DATABASE
The Journal of Biological Databases and Curation

Database, 2021, 1–10
doi:10.1093/database/baab013
Database tool

Database tool

MENSAdb: a thorough structural analysis of membrane protein dimers

Pedro Matos-Filipe[1,†], António J. Preto[1,2,†], Panagiotis I. Koukos[3], Joana Mourão[1], Alexandre M.J.J. Bonvin[3] and Irina S. Moreira[4,5,*]

[1]Center for Neuroscience and Cell Biology, University of Coimbra, Coimbra 3005-504, Portugal, [2]PhD Programme in Experimental Biology and Biomedicine, Institute for Interdisciplinary Research, University of Coimbra, Coimbra, 3030-789, Portugal, [3]Bijvoet Centre for Biomolecular Research, Faculty of Science—Chemistry, Utrecht University, Utrecht, 3584, CH, Netherlands, [4]Department of Life Sciences, University of Coimbra, Coimbra, 3000-456, Portugal and [5]Center for Neuroscience and Cell Biology, Center for Innovative Biomedicine and Biotechnology, University of Coimbra, Coimbra, Portugal

*Corresponding author: Tel: +351-239240227 or +351- 231 249 170; Fax: +351- 231 249 179; Email: irina.moreira@cnc.uc.pt

†Co-first authors.

Citation details: Matos-Filipe, P., Preto, A.J., Koukos, P.I. *et al.* MENSAdb: a thorough structural analysis of membrane protein dimers. *Database* (2021) Vol. 2021: article ID baab013; doi:10.1093/database/baab013

Received 8 May 2020; Revised 19 January 2021; Accepted 1 March 2021

Abstract

Membrane proteins (MPs) are key players in a variety of different cellular processes and constitute the target of around 60% of all Food and Drug Administration–approved drugs. Despite their importance, there is still a massive lack of relevant structural, biochemical and mechanistic information mainly due to their localization within the lipid bilayer. To help fulfil this gap, we developed the MEmbrane protein dimer Novel Structure Analyser database (MENSAdb). This interactive web application summarizes the evolutionary and physicochemical properties of dimeric MPs to expand the available knowledge on the fundamental principles underlying their formation. Currently, MENSAdb contains features of 167 unique MPs (63% homo- and 37% heterodimers) and brings insights into the conservation of residues, accessible solvent area descriptors, average *B*-factors, intermolecular contacts at 2.5 Å and 4.0 Å distance cut-offs, hydrophobic contacts, hydrogen bonds, salt bridges, π–π stacking, T-stacking and cation–π interactions. The regular update and organization of all these data into a unique platform will allow a broad community of researchers to collect and analyse a large number of features efficiently, thus facilitating their use in the development of prediction models associated with MPs.

Database URL: http://www.moreiralab.com/resources/mensadb.

Introduction

Membrane proteins (MPs) account for around 15–39% of the human proteome (1, 2). They assume a critical role in a vast set of cellular and physiological mechanisms, including molecular transport, nutrient uptake, toxin and waste product clearance, respiration and signalling (3).

While roughly 60% of all Food and Drug Administration (FDA)–approved drugs target MPs, there is a shortage of structural and biochemical data about them mainly due to their localization in the lipid bilayer (4, 5). In the last years, a daunting challenge of drug discovery has been the development of compounds that can target the 'undruggable' regions of MPs, enabling the modulation of protein–lipid, protein–nucleic acid and protein–protein interactions (PPIs) (6, 7). In this respect, being able to characterize the structural and physicochemical properties of MPs as well as their interactions and interfaces is essential to develop improved and more targeted therapies as well as to discover new drug targets. Particular features of proteins, such as electrostatic interactions (8), hydrophobic effects (9) or 'hot-spot' residues (10–13), were shown to contribute to the affinity and specificity of PPIs. Other well-characterized properties of proteins are the evolutionary conservation and distribution of their amino acids. These two features contribute the most to the prediction of functionally essential residues, as highlighted by several publications (14–17). While many studies have dealt with soluble systems, there is a significant lack of in-depth analysis of MP complexes and their interactions.

We present here the MEmbrane protein dimer Novel Structure Analyser database (MENSAdb), the first interactive web application exposing a comprehensive and thorough array of fundamental features of dimer surfaces of MPs and their interfacial regions. Users can easily access a thorough, systematic analysis of sequence–structure relationships (Figure 1) based on a curated database of 201 protein dimers obtained from the Membrane Proteins of Known 3D structure (MPSTRUC) (18). MENSAdb delivers tabular and graphical data formats that can be visually explored for a large number of MP features based on conservation, accessible solvent area (ASA) descriptors, average *B*-factors, intermolecular contacts at 2.5 Å and 4.0 Å distance cut-offs, hydrophobic contacts, hydrogen bonds, salt bridges, π–π stacking, T-stacking and cation–π interactions. Additionally, users can inspect differences in these features between three distinctive residue classes: (i) non-surface, (ii) surface and non-interfacial and (iii) surface and interfacial. The web server relies on a custom front-end application that provides the results to the user. The resulting knowledge and full database can be easily assessed and downloaded.

Our main goal with the integration of these features into a single platform is to assist the development of prediction models associated with MPs, either for classification or for regression tasks, as well as to help researchers to better understand MP interfacial characteristics. Our database is freely available at www.moreiralab.com/resources/mensadb.

Materials and methods

Data collection and pre-processing

Experimental structures of 167 unique transmembrane (TM) proteins that included β-barrel TMs and α-helix TMs were obtained from MPSTRUC (http://blanco.biomol.uci.edu/mpstruc/) (18). These correspond to structures achieved mainly from X-ray crystallography (91%) or electron microscopy (4%), with a resolution below or equal to 4.50 Å, and less frequently from nuclear magnetic resonance (5%). We discarded all non-TM, monomeric and monotopic (not embedded in the lipid bilayer) proteins. Pre-processing of the database was performed by excluding dimers in which one of the chains was a soluble protein, single MPs interacting with small soluble peptides (protein–peptide), pores, protein–antibodies (since antibodies are soluble proteins) and proteins with small organic or non-organic ligands (protein–ligand). In the previous case, the complex was maintained if the presence of more than one MPs chain was observed. Additionally, structures with unknown residues or with many incomplete amino acids were also excluded, as were structures with interfaces interacting highly with lipids. Sequences were filtered to ensure at most 35% sequence redundancy in each interface by using the PISCES web server (19). The final database was composed of 63% ($n = 105/167$) homodimers and 37% ($n = 62/167$) heterodimers. From the Protein Data Bank (PDB) files, all possible dimer combinations were extracted for the structures in which the number of chains was higher than two (functional high-order oligomers) and it is constituted by 201 protein dimer combinations (Supplementary File 1). The selected structures were then subjected to further processing. In particular, we (i) identified and removed residues outside the TM domain according to the MPSTRUC (18) annotation of α-helix and β-barrel amino acids available in the PDB (20) in conjunction with visual inspection; (ii) removed unnecessary heteroatoms; (iii) reversed mutated non-standard amino acids (e.g. selenomethionine was mutated to methionine); and (iv) added hydrogens to the structures. In-house PyMOL (20) and Visual Molecular Dynamics (VMD) scripts (21) were used to perform these pre-processing steps.

Definition of interfacial and non-interfacial residues

The relative solvent accessibility (RSA) defined as the ratio between an amino acid ASA value and its corresponding area in a Gly-X-Gly peptide was calculated using an in-house pipeline with Database of Secondary Structure

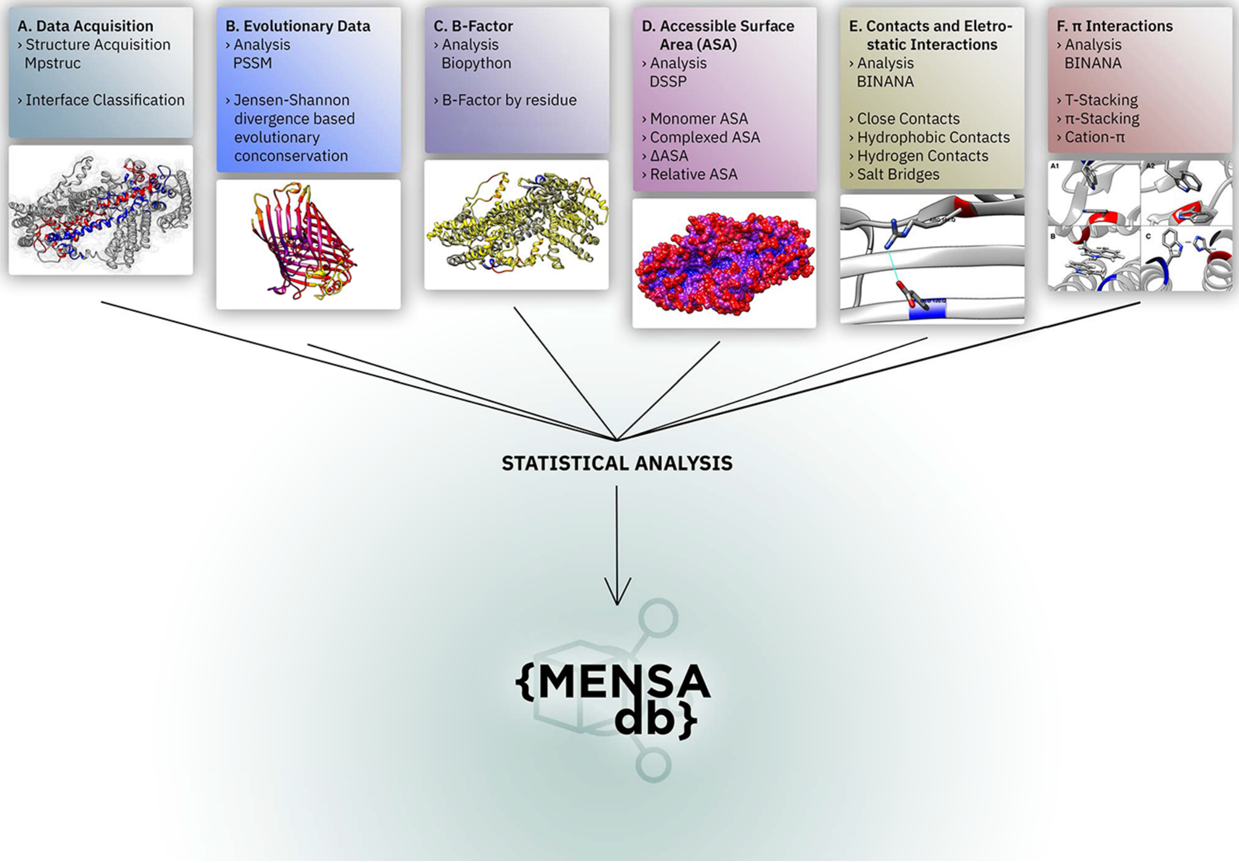


Figure 1. Overall representation of MENSAdb. Boxes A–F illustrate the steps involving the data collection, evolutionary conservation, *B*-factor, accessible surface and PPI analysis. Each box contains an example of the proteinic motifs under the scope of this work. (A) Interface between chains A and B of the STRA6 receptor for retinol uptake in *Danio rerio* (PDBid: 5SY1) (55). (B) Representation of evolutionary conservation of protein motifs (purple being more conserved and yellow less conserved) in the chain P of a hedgehog auto-processing domain in *Drosophila melanogaster* (PDBid: 1AT0) (56). (C) and (D) Average *B*-factor and complexed accessible surface area, respectively, of the chains A and B of 5SY1 (55). (E) Salt bridge between GLU120 and ARG161 of the chain Q of the sucrose-specific porin (PDBid: 1A0T) of *Salmonella typhimurium* (57). (F) The spectrum of π systems predicted: (A1 and A2) T-stacking motif between TRP25 (chain L) and TRP255 (chain M) from *Rattus norvegicus* S100B protein (PDBid: 1XYD) (58) is represented from two perspectives; (B) illustration of a π–π stacking structure between TRP262 (chain A) and TRP262 (chain B) from *Archaeoglobus fulgidus* CDP-alcohol phosphotransferase (PDBid: 4O6M) (59) and (C) cation–π interaction between HIS275 (chain B) and TRP175 (chain C) from *Escherichia coli* formate dehydrogenase-N (PDNid: 1KQF) (60).

assignments for all Proteins entries (DSSP) (22). Residues above a 0.20 RSA cut-off were considered as surface residues (23). We obtained 55 008 possible surface residues from a total of 91 861, while the remaining ones were considered core residues. Secondly, we considered those for which the pairwise distance between any atom of chain A and any atom of chain B was below 5 Å as interfacial residues, splitting surface residues into two classes: interfacial (15 277 residues) and non-interfacial ones (39 731 residues).

Determination of sequence and structural features of all residues

Evolutionary conservation of all sites was calculated using the Jensen–Shannon divergence (JSD) measure, a symmetrized and smoothed version of the Kullback–Leibler divergence (24), of the Position-Specific Scoring Matrix (PSSM), which itself was calculated with a local deployment of PSI-BLAST against the NCBI non-redundant database with parameters num_iterations = 3 and evalue = 0.001 (25). Equation 1 was used to quantify the similarity between two probability distributions and compares the amino acid distribution observed in PSSM p_{ia} with a background distribution f_a.

$$JSD = H\left(\frac{p_{ia}+p_a}{2}\right) - \frac{1}{2}H(p_{ia}) - \frac{1}{2}H(f_a) \qquad (1)$$

$H(\cdot)$ denotes the entropy of amino acid distribution. The code provided by Capra *et al.* was introduced into the pipeline due to its high performance in comparison with other methods (16). This metric works on the premise that the highest JSD value corresponds to a more conserved residue. We tested three different background distributions, BLOSUM62 (the PSI-BLAST default one), SLIM (26) and bbTM (27) to assess which one of them was the most suitable for MPs interface prediction. SLIM is a non-symmetric matrix optimized for TM protein segments, whereas bbTM is a set of matrices optimized for β-barrel proteins that uses three different matrices (one for intracellular segments, one for extracellular segments and another for TM residues). Herein, we only used the matrix developed for TM segments, since the remaining residues were already excluded from the analysis. We also generated a new column named 'appropriate JSD' in which we selected SLIM and bbTM depending on the presence and absence of an α-helix or β-barrel protein, respectively.

DSSP was used to calculate the RSA not only in the complexed form but also in the monomeric form, which were then multiplied by Sander and Rost amino acid constants (ALA: 106, ARG: 248, ASN: 157, ASP: 163, CYS:135, GLN: 198, GLU: 194, GLY: 84, HIS: 184, ILE: 169, LEU: 165, LYS: 205, MET: 188, PHE: 197, PRO: 136, SER: 130, THR: 142, TRP: 227, TYR: 222 and VAL: 142) (28) to calculate ASA of each amino acid, 'i', in the complexed ($_{\text{comp}}\text{ASA}_\text{i}$) and monomeric ($_{\text{mon}}\text{ASA}_\text{i}$) systems, respectively. These values were also used to calculate $\Delta\text{ASA}_\text{i}$ (Equation 2).

$$\Delta ASA_i = compASA_i - mon\ ASA_i \tag{2}$$

For further clarification, we also listed all $_{\text{rel}}\text{ASA}_\text{i}$ values (Equation 3), which allows the differentiation of residues with equal $\Delta\text{ASA}_\text{i}$ but with different absolute monomer ASA values (29–31).

$$rel\ ASA_i = \frac{\Delta ASA_i}{mon\ ASA_i} \tag{3}$$

We also extracted the temperature factor (*B*-factor) value for each residue from the PDB file of the analysed structures (obtained directly from MPSTRUC) using Biopython (32).

Determination of structural descriptors of MP–protein interface

Close and hydrophobic contacts, hydrogen bonds, salt bridges and π-interactions (π–π stacking, T-stacking and cation–π interactions) were described using BINANA—Binding Analyzer, a Python-implemented algorithm that characterizes protein complexes (33). Close contacts correspond to the number of pairs of atoms formed within 2.5 and 4.0 Å radius.

Data treatment

Since the composition of the database was not equally distributed across the three classes of MPs presented here, we defined a correction factor (C_{factor}), Equation 4, based on the concept of propensity score calculation, as shown by Huang (34). This factor is defined as the ratio between the frequency of occurrence of residue *i* in each one of the classes (fi_{CLAS}) and the frequency of occurrence of the total number of amino acids in that class (fi_{TOT}). The obtained MP-class-specific C_{factor} was used to correct the various metrics described in the 'Results' section by multiplying them by their respective C_{factor} except that of $_{\text{rel}}\text{ASA}$.

$$C_{factor} = \frac{fi_{CLAS}}{fi_{TOT}} \tag{4}$$

Statistics

For all plots, residues are ordered by increasing hydrophobicity based on the Kyte and Doolittle hydropathy index (35). Descriptive statistics such as three quartiles (Q1, Q2 and Q3), average and standard deviation were obtained using Pandas, a Python library (36). *P*-values were calculated through SciPy (https://docs.scipy.org/) with the independent *t*-test and one-way ANOVA. Further statistics were calculated for amino acids sets split according to the hydrophilic and hydrophobic potential as (i) charged—Asp, Glu, Lys and Arg; (ii) positively charged—Lys and Arg; (iii) negatively charged—Asp and Glu; (iv) polar—Ser, Thr, Asn, Gln, Tyr and His; (v) non-polar—Ala, Val, Ile, Leu, Met, Phe and Trp; aromatic—Phe, Trp and Tyr. Cys, Gly and Pro were not included in those subsets.

Code availability

MENSAdb code used for all the structural and physicochemical analyses of MP dimers is freely distributed as a GitHub repository at https://github.com/MoreiraLAB/mensadb-open. The available Python code allows users to perform feature extraction using a pre-processed PDB file easily. For detailed information on all the pre-processing steps (trimming of non-TM residues, removal of heteroatoms, mutation of exotic residues, modelling of incomplete structures and dimer extraction from the structure files), please see Preto *et al.* (37). The addition of hydrogens was implemented within the pipeline available in the GitHub repository. The original code was tested in a 64-bit version of Linux Ubuntu 18.04 (Intel Xeon 40 Core 2.2 GHz, 126 GB RAM) and required the installation of Python version 3.7.2 with the following free and open-source packages: NumPy ≥ 1.16.1, pandas ≥ 0.23.4, vmd-python ≥ 3.0.6, dit ≥ 1.2.3, Biopython ≥ 1.7.3 and standalone software: BLAST+ ≥ 2.9.0, BINANA ≥ 1.2.0, DSSP ≥ 3.0.7, MGTools ≥ 1.5.6 and AutoDock ≥ 3.0.7. The JSD measure we determined using a non-redundant protein database for comparison (for download options, please see https://ftp.ncbi.nlm.nih.gov/blast/db/).

Database development

Data resulting from this work are available through MENSAdb (www.moreiralab.com/resources/mensadb), without the need for login, registration or license, a rich data visualization web application built using Python's 'Flask'-based 'Dash' visualization framework (by 'Plotly'). MENSAdb's real-time query features are supported by a MongoDB back end, which enables the application to query, filter and aggregate the data in multiple meaningful ways. To boost performance, a 'Flask' caching layer is applied to support the complex queries required for visualization. To further ensure performance and security and support high-availability scenarios, all HTTP traffic directed at MENSAdb is served by the NGINX high-performance webserver and load balancer, which then routes it to multiple MENSAdb application instances. The final database of MENSAdb containing all the raw data of structural and physicochemical properties of MPs is publicly available from Figshare (Data Citation 1; dx.doi.org/10.6084/m9.figshare.7808909), and the full membrane dimer structures listed according to PDB code can be found in Supplementary File 1.

Results and discussion

MP dimer composition and characteristics

The overall residue distribution in Figure 2A and B shows that MPs have a higher content of hydrophobic and aromatic residues, such as leucine (13.2%), alanine (9.4%), valine (8.6%), glycine (8.4%), isoleucine (8.3%) and phenylalanine (6.9%) that account for 54.8% of all detected residues. For a better clarification the percentages presented in this sub-section, oppositely to remaining sub-sections are listed without correction factor. Indeed, these residues were shown to contribute the most to the accuracy of machine learning (ML) models developed for predicting protein–protein binding sites (38). This high content in hydrophobic residues, also previously reported in other studies (14, 38–43), is essential since it favours the thermodynamic interactions with the lipid bilayer. Figure 2A and B also show that GAS residues are significantly enriched at the MPs core (12.3%) and non-interfacial surface locations (8.5%), in comparison to interfacial surface (3.0%). These small residues are the strong driving force for membrane folding (44, 45). As expected, charged residues (arginine, aspartate, glutamate and lysine) are typically excluded from the MPs interface (surface: 7.4%; core: 2.6%; interface: 2.3%).

Evolutionary conservation of protein sequences is a key feature to better understand and characterize the functionally and structurally important residues at PPIs. Herein, we used three different background matrices to calculate conservation, namely BLOSUM62 (PSSM_JSD), SLIM and bbTM as well as the 'appropriate_JSD'. Figure 3 illustrates their distribution split into three different protein regions: core/non-surface, interfacial surface and non-interfacial

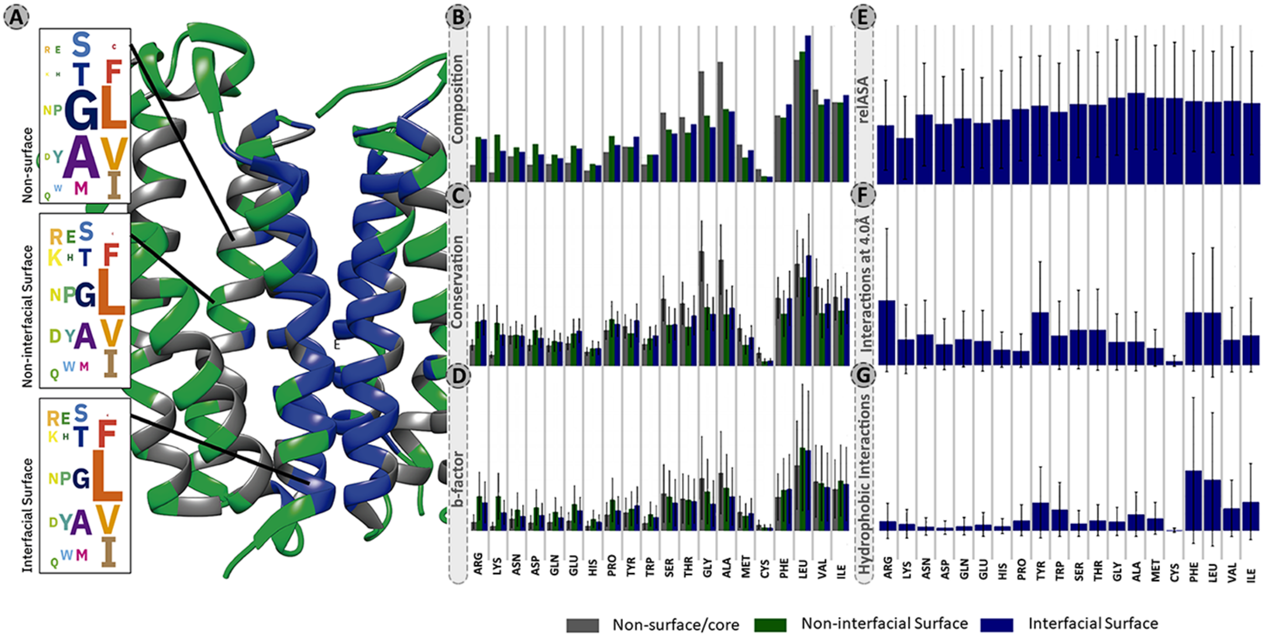


Figure 2. Panel of selected structural and physicochemical properties of MPs and their interactions. (A)—residue distribution of the translocator membrane protein (PDBid: 4UC1) from *Rhodobacter sphaeroides* (61). Amino acids are coloured according to the protein region within which they are embedded: grey—non-surface residues; green—non-interfacial surface residues; blue—interfacial surface residues. (B)—residue composition of the database. The correction factor described in section "Data treatment" of Material and methods was not applied here. (C)—normalized evolutionary conservation scores. (D)—normalized B-factor scores. (E)—normalized $_{rel}$ASA. (F)—normalized intermolecular contacts at 4 Å. (G)—normalized hydrophobic contacts.

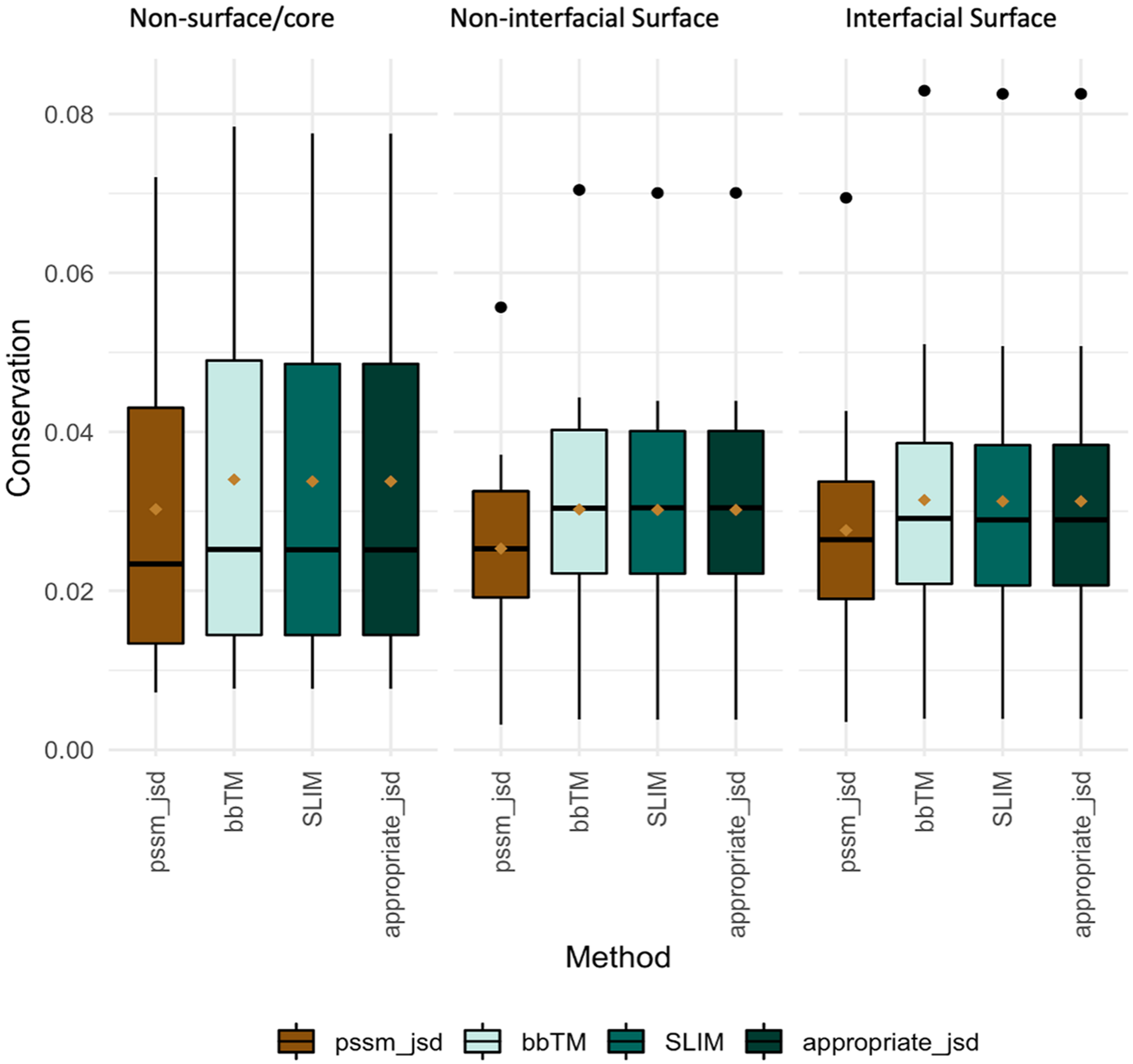


Figure 3. Conservation JSD distribution using BLOSUM62, SLIM, bbTM and the appropriate JSD background matrices (SLIM and bbTM were considered for α-helix and β-barrel proteins, respectively). Mean values are represented as a brown diamond. The results from the multiple pairwise test against all three background matrices yielded non-significant.

surface. The three different background matrices yielded similar results, which were non-significant according to multiple pairwise test. The same pattern was observed for all, with conservation being lower for surface, followed by interface and then protein core. As the used background matrix does not change the main conclusions about conservation at a MP dimer, we decided to follow up with the BLOSUM62 matrix for an easier implementation by the reader. Figure 2C reveals that for MPs, the more conserved JSD normalized values were found in the non-surface (0.05 ± 0.03) and in the interface (interface: 0.04 ± 0.02, surface non-interfacial 0.03 ± 0.02). The highest differences were for the GAS residues of the core region (core: 0.06 ± 0.03, surface: 0.03 ± 0.02; interface: 0.03 ± 0.02) and for the non-polar residues at protein core (core: 0.05 ± 0.02; surface: 0.04 ± 0.02; interface: 0.05 ± 0.03). These results, albeit not remarked different, support that the core and the interface are the most conserved regions, granting the necessary structural stability at specific PPIs, as previously observed (46). Additional results are available in the 'Conservation' option in the MENSAdb webserver.

B-factor (Figure 2D), related to the displacement of an atom from its reference position due to thermal motion and positional disorder (47), is typically used in a variety of applications including as a measure of atoms mobility for

PPIs prediction (48, 49). We observed a decrease in normalized average *B*-factor values of the interfacial residues compared to the non-interfacial surface ones (5.71 ± 6.10 Å^2 vs 6.25 ± 6.16 Å^2), putting their average closer to the non-surface MP residues (6.02 ± 5.69 Å^2). Also, positively charged residues are one of the most dissimilar ones (*B*-factor core: 1.19 ± 0.96 Å^2; *B*-factor surface: 5.34 ± 3.72 Å^2; *B*-factor interface: 3.74 ± 2.86 Å^2). This is in agreement with the fact that residues participating in PPIs are usually less flexible in comparison with the ones from the surface, which is reflected in lower *B*-factor values (49–51). Leucine, very conserved at the interface, seems to also have a higher mobility at PPI-associated locations (surface: 12.66 ± 9.86 Å^2; interface: 12.25 ± 9.52 Å^2 vs core: 9.88 ± 6.83 Å^2). Previous studies have suggested that leucine and isoleucine have an important role in flexible loop-mediated PPIs (52). Users can find illustrative plots of average *B*-factor values (by residue) in the 'Average *B*-factor' option in the MENSAdb web server.

The ASA descriptors detect protein regions that, when interacting or aggregating, lose solvent accessible area, while relASA indicates the relative exposed solvent surface area. MENSAdb and Figure 2E show that $_{rel}$ASA, which is the fraction of ΔASA by $_{mon}$ASA, is increased upon complex formation. These seem to be particularly relevant for non-polar residues (core: 5.27 ± 19.78 Å^2; surface: 0.00 ± 0.09 Å^2; interface: 52.01 ± 32.49 Å^2). Additional and detailed information about 'Monomer Accessible Surface Area' ($_{mon}$ASA), 'Complex Accessible Surface Area' ($_{comp}$ASA), 'Delta Accessible Surface Area' (ΔASA) and 'Relative Accessible Surface Area' ($_{rel}$ASA) can be viewed in MENSAdb web server options.

Characteristics of interfacial residues

Identification and characterization of critical features of membrane PPI dimers can provide important clues to pinpoint residues or interactions, important for drug development. For this, additional interfacial structural characteristics were quantified to better understand MP dimers. Concerning the intermolecular atomic contacts per amino acid type, we observed that the aromatic residues (Figure 2F, corrected contacts at 4 Å: 0.56 ± 0.61) are much more prone to establish close contacts at short distance than other residues. Arg was also highlighted in our results (corrected contacts at 4 Å: 0.75 ± 0.82). For further information, check the 'Interactions at 2.5 Angstroms' and 'Interactions at 4.0 Angstroms' options in the MENSAdb web server.

Hydrophobicity involving large aromatic residues is key in MP dimers and aromatic residues, and non-polar residues show a high number of hydrophobic contacts (Figure 2G, aromatic: 0.25 ± 0.34 and non-polar: 0.23 ± 0.32). In particular, Phe and Tyr establish π–π stacking, T-stacking and cation–π interactions in different dimers. Cation–π interactions are also particularly relevant for Arg (for a closer detailed view, please see the 'Hydrophobic Interactions', 'Pi–Pi Interactions', 'T-Stacking Interactions' and 'Cation–Pi Interactions' options in the MENSAdb).

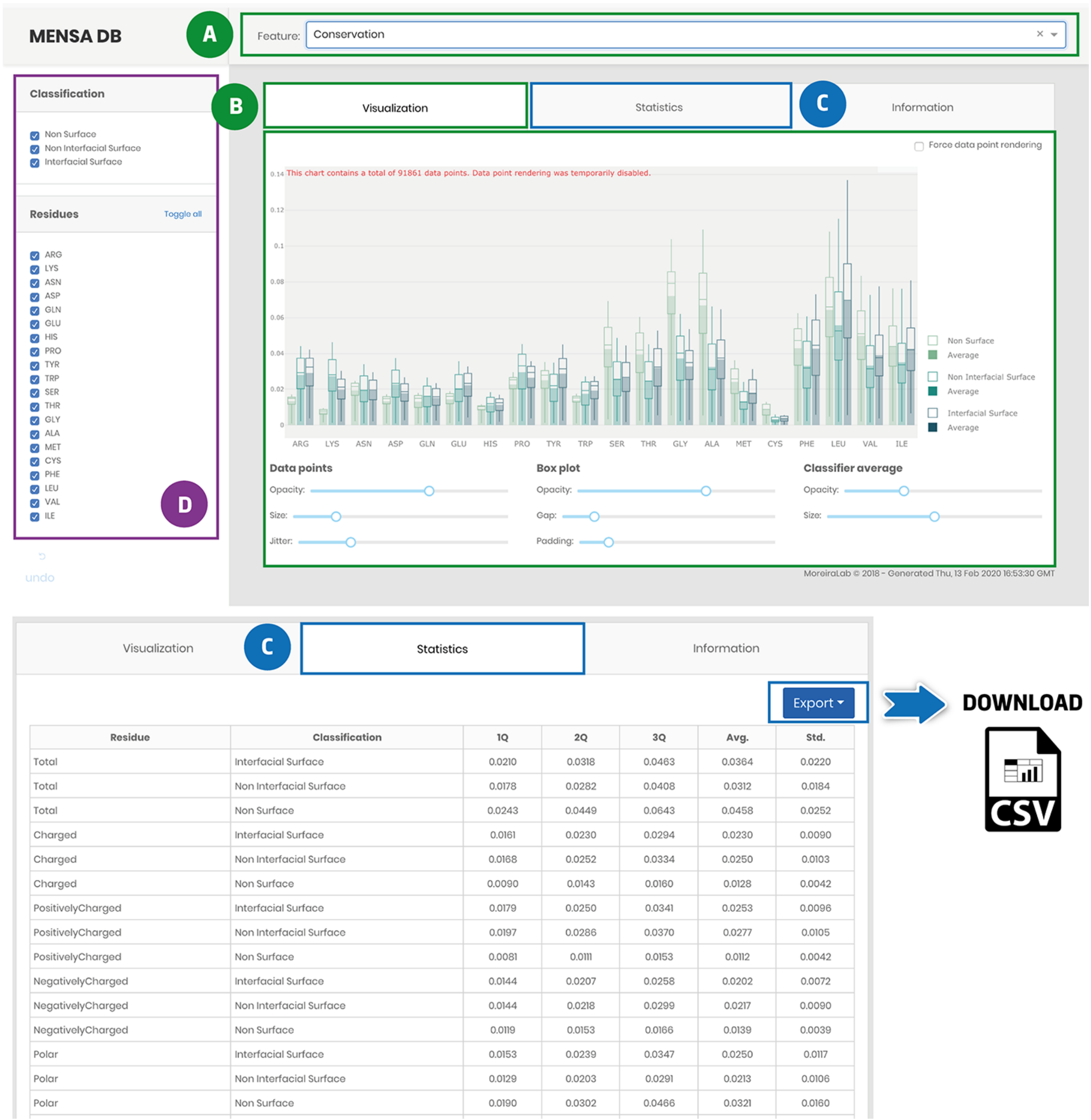
Additionally, although MP residues reside in a non-polar (low dielectric) environment (8, 53), both salt bridges between charged residues and hydrogen bonds through almost all amino acids are common to stabilize the interface and promote complex formation. Hydrogen bonds measured here involving both side chains and backbone are particularly important not only for charged residues (0.01 ± 0.03) but also for aromatic ones (0.01 ± 0.02), in particular tyrosine (0.01 ± 0.03) and tryptophan (0.01 ± 0.01). For a closer detailed view, please see the 'Salt-bridge Interactions' and 'Hydrogen-bond Interactions' options in the MENSAdb web server.

All different values presented herein showed statistical relevance.

All the results presented herein were obtained under the assumption that the interfaces in this study were biologically relevant, and utmost care was taken to ensure this (Supplementary File 1). Further limitations could arise from possible crystallographic artefacts.

MENSAdb interface and usability

The developed application enables users to explore the MP-dimer database (Figure 4). Access to evolutionary and physicochemical features is provided through a drop-down menu on the main page (Figure 4A). The data are presented in downloadable box plots for visual inspection that can be easily changed, for example, by filtering, zooming or panning (Figure 4B). Besides, data-associated statistics are also accessible in a tabular format [Q1, Q2, Q3, Average (Avg.) and Standard Deviation (Std.)] (Figure 4C). Stats and raw data can be downloaded as a .csv file using the export button for further reuse and integration in other studies. Users can also filter data for each selected feature by classification (non-surface, non-interfacial surface and interfacial surface) or residue type (Figure 4D). The database also has an 'Information' tab with general information for each included feature and a brief description of the underlying methods for their acquisition and pre-processing to help first-time users. MENSAdb will continue to be updated at least annually, and we expect, shortly, to integrate a new model for the prediction of MP interfaces.



This chart contains a total of 91861 data points. Data point rendering was temporarily disabled.

MoreiraLab © 2018 - Generated Thu, 13 Feb 2020 16:53:30 GMT

Residue	Classification	1Q	2Q	3Q	Avg.	Std.
Total	Interfacial Surface	0.0210	0.0318	0.0463	0.0364	0.0220
Total	Non Interfacial Surface	0.0178	0.0282	0.0408	0.0312	0.0184
Total	Non Surface	0.0243	0.0449	0.0643	0.0458	0.0252
Charged	Interfacial Surface	0.0161	0.0230	0.0294	0.0230	0.0090
Charged	Non Interfacial Surface	0.0168	0.0252	0.0334	0.0250	0.0103
Charged	Non Surface	0.0090	0.0143	0.0160	0.0128	0.0042
PositivelyCharged	Interfacial Surface	0.0179	0.0250	0.0341	0.0253	0.0096
PositivelyCharged	Non Interfacial Surface	0.0197	0.0286	0.0370	0.0277	0.0105
PositivelyCharged	Non Surface	0.0081	0.0111	0.0153	0.0112	0.0042
NegativelyCharged	Interfacial Surface	0.0144	0.0207	0.0258	0.0202	0.0072
NegativelyCharged	Non Interfacial Surface	0.0144	0.0218	0.0299	0.0217	0.0090
NegativelyCharged	Non Surface	0.0119	0.0153	0.0166	0.0139	0.0039
Polar	Interfacial Surface	0.0153	0.0239	0.0347	0.0250	0.0117
Polar	Non Interfacial Surface	0.0129	0.0203	0.0291	0.0213	0.0106
Polar	Non Surface	0.0190	0.0302	0.0466	0.0321	0.0160

Figure 4. Main landing page of MENSAdb web server. Screenshot of the home page (A)—quickly query by evolutionary or physicochemical features. (B)—In the visualization tab, the results are shown in a graphical format. Users can easily change visual properties (opacity, size, jitter, gap and padding) by interacting with the lower panel. (C)—Statistics tab displays the data in a tabular format with associated metrics (Q1, Q2, Q3, Average-Avg. and Standard Deviation-Std.). Stats and raw data can be downloaded using the Export button in the top right corner, as a .csv file. (D)—In the left panel, users can filter graphic data by classification and residue type.

MENSAdb is the first comprehensive resource dedicated explicitly to exposing the evolutionary and physicochemical features of dimeric MP structures. Our main goal with the integration of these features into a single platform is to assist the development of experimental and computational assays, relevant for a better understanding of dimeric MP interactions and interfaces of this largest but poorly studied type of proteins. In the last years, some studies used evolutionary and physicochemical properties similar to the ones provided in our database to train ML for the prediction of MP complex binding sites (38, 46, 54). Nevertheless, as far as we know, herein we offer original features such as the ones from membrane PPI analysis not yet used or provided by other databases more dedicated to MP structures (PDBTM, OPM, MemProtMD and, MPSTRUC) or classification (TCDB).

Supplementary data

Supplementary data are available at *Database* online.

Acknowledgements

A.J.P. was also supported by FCT through PhD scholarship SFRH/BD/144966/2019. I.S.M. was funded by the FCT Investigator programme—IF/00578/2014 (co-financed by the European Social Fund and Programa Operacional Potencial Humano). A.M.J. and P.I.K acknowledge funding from the Dutch Foundation for Scientific Research (NWO) (TOP-PUNT grant 718.015.001). The authors would also like to acknowledge ERNEST (European Research Network on Signal Transduction, CA18133) and STRATAGEM (New diagnostic and therapeutic tools against multidrug-resistant tumours, CA17104).

Funding

COMPETE 2020—Operational Programme for Competitiveness and Internationalisation—and Portuguese national funds via FCT—Fundação para a Ciência e a Tecnologia—under projects POCI-01-0145-FEDER-031356, PTDC/QUI-OUT/32243/2017 and UIDB/04539/2020.

Author contributions

P.M.F., A.J.P, P.I.K. and I.S.M performed the acquisition of data. P.M.F., A.J.P, J.M. and I.S.M. processed the data. I.S.M. and A.M.J. conceived the study. All authors wrote and approved the final version of the manuscript.

Conflict of interest. None declared.

Data citation

Matos-Filipe,P., Preto,A.J., Koukos,P.I., Mourão,J., Bonvin,A.M.J.J., Moreira,I.S. *figshare* dx.doi.org/10.6084/m9.figshare.7808909.

References

1. Almén,M.S., Nordström,K.J.V., Fredriksson,R. *et al.* (2009) Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol.*, **7**, 50.
2. Michael Gromiha,M. and Ou,Y.Y. (2014) Bioinformatics approaches for functional annotation of membrane proteins. *Brief. Bioinformatics*, **15**, 155–168.
3. Tan,S., Tan,H.T. and Chung,M.C.M. (2008) Membrane proteins and membrane proteomics. *Proteomics*, **8**, 3924–3932.
4. Overington,J.P., Al-Lazikani,B. and Hopkins,A.L. (2006) How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.
5. Yıldırım,M.A., Goh,K.-I., Cusick,M.E. *et al.* (2007) Drug-target network. *Nat. Biotechnol.*, **25**, 1119–1126.
6. Yin,H. and Flynn,A.D. (2016) Drugging membrane protein interactions. *Annu. Rev. Biomed. Eng.*, **18**, 51–76.
7. Feng,Y., Wang,Q. and Wang,T. (2017) Drug target protein-protein interaction networks: a systematic perspective. *Biomed. Res. Int.*, **2017**, 1289259.
8. Zhang,Z., Witham,S. and Alexov,E. (2011) On the role of electrostatics on protein-protein interactions. *Phys. Biol.*, **8**, 035001.
9. Chanphai,P., Bekale,L. and Tajmir-Riahi,H.A. (2015) Effect of hydrophobicity on protein-protein interactions. *Eur. Polym. J.*, **67**, 224–231.
10. Darnell,S.J., LeGault,L. and Mitchell,J.C. (2008) KFC server: interactive forecasting of protein interaction hot spots. *Nucleic Acids Res.*, **36**, 265–269.
11. Moreira,I.S., Fernandes,P.A. and Ramos,M.J. (2007) Hot spots–a review of the protein-protein interface determinant amino-acid residues. *Proteins Struct. Funct. Bioinform.*, **68**, 803–812.
12. Moreira,I.S., Koukos,P.I., Melo,R. *et al.* (2017) SpotOn: high accuracy identification of protein-protein interface hot-spots. *Sci. Rep.*, **7**, 1–11.
13. Rosell,M. and Fernández-Recio,J. (2018) Hot-spot analysis for drug discovery targeting protein-protein interactions. *Expert. Opin. Drug. Discov.*, **13**, 327–338.
14. Ulmschneider,M.B. and Sansom,M.S.P. (2001) Amino acid distributions in integral membrane protein structures. *Biochimica Et Biophysica Acta Biomembranes*, **1512**, 1–14.
15. Caffrey,D.R. (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.*, **13**, 190–202.
16. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
17. Zhang,Q.C., Petrey,D., Norel,R. *et al.* (2010) Protein interface conservation across structure space. *Proc. Natl. Acad. Sci.*, **107**, 10896–10901.
18. White,S.H. (2009) Biophysical dissection of membrane proteins. *Nature*, **459**, 344–346.
19. Wang,G. and Dunbrack,R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
20. DeLano,W.L. (2015) The PyMOL molecular graphics system. Version 2.2 Schrödinger, LLC.
21. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
22. Touw,W.G., Baakman,C., Black,J. *et al.* (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.*, **43**, D364–D368.
23. Lins,L., Thomas,A. and Brasseur,R. (2003) Analysis of accessible surface of residues in proteins. *Protein Sci.*, **12**, 1406–1417.
24. Lin,J. (1991) Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, **37**, 145–151.
25. Altschul,S.F., Madden,T.L., Schäffer,A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
26. Müller,T., Rahmann,S. and Rehmsmeier,M. (2001) Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics (Oxford, England)*, **17**, S182–S189.
27. Jimenez-Morales,D., Adamian,L. and Liang,J. (2008) Detecting remote homologues using scoring matrices calculated from the estimation of amino acid substitution rates of beta-barrel membrane proteins. *Annu. Int. Conf. IEEE Eng. Med. Biol Soc.*, **2008**, 1347–1350.

28. Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins Struct. Funct. Bioinform.*, **20**, 216–226.
29. Melo,R., Fieldhouse,R., Melo,A. *et al.* (2016) A machine learning approach for hot-spot detection at protein-protein interfaces. *Int. J. Mol. Sci.*, **17**, 1–14.
30. Munteanu,C.R., Pimenta,A.C., Fernandez-Lozano,C. *et al.* (2015) Solvent accessible surface area-based hot-spot detection methods for protein-protein and protein-nucleic acid interfaces. *J. Chem. Inf. Model*, **55**, 1077–1086.
31. Martins,J.M., Ramos,R.M., Pimenta,A.C. *et al.* (2014) Solvent-accessible surface area: how well can be applied to hot-spot detection? *Proteins Struct. Funct. Bioinform.*, **82**, 479–490.
32. Cock,P.J.A., Antao,T., Chang,J.T. *et al.* (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
33. Durrant,J.D. and McCammon,J.A. (2011) BINANA: a novel algorithm for ligand-binding characterization. *J. Mol. Graph. Model.*, **29**, 888–893.
34. Huang,H.L. (2014) Propensity scores for prediction and characterization of bioluminescent proteins from sequences. *PLoS ONE*, **9**, e97158.
35. Kyte,J., Doolittle,R.F., Diego,S. *et al.* (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
36. McKinney,W. (2010) Data structures for statistical computing in python. In: Van Der Walt S, Millman J (eds). *Proceedings of the 9th Python in Science Conference*. Austin, Texas, USA, pp. 51–56.
37. Preto,A., Matos-Filipe,P., Koukos,P. *et al.* (2019) Structural characterization of membrane protein dimers. In: Kister AE (ed). *Protein Supersecondary Structures. Methods in Molecular Biology*. Humana Press, New York, NY, pp. 403–436.
38. Bordner,A.J. (2009) Predicting protein-protein binding sites in membrane proteins. *BMC Bioinform.*, **10**, 312.
39. Eilers,M., Patel,A.B., Liu,W. *et al.* (2002) Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys. J.*, **82**, 2720–2736.
40. Saidijam,M., Azizpour,S. and Patching,S.G. (2018) Comprehensive analysis of the numbers, lengths and amino acid compositions of transmembrane helices in prokaryotic, eukaryotic and viral integral membrane proteins of high-resolution structure. *J. Biomol. Struct. Dyn.*, **36**, 443–464.
41. Mbaye,M.N., Hou,Q., Basu,S. *et al.* (2019) A comprehensive computational study of amino acid interactions in membrane proteins. *Sci. Rep.*, **9**, 12043.
42. Duarte,J.M., Biyani,N., Baskaran,K. *et al.* (2013) An analysis of oligomerization interfaces in transmembrane proteins. *BMC Struct. Biol.*, **13**, 1–11.
43. Yan,C., Wu,F., Jernigan,R.L. *et al.* (2008) Characterization of protein-protein interfaces. *Protein J.*, **27**, 59–70.
44. Zhang,S.Q., Kulp,D.W., Schramm,C.A. *et al.* (2015) The membrane- and soluble-protein helix-helix interactome: similar geometry via different interactions. *Structure*, **23**, 527–541.
45. Zhang,Y., Kulp,D.W., Lear,J.D. *et al.* (2009) Experimental and computational evaluation of forces directing the association of transmembrane helices. *J. Am. Chem. Soc.*, **131**, 11341–11343.
46. Li,B., Mendenhall,J. and Meiler,J. (2019) Interfaces between alpha-helical integral membrane proteins: characterization, prediction, and docking. *Comput. Struct. Biotechnol. J.*, **17**, 699–711.
47. Trueblood,K.N., Bürgi,H.B., Burzlaff,H. *et al.* (1996) Atomic dispacement parameter nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature. *Acta. Crystallogr. A Found Crystallogr.*, **52**, 770–781.
48. Chung,J.-L., Wang,W. and Bourne,P.E. (2005) Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins*, **62**, 630–640.
49. Liu,R., Jiang,W. and Zhou,Y. (2010) Identifying protein–protein interaction sites in transient complexes with temperature factor, sequence profile and accessible surface area. *Amino Acids*, **38**, 263–270.
50. Chakravarty,D., Janin,J., Robert,C.H. *et al.* (2015) Changes in protein structure at the interface accompanying complex formation. *IUCrJ*, **2**, 643–652.
51. Jones,S. and Thornton,J.M. (1995) Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.*, **63**, 31–65.
52. Gavenonis,J., Sheneman,B.A., Siegert,T.R. *et al.* (2014) Comprehensive analysis of loops at protein-protein interfaces for macrocycle design. *Nat. Chem. Biol.*, **10**, 716–722.
53. Lomize,A.L., Pogozheva,I.D., Lomize,M.A. *et al.* (2007) The role of hydrophobic interactions in positioning of peripheral proteins in membranes. *BMC Struct. Biol.*, **7**, 1–30.
54. Zeng,B., Hönigschmid,P. and Frishman,D. (2019) Residue co-evolution helps predict interaction sites in α-helical membrane proteins. *J. Struct. Biol.*, **206**, 156–169.
55. Chen,Y., Clarke,O.B., Kim,J. *et al.* (2016) Structure of the STRA6 receptor for retinol uptake. *Science*, **353**, aad8266.
56. Hall,T.M.T., Porter,J.A., Young,K.E. *et al.* (1997) Crystal structure of a hedgehog autoprocessing domain: homology between hedgehog and self-splicing proteins. *Cell*, **91**, 85–97.
57. Forst,D., Welte,W., Wacker,T. *et al.* (1998) Structure of the sucrose-specific porin ScrY from *Salmonella typhimurium* and its complex with sucrose. *Nat. Struct. Biol.*, **5**, 37–46.
58. Wilder,P.T., Varney,K.M., Weiss,M.B. *et al.* (2005) Solution structure of zinc- and calcium-bound rat S100B as determined by nuclear magnetic resonance spectroscopy. *Biochemistry*, **44**, 5690–5702.
59. Sciara,G., Clarke,O.B., Tomasek,D. *et al.* (2014) Structural basis for catalysis in a CDP-alcohol phosphotransferase. *Nat. Commun.*, **5**, 4068.
60. Jormakka,M., Törnroth,S., Byrne,B. *et al.* (2002) Molecular basis of proton motive force generation: structure of formate dehydrogenase-N. *Science*, **295**, 1863–1868.
61. Li,F., Liu,J., Zheng,Y. *et al.* (2015) Crystal structures of translocator protein (TSPO) and mutant mimic of a human polymorphism. *Science*, **347**, 555–558.