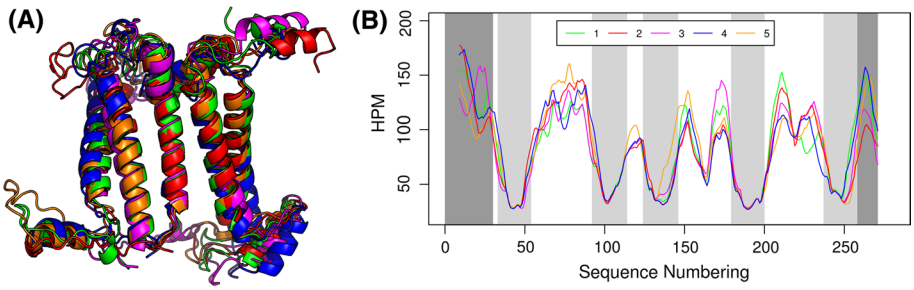


Fig. 9 HPM profile for ranking structural models. **a** Cartoon representations of five models generated by I-TASSER (Zhang 2008) and aligned on the X-ray structure (PDB code: 1EYS (Nogi et al. 2000)). The colour code is the same than the one used on the right profile. **b** HPM Score profile for 5 structural models. The *light grey* regions correspond to transmembrane regions defined by OPM. The *dark grey* zones highlight the N- and C-termini of the protein

HPM does not perform better: model 2 is ranked in first position whereas it is in the worst rank in E-G measures. This clearly illustrates that the identification of the correct model and the ranking models remain a difficult task.

These differences led us to examine in more details HPM scores along the sequence as described above (see Fig. 9b) and to propose a new strategy for ranking the models. We searched for the model with the largest number of minimal values, which is assumed to be the best one. Interestingly, model 3 has by far the largest number of minimal values along the sequences (76 compared to 57 for models 2 and 4), even though the global HPM score, i.e. the sum of local HPM Scores, was not the lowest one. Finally, on average, models 3 and 4 are considered as the best ones by the E-G methods, in agreement with the largest number of HPM minimal values along the sequence. Hence, rather than the global HPM score alone, we suggest to use this measure to select the best model.

This preliminary study is rather encouraging considering that no optimisation was performed for the present goal, compared to strategies specifically designed for. Together with the MsbA study, it shows that local sequence–structure compatibility trapped by HPM strategy might be helpful in detecting 3D regions that would require further refinement and/or could complement other ranking methodologies. So, according to these results, a protein design, based on most favourable fragments amongst models, could improve the quality of a final structure. The local HPM Score is already informative to highlight some interesting regions. Nonetheless, the total HPM Score could be further improved by normalization on the protein length or other parameters, like done in TM Score for example. Moreover, these results need to be confirmed by a larger study that will necessitate the construction and evaluation of many models and decoys.

Table 3 Assessment of the quality of structural models by ranking

	HPM	CS	ProQM	RMSD (local)[a]	GDT-TS (local)[a]	RMSD (ref)[a]	GDT-TS (ref)[a]
1st rank	2	1	5	4	4	3	4
2nd rank	3	4	4	1	3	4	3
3rd rank	1	5	3	5	1	1	6
4th rank	4	2	1	2	5	1	5
5th rank	5	3	2	3	2	2	2

	TM-align	GDT-HA[b]	RMSD_TMscore[b]	TM-Score[b]	GDT-TS[b]	Max Sub[b]
1st rank	4	3	4	4	3	3
2nd rank	3	4	1	3	4	4
3rd rank	1	1	5	1	5	1
4th rank	5	5	2	5	1	5
5th rank	2	2	3	2	2	2

Five models were generated using I-TASSER webserver (Zhang 2008), the table gives the number of models ranked from the best to the worst. The three first columns correspond to prediction tools to rank structural models, so HPM was benchmarked with I-TASSER (CS score) (Zhang 2008) and ProQM (Ray et al. 2010). Then, HPM was benchmarked with objective measures, i.e. based on a reference (the sequence of chain L of 1EYS), GDT-TS, GDT-HA, RMSD, TM-align, TM Score and Maxsub

[a] The measures computed from the iPBA webserver (Gelly et al. 2011)

[b] The measures computed from TM Score webserver of Zhang's lab

Conclusion

Better understanding of membrane protein structures and improving their quality is an emerging challenge during these last years. Here, we propose to use an unsupervised learning approach, called HPM, which has been already shown its efficiency for analysing globular proteins. Applying to all-α transmembrane protein databank, HPM allowed studying their sequence–structure relationship by clustering close protein fragments together. Due to various functions of membrane proteins and their environment, we suggest that some patterns could be highlighted, and we show that HPM may be a promising tool able to separate channels and non-channels. As shown, it could also be used to perform specific sequence alignments. Finally, HPM is a learning approach based on the minimization of a scoring function; we show that this former could be used to evaluate protein structures, i.e. quality of structures or ranking of models, at a local or global level. This last application offers a way to check and improve the structure quality of membrane proteins.

Acknowledgments We would like to thank our colleagues Jean-Christophe Gelly and Stéphane Téletchéa for their precious advice on this article. This work was supported by grants from the Ministry of Research (France), University Paris Diderot, Sorbonne, Paris Cité (France), National Institute for Blood Transfusion (INTS, France), National Institute for Health and Medical Research (INSERM, France) and labex GR-Ex to JE, CE and AdB, National Institute for Agricultural Research (INRA, France) to AU and National Center for Scientific Research (CNRS) to JE. JE also acknowledges an ATER (research and teaching) position from University Paris Diderot (France). The labex GR-Ex, reference ANR-11-LABX-0051 is funded by the program "Investissements d'avenir" of the French National Research Agency, reference ANR-11-IDEX-0005-02.

Conflict of interest The authors declare that they have no conflict of interest.

References

Arinaminpathy Y, Khurana E, Engelman DM, Gerstein MB (2009) Computational analysis of membrane proteins: the largest class of drug targets. Drug Discov Today 14(23–24):1130–1135. doi:10.1016/j.drudis.2009.08.006

Bansal M, Kumar S, Velavan R (2000) HELANAL: a program to characterize helix geometry in proteins. J Biomol Struct Dyn 17(5):811–819. doi:10.1080/07391102.2000.10506570

Benros C, de Brevern AG, Etchebest C, Hazout S (2006) Assessing a novel approach for predicting local 3D protein structures from sequence. Proteins 62(4):865–880. doi:10.1002/prot.20815

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242 **(gkd090 [pii])**

Bornot A, Etchebest C, de Brevern AG (2009) A new prediction strategy for long local protein structures using an original description. Proteins 76(3):570–587. doi:10.1002/prot.22370

Bornot A, Etchebest C, de Brevern AG (2011) Predicting protein flexibility through the prediction of local structures. Proteins 79(3):839–852. doi:10.1002/prot.22922

Burgess SM, Delannoy M, Jensen RE (1994) MMM1 encodes a mitochondrial outer membrane protein essential for establishing and maintaining the structure of yeast mitochondria. J Cell Biol 126(6):1375–1391

Cline M, Hughey R, Karplus K (2002) Predicting reliable regions in protein sequence alignments. Bioinformatics 18(2):306–314

Cohen RS (2005) The role of membranes and membrane trafficking in RNA localization. Biol Cell 97(1):5–18. doi:10.1042/BC20040056

Cordes FS, Bright JN, Sansom MS (2002) Proline-induced distortions of transmembrane helices. J Mol Biol 323(5):951–960 **(S0022283602010069 [pii])**

Dayhoff MO, Schwartz RM (1978) A model of evolutionary change in proteins. Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington, pp 345–358

de Brevern AG (2005) New assessment of a structural alphabet. Silico Biol 5(3):283–289

de Brevern AG, Hazout S (2000) Hybrid Protein Model (HPM): a method to compact protein 3D-structure information and physicochemical properties. IEEE Comp Soc (SPIRE 2000) S1:49–54

de Brevern AG, Hazout S (2001) Compacting local protein folds with a "hybrid protein model". Theor Chem Acc 106(1–2):36–47

de Brevern AG, Hazout S (2003) 'Hybrid protein model' for optimally defining 3D protein structure fragments. Bioinformatics 19(3):345–353

de Brevern AG, Etchebest C, Hazout S (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41(3):271–287. doi:10.1002/1097-0134(20001115)41:3<271:AID-PROT10>3.0.CO;2-Z

de Brevern AG, Valadie H, Hazout S, Etchebest C (2002) Extension of a local backbone description using a structural alphabet: a new approach to the sequence–structure relationship. Protein Sci 11(12):2871–2886. doi:10.1110/ps.0220502

de Brevern AG, Bornot A, Craveur P, Etchebest C, Gelly JC (2012) PredyFlexy: flexibility and local structure prediction from sequence. Nucleic Acids Res 40((Web Server issue)):W317–W322. doi:10.1093/nar/gks482

Edgar RC (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform 5:113. doi:10.1186/1471-2105-5-113

Edgar RC (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797. doi:10.1093/nar/gkh340

Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 179(1):125–142 **(0022-2836(84)90309-7 [pii])**

Elofsson A, von Heijne G (2007) Membrane protein structure: prediction versus reality. Annu Rev Biochem 76:125–140. doi:10.1146/annurev.biochem.76.052705.163539

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A (2006) Comparative protein structure modeling using Modeller. Curr Protoc Bioinform Chapter 5:5.6. doi:10.1002/0471250953.bi0506s15

Forrest LR, Tang CL, Honig B (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys J 91(2):508–517. doi:10.1529/biophysj.106.082313

Fuchs A, Kirschner A, Frishman D (2009) Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks. Proteins 74(4):857–871. doi:10.1002/prot.22194

Gabdoulline RR, Hoffmann R, Leitner F, Wade RC (2003) ProSAT: functional annotation of protein 3D structures. Bioinformatics 19(13):1723–1725

Gabdoulline RR, Ulbrich S, Richter S, Wade RC (2006) ProSAT2—Protein Structure Annotation Server. Nucleic Acids Res 34(Web Server issue):W79–W83. doi:10.1093/nar/gkl216

Gamper N, Shapiro MS (2007) Regulation of ion transport proteins by membrane phosphoinositides. Nat Rev Neurosci 8(12):921–934. doi:10.1038/nrn2257

Gelly JC, Joseph AP, Srinivasan N, de Brevern AG (2011) iPBA: a tool for protein structure comparison using sequence alignment strategies. Nucleic Acids Res 39(Web Server issue):W18–W23. doi:10.1093/nar/gkr333

Giacomini KM, Huang SM, Tweedie DJ, Benet LZ, Brouwer KL, Chu X, Dahlin A, Evers R, Fischer V, Hillgren KM, Hoffmaster KA, Ishikawa T, Keppler D, Kim RB, Lee CA, Niemi M, Polli JW, Sugiyama Y, Swaan PW, Ware JA, Wright SH, Yee SW, Zamek-Gliszczynski MJ, Zhang L (2010) Membrane transporters in drug development. Nat Rev Drug Discov 9(3):215–236. doi:10.1038/nrd3028

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Hall SE, Roberts K, Vaidehi N (2009) Position of helical kinks in membrane protein crystal structures and the accuracy of computational prediction. J Mol Graph Model 27(8):944–950. doi:10.1016/j.jmgm.2009.02.004

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919

Hill JR, Kelm S, Shi J, Deane CM (2011) Environment specific substitution tables improve membrane protein alignment. Bioinformatics 27(13):i15–i23. doi:10.1093/bioinformatics/btr230

Ikeda M, Arai M, Okuno T, Shimizu T (2003) TMPDB: a database of experimentally-characterized transmembrane topologies. Nucleic Acids Res 31(1):406–409

Jones DT, Taylor WR, Thornton JM (1994) A mutation data matrix for transmembrane proteins. FEBS Lett 339(3):269–275 **(0014-5793(94)80429-X [pii])**

Joseph AP, Agarwal G, Mahajan S, Gelly JC, Swapna LS, Offmann B, Cadet F, Bornot A, Tyagi M, Valadie H, Schneider B, Etchebest C, Srinivasan N, De Brevern AG (2011) A short survey on protein blocks. Biophys Rev 2(3):137–147. doi:10.1007/s12551-010-0036-1

Kelm S, Shi J, Deane CM (2010) MEDELLER: homology-based coordinate generation for membrane proteins. Bioinformatics 26(22):2833–2840. doi:10.1093/bioinformatics/btq554

Kohonen T (2013) Essentials of the self-organizing map. Neural Netw 37:52–65. doi:10.1016/j.neunet.2012.09.018

Kullback S, Leibler RA (1951) On information andsufficiency. Ann Math Stat 22:79–86

Langelaan DN, Wieczorek M, Blouin C, Rainey JK (2010) Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. J Chem Inf Model 50(12):2213–2220. doi:10.1021/ci100324n

Laskowski RA, Watson JD, Thornton JM (2005a) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33 (Web Server issue):W89–W93. doi:10.1093/nar/gki414

Laskowski RA, Watson JD, Thornton JM (2005b) Protein function prediction using local 3D templates. J Mol Biol 351(3):614–626. doi:10.1016/j.jmb.2005.05.067

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659. doi:10.1093/bioinformatics/btl158

Liu Y, Engelman DM, Gerstein M (2002) Genomic analysis of membrane protein families: abundance and conserved motifs. Genome Biol 3(10):research0054

Lo A, Chiu YY, Rodland EA, Lyu PC, Sung TY, Hsu WL (2009) Predicting helix-helix interactions from residue contacts in membrane proteins. Bioinformatics 25(8):996–1003. doi:10.1093/bioinformatics/btp114

Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006) OPM: orientations of proteins in membranes database. Bioinformatics 22(5):623–625. doi:10.1093/bioinformatics/btk023

Marsico A, Henschel A, Winter C, Tuukkanen A, Vassilev B, Scheubert K, Schroeder M (2010a) Structural fragment clustering reveals novel structural and functional motifs in alpha-helical transmembrane proteins. BMC Bioinform 11:204. doi:10.1186/1471-2105-11-204

Marsico A, Scheubert K, Tuukkanen A, Henschel A, Winter C, Winnenburg R, Schroeder M (2010b) MeMotif: a database of linear

motifs in alpha-helical transmembrane proteins. Nucleic Acids Res 38 (Database issue):D181–D189. doi: 10.1093/nar/gkp1042

Matthews BW (2007) Five retracted structure reports: inverted or incorrect? Protein Sci 16(6):1013–1016. doi:10.1110/ps.072888607

Meruelo AD, Samish I, Bowie JU (2011) TMKink: a method to predict transmembrane helix kinks. Protein Sci 20(7):1256–1264. doi:10.1002/pro.653

Nagarathnam B, Sankar K, Dharnidharka V, Balakrishnan V, Archunan G, Sowdhamini R (2011) TM-MOTIF: an alignment viewer to annotate predicted transmembrane helices and conserved motifs in aligned set of sequences. Bioinformation 7(5):214–221

Nam HJ, Jeon J, Kim S (2009) Bioinformatic approaches for the structure and function of membrane proteins. BMB Rep 42(11):697–704

Ng PC, Henikoff JG, Henikoff S (2000) PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. Bioinformatics 16(9):760–766

Nogi T, Fathir I, Kobayashi M, Nozawa T, Miki K (2000) Crystal structures of photosynthetic reaction center and high-potential iron-sulfur protein from *Thermochromatium tepidum*: thermostability and electron transfer. Proc Natl Acad Sci USA 97(25):13561–13566. doi:10.1073/pnas.240224997

Nugent T, Jones DT (2009) Transmembrane protein topology prediction using support vector machines. BMC Bioinform 10:159. doi:10.1186/1471-2105-10-159

Nugent T, Jones DT (2010) Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. PLoS Comput Biol 6(3):e1000714. doi:10.1371/journal.pcbi.1000714

Nugent T, Jones DT (2012) Membrane protein structural bioinformatics. J Struct Biol 179(3):327–337. doi:10.1016/j.jsb.2011.10.008

Nugent T, Ward S, Jones DT (2011) The MEMPACK alpha-helical transmembrane protein structure prediction server. Bioinformatics 27(10):1438–1439. doi:10.1093/bioinformatics/btr096

Papaloukas C, Granseth E, Viklund H, Elofsson A (2008) Estimating the length of transmembrane helices using *Z*-coordinate predictions. Protein Sci 17(2):271–278. doi:10.1110/ps.073036108

Persson B, Argos P (1994) Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J Mol Biol 237(2):182–192. doi:10.1006/jmbi.1994.1220

Pieper U, Schlessinger A, Kloppmann E, Chang GA, Chou JJ, Dumont ME, Fox BG, Fromme P, Hendrickson WA, Malkowski MG, Rees DC, Stokes DL, Stowell MH, Wiener MC, Rost B, Stroud RM, Stevens RC, Sali A (2013) Coordinating the impact of structural genomics on the human alpha-helical transmembrane proteome. Nat Struct Mol Biol 20(2):135–138. doi:10.1038/nsmb.2508

Pirovano W, Feenstra KA, Heringa J (2008) PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. Bioinformatics 24(4):492–497. doi:10.1093/bioinformatics/btm636

Ray A, Lindahl E, Wallner B (2010) Model quality assessment for membrane proteins. Bioinformatics 26(24):3067–3074. doi:10.1093/bioinformatics/btq581

Reyes CL, Chang G (2005) Structure of the ABC transporter MsbA in complex with ADP.vanadate and lipopolysaccharide. Science 308(5724):1028–1031. doi:10.1126/science.1107733

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4):725–738. doi:10.1038/nprot.2010.5

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815. doi:10.1006/jmbi.1993.1626

Sansom MS, Weinstein H (2000) Hinges, swivels and switches: the role of prolines in signalling via transmembrane alpha-helices. Trends Pharmacol Sci 21(11):445–451 **(S0165614700015534 [pii])**

Sauder JM, Arthur JW, Dunbrack RL Jr (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 40(1):6–22. doi:10.1002/(SICI)1097-0134(20000701)40:1<6:AID-PROT30>3.0.CO;2-7

Shafrir Y, Guy HR (2004) STAM: simple transmembrane alignment method. Bioinformatics 20(5):758–769. doi:10.1093/bioinformatics/btg482

Siew N, Elofsson A, Rychlewski L, Fischer D (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16(9):776–785

Stamm M, Staritzbichler R, Khafizov K, Forrest LR (2013) Alignment of helical membrane protein sequences using AlignMe. PLoS One 8(3):e57731. doi:10.1371/journal.pone.0057731

Stamm M, Staritzbichler R, Khafizov K, Forrest LR (2014) AlignMe—a membrane protein sequence alignment web server. Nucleic Acids Res 42(Web Server issue):W246–W251. doi:10.1093/nar/gku291

Sutormin RA, Rakhmaninova AB, Gelfand MS (2003) BATMAS30: amino acid substitution matrix for alignment of bacterial transporters. Proteins 51(1):85–95. doi:10.1002/prot.10308

Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22(12):1540–1542. doi:10.1093/bioinformatics/btl117

Szalontai B (2009) Membrane protein dynamics: limited lipid control. PMC Biophys 2(1):1. doi:10.1186/1757-5036-2-1

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680

Thompson JD, Plewniak F, Poch O (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15(1):87–88 **(btc017 [pii])**

Thompson JD, Koehl P, Ripp R, Poch O (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins 61(1):127–136. doi:10.1002/prot.20527

Tress ML, Jones D, Valencia A (2003) Predicting reliable regions in protein alignments from sequence profiles. J Mol Biol 330(4):705–718 **(S0022283603006223 [pii])**

Tusnady GE, Dosztanyi Z, Simon I (2004) Transmembrane proteins in the Protein Data Bank: identification and classification. Bioinformatics 20(17):2964–2972. doi:10.1093/bioinformatics/bth340

Tusnady GE, Dosztanyi Z, Simon I (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 33(Database issue):D275–D278. doi:10.1093/nar/gki002

Viklund H, Granseth E, Elofsson A (2006) Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes. J Mol Biol 361(3):591–603. doi:10.1016/j.jmb.2006.06.037

Visser WF, van Roermund CW, Ijlst L, Waterham HR, Wanders RJ (2007) Metabolite transport across the peroxisomal membrane. Biochem J 401(2):365–375. doi:10.1042/BJ20061352

von Heijne G (2011) Introduction to theme "membrane protein folding and insertion". Annu Rev Biochem 80:157–160. doi:10.1146/annurev-biochem-111910-091345

Wallin E, von Heijne G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 7(4):1029–1038. doi:10.1002/pro.5560070420

Walters RF, DeGrado WF (2006) Helix-packing motifs in membrane proteins. Proc Natl Acad Sci USA 103(37):13658–13663. doi:10.1073/pnas.0605878103

Wang XF, Chen Z, Wang C, Yan RX, Zhang Z, Song J (2011) Predicting residue-residue contacts and helix-helix interactions in

transmembrane proteins using an integrative feature-based random forest approach. PLoS One 6(10):e26767. doi:10.1371/journal.pone.0026767

Ward A, Reyes CL, Yu J, Roth CB, Chang G (2007) Flexibility in the ABC transporter MsbA: Alternating access with a twist. Proc Natl Acad Sci USA 104(48):19005–19010. doi:10.1073/pnas.0709388104

Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15(3):275–284. doi:10.1016/j.sbi.2005.04.003

White SH (2009) Biophysical dissection of membrane proteins. Nature 459(7245):344–346. doi:10.1038/nature08142

Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M (2007) Drug-target network. Nat Biotechnol 25(10):1119–1126. doi:10.1038/nbt1338

Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU (2004a) The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proc Natl Acad Sci USA 101(4):959–963. doi:10.1073/pnas.0306077101

Yohannan S, Yang D, Faham S, Boulting G, Whitelegge J, Bowie JU (2004b) Proline substitutions are not easily accommodated in a membrane protein. J Mol Biol 341(1):1–6. doi:10.1016/j.jmb.2004.06.025

Zamyatnin AA (1984) Amino acid, peptide, and protein volume in solution. Annu Rev Biomed Eng 13:145–165

Zdobnov EM, Apweiler R (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17(9):847–848

Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinform 9:40. doi:10.1186/1471-2105-9-40

Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57(4):702–710. doi:10.1002/prot.20264

Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33(7):2302–2309. doi:10.1093/nar/gki524