TransINT: an interface-based prediction of membrane protein-protein interactions

Khazen, G.[1,*], Gyulkhandanian, A.[2], Issa, T.[1] and Maroun, R.C. [2,*]

[1] Computer Science and Mathematics Department, Lebanese American University, Byblos-Lebanon

[2] Inserm U1204/ Université d'Evry/Université Paris-Saclay/

Structure-Activité des Biomolécules Normales et Pathologiques

91025 Evry

France.

ABSTRACT

Membrane proteins account for about one-third of the proteomes of organisms and include structural proteins, channels, and receptors. The mutual interactions they establish in the membrane play crucial roles in organisms, as they are behind many processes such as cell signaling and protein function. Because of their number and diversity, these proteins and their macromolecular complexes, along with their physical interfaces represent potential pharmacological targets par excellence for a variety of diseases, with very important implications for the design and discovery of new drugs or peptides modulating the interaction. Yet, structural data coming from experiments is very scarce for this family of proteins and their study at the molecular level goes necessarily through extraction from cell membranes, a step

that reveals itself noxious for their structure and function. To overcome this problem, we propose a computational approach for the prediction of alpha-helical transmembrane protein higher-order structures through data mining, sequence analysis, motif search, extraction, identification and characterization of the amino acid residues at the interface of the complexes. This method leads us to the formulation of protein-protein binding sites that we use to scan protein sequence datasets for generating new potential interacting protein couples. Our template motif-based approach using experimental recognition sites leads us to predict thousands of binary complexes between membrane proteins. We generate an online database of the annotated predicted interactions that can be accessed on https://transint.shinyapps.io/transint/.

Key words: protein-protein interactions; bioinformatics; biostatistics; integral membrane proteins; computational biology; prediction method; binary interactions; molecular recognition; protein complexes

INTRODUCTION

As proteins are the core of the cell machinery in organisms and encompass a large variety of functions, including but not limited to, activation, transport, degradation, stabilization, and participation in the production of other proteins. Since their complexes are the functional units of cells, it is fundamental to understand the interactions between them. In other words, knowledge of the 3D structure of the proteins involved and of the interfaces needed for complex formation is a fundamental phenomenon governing all processes of life. [1]

The membrane protein interactome is the complete set of direct interactions that take place among integral-to-the-membrane proteins (single- and multiple-spanning). These transmembrane proteins (TM) span the cell membrane in its entirety and constitute the membranome, i.e. the proteome of the membrane. TMs represent around one third of the proteomes of organisms [2] and in eukaryotes come mostly in the form of alpha-helical domains that cross different types of cell membranes. They include many important enzymes, receptors, and transporters (about 3×10^5 in number, excluding polymorphisms or rare mutations). The number of possible binary and multiple interactions between them is thus vastly larger.

TM proteins represent potential pharmacological targets par excellence in human health and disease because they include many families involved in protein-protein interaction (PPI) networks, leading to many different physiological processes (signal transduction, apoptosis…). Thus, TM interactions, through the lipid-embedded domains, lead to oligomer formation and guide the assembly and function of many cell proteins and receptors. In addition, the assembly of TM proteins may lead to emergent properties, a relationship existing between oligomerization and function. Indeed, deficient oligomerization is associated with known diseases. [3–5]

Estimates of the total number of human PPIs range from 130,000 to 600,000 [6–8] to 3,000,000 [9], several orders of magnitude bigger than the *D melanogaster* interactome. High-

throughput experimental and theoretical approaches are being used to build PPI networks and thus elucidate the rules that govern these interactions. However, traditional techniques like yeast two-hybrid (Y2H) assays [10] are not well suited for identifying membrane protein interactions. In addition, the data covering PPIs is increasing exponentially -in year 2012, there were more than 235,000 binary interactions reported. [11] Most protein interaction databases (DB) – bioGRID, [12] BIND, [13] STRING, [14] IntAct, [15] MINT, [11] MENTHA[16] and others - offer general information about experimentally validated PPIs of all types. The IMEx Consortium groups all nonredundant protein interaction data in one interface. [17] Nevertheless, these DBs are mostly concerned with water-soluble globular proteins and the juxtamembrane interactions of TM proteins. But, unlike globular proteins, TM proteins are water-insoluble and for technical reasons their large-scale investigation has lagged behind. [18]

Proteome-wide maps of the human interactome network have been generated in the past. [19–21] Nevertheless, these assays, just like the Y2H assay, are depleted of membrane proteins or unreliable.[22] And even though a new biochemical technique has been developed using a split-ubiquitin membrane two-hybrid (MYTH) system, so far only a very limited number of membrane complexes have been determined using it. [19,20] This procedure has been further extended with a mammalian-membrane two-hybrid assay (MaMTH) technique for the identification and characterization of the interaction partners of integral membrane proteins under desired conditions. [23] But to the best of our knowledge, MaMTH has not been used as a systematic screening assay to map the human membrane protein interactome. On the other hand, using a variety of machine learning classifiers there are several methods for the prediction of PPIs from: i) sequence alone (PIPE2; [24] SigProd; [25] MULTIPROSPECTOR; [26] PCA-EELM; [27] MCDPPI; [28] amino acid composition; [29]; and ii) from template structures are available. For a review on the interactome, see. [30]

But, again, most of the approaches address soluble globular proteins only.[31] Qi et al developed a random forest learning machine approach but limited to the human membrane receptor interactome, HMRI. [32] On another hand, a*b-initio* prediction of membrane protein interfaces is rendered difficult as these have amino acid compositions that are not very different from the rest of the protein surface, decorated by hydrophobic residues in the membrane-exposed surface.

To circumvent this problem, we developed a knowledge-based predictive approach based on the detection of alpha-helical transmembrane contact residues issued from membrane protein multimers, or their portions thereof reported in the PDB [33] and validated by the OPM DB, [34] in the context of the lipid bilayer. Querying thereafter the PDBsum DB, [35,36] we obtained the atomic details of the protein-protein interfaces, i.e. the residues that are in contact at the interface of the complexes. We then gathered those amino acids at the recognition interface to generate regular expressions, or patterns, that represent in a linear form the interaction motifs in space, including wildcard regions in addition to the interface residues. With this information in hand, we proceeded to find the obtained motifs in other TM proteins by allowing a certain degree of mutations in the interface residues of the motifs. The permitted mutation rates of the interface residues allowed us to obtain local, spatial non-homologous interface sequences. Our assumption is that homologs of the template motifs are expected to interact in an analogous manner, [37] as opposed to a global approach that would limit the search to functionally-related partners without paying attention to the particular sequence at the interface. In all cases, it is reasonable to assume that the number of interface motifs is limited in nature. [38] Thus, our approach is focused on the structural and sequence homology of the recognition site residues found at the membrane-embedded interfaces of macromolecular complexes of the TM proteins, and not in the overall sequence homology, such as in other template-based predictions.[39,40] In other words, we conduct binding-site sequence homology-based inferences of pairwise PPIs.

As such, our method includes implicitly the spatial arrangement of these residues, but it expresses the relationship as a regular expression representing a linear interaction motif. The linear 1D motifs we obtain represent thus 3D epitopes. Given that membrane proteomes may vary between species, tissues, cell types, metabolic pathways and developmental stages, in this work we focus on the plasma membrane of eukaryotes, ensuring thus that we are probing proteins with the same sub-cellular localization.

RESULTS

Fully automated pipeline to predict and analyze hundreds of novel membrane protein-protein interactions (MPPI)

UniProt provided us with 13,352 eukaryote plasma membrane TM proteins that have the GO annotations mentioned in the M&M section. Overall, these proteins mapped to 9845 distinct oligomer TM PDB structures. As we were only interested in structures that satisfy the requirements signified in the M&M section, we validated and used 50 PDB files of membrane protein-protein complexes. After checking which corresponding PDBsum files to consider, we ended up with 53 distinctly different template interactions, associated to 48 unique reviewed UniProt entries across species, of which 50 are structural homomers and 3 structural heteromers (Table S1). The set includes X-ray, solution NMR and electron microscopy structures. The TM proteins include classes such as receptor tyrosine kinases, TLRs, ion channels, Cys-Loop and immune receptors, gap junctions, transporters, and GPCRs. Of these, besides *H. sapiens*, a taxonomically diverse set of organisms across the eukaryote branch of the tree of life are represented, such as *A. thaliana*, *B. taurus*, *G. gallus*, *M. musculus*, *O. sativa*, *R. norvegicus*, and *S. cerevisiae*. On another hand, there are 21 structures of complexes between bitopic proteins, 32 between polytopic proteins, and 1 mixed bitopic-polytopic protein complex. The oligomerization order is well represented, with 23 homo 2-mers, five homo 3-mers, 12 homo

4-mers, one hetero 4-mer, four homo 5-mers, three hetero 6-mers, one hetero 10-mer, and one homo 12-mer. The following EC numbers are represented in this set: 2 (ten transferases); 3 (one hydrolases) and 7 (one translocase). When verifying the protein-protein interfaces in the complexes with EPPIC or PRODIGY for the type of assembly they form (crystallographic or biological), we found that all the X-ray or electron microscopy complexes are classified as biological (Table S2). The NMR-determined complexes are not submitted to the test, since they are formed in solution as monomers. This is thus the number of experimentally solved structures of TM protein complexes that we used as the template DB.

The total number of motifs found after removing redundancies due to different chains of the same structures interacting more than once was of 98 (Table S3). Moreover, we observed that some amino acid residues were more favored than others in the TM recognition sites. For instance, the hydrophobic side chains Leu, Ile, Val and Phe were the most common, with Leu being found more than 300 times, making about more than a third of all contact residues (Fig. 1). The physicochemical properties of TM PPI binding sites are therefore different from those of soluble proteins. The amino acid residue abundancies we found in the motifs match those reported by Mayol et al. [41] Fig. 2 shows, as a symmetric "heat map" the couples of contact residues at the interface for our template set, as they come out from PDBsum. These residues are exposed to the lipid bilayer. We can see that the largest value corresponds to the contact between two Leu residues, followed by contact between Phe residues, and then the Leu-Phe interaction. The least observed interactions include, for example His-His for a homotype pair, and Trp-Cys for a heterotype pair. This outcome leads us to conclude that residues tend to contact other residues sharing the same physicochemical properties, and agrees with the statistics obtained for inter-residue interactions in the TM bundles DB for alpha- helical MPs. [41] In addition, these statistical trends in the contacts imply correlated mutations.

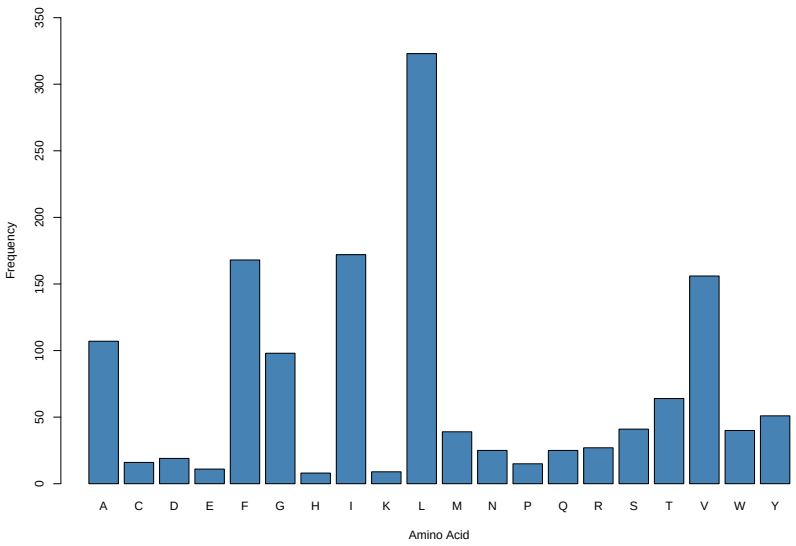


Fig. 1 Frequency of amino acids in the extracted motifs from the contact maps at the interfaces obtained from the PDBsum server.

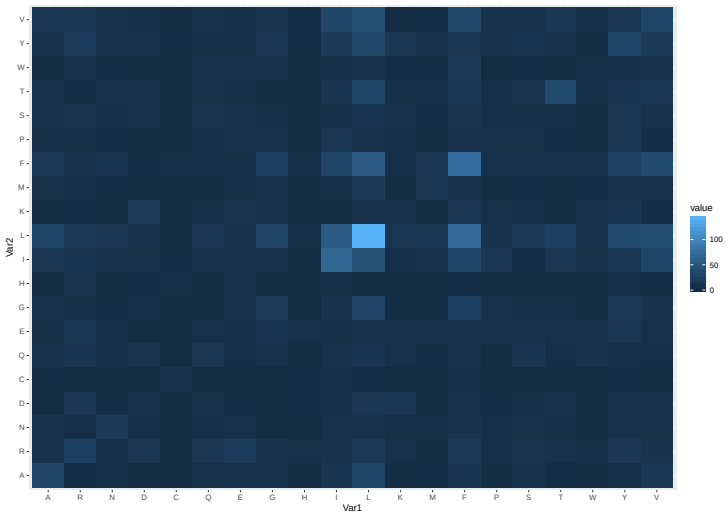
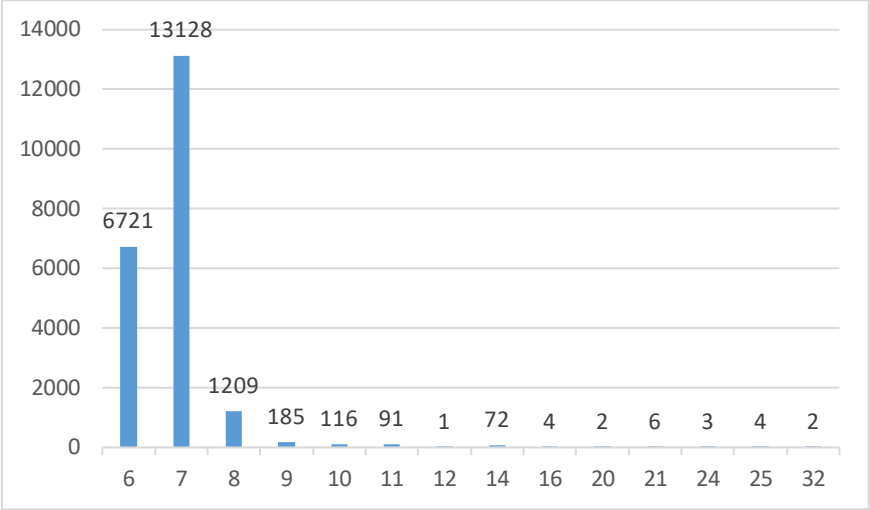
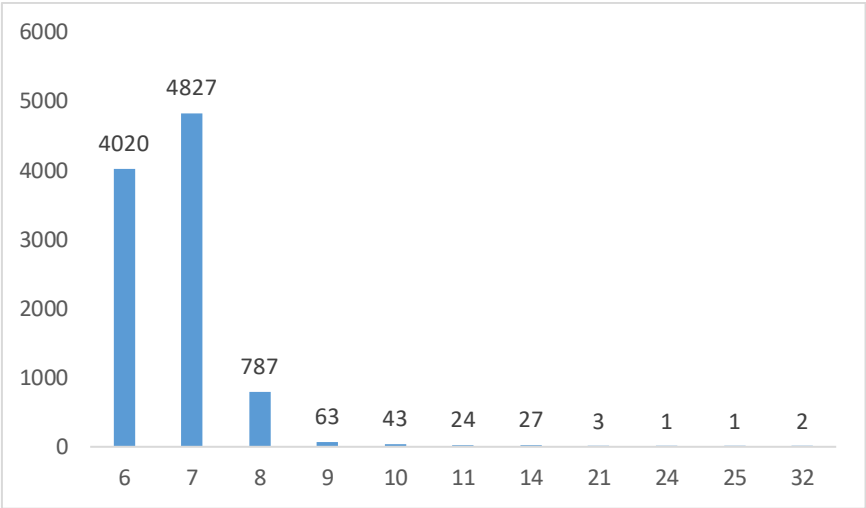


Fig. 2 Symmetrical heat map of pairs of contact residues at the MP-MP interface of the template complexes. Amino acid residue names are represented by the one-letter code.

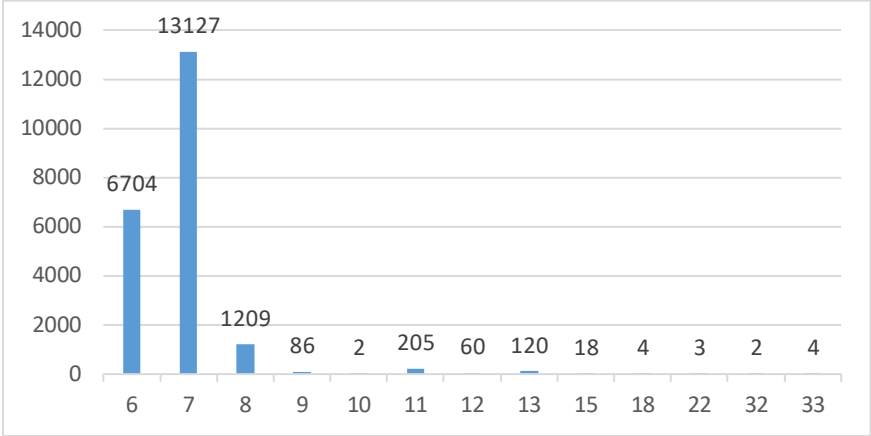
We then wanted to look at the number of motifs resulting for each number n of contact residues for the species recorded in the PDB and UniProt, as well as for *H. sapiens.* As seen in Fig. 3a and 3b, the count occurrence of contact residues is largest for n = 7: 13128 for motifs belonging to protein A of the interacting pair, and 13127 for motifs of protein B (4827 and 4828 for *H. sapiens*, respectively).

a)





b)



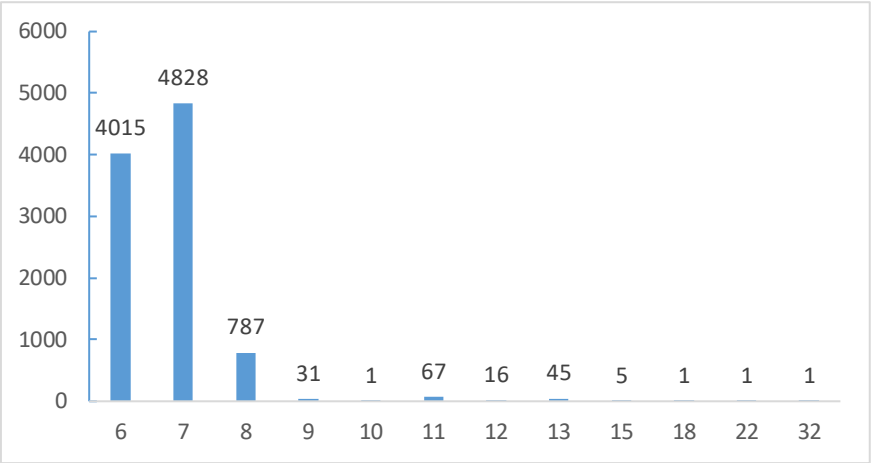


Fig. 3 Count occurrence of contact residues for six or more contact residues. a) Protein A motifs; b) Protein B motifs. Mutation rate 0-20%. Upper graph: all organisms drawn out from the PDB and UniProt; lower graph: *H. sapiens*. Motifs with seven contact residues occur the most, which corresponds to about one TM alpha-helix.

The protein-protein docking benchmark 5.0 (PPDB) [42] assembles non-redundant high-resolution complex structures and for which the X-ray or NMR unbound structures of the constituent proteins are also available (https://zlab.umassmed.edu/benchmark/). However, none of the complexes in PPDB v5.0 correspond to TM proteins. From our 53 non-redundant template structures of protein-protein complexes, we could extract a subset of them along with

the unbound structures of their components in order to define a true benchmark made up of 10 sets of structures (Table 1; this table is derived from Table S2, column C).

Table 1 Benchmark of membrane protein-protein complexes and their unbound components

Complex PDB code	Unbound PDB code protein A	Unbound PDB code B protein B	Type Homo- (Hm) or Hetero- (Ht) meric
2LOH	2LLM	2LLM	Hm
2QTS	3IJ4	3IJ4	Hm
	4FZ1	4FZ1	Hm
	4NTW	4NTW	Hm
	4NTX	4NTX	Hm
	4NTY	4NTY	Hm
	4NYK	4NYK	Hm
2ZW3	5ER7	5ER7	Hm
	5ERA	5ERA	Hm
3SYQ	3SYA	3SYA	Hm
	3SYC	3SYC	Hm
	3SYO	3SYO	Hm
	3SYP	3SYP	Hm
	4KFM	4KFM	Hm
4JKV	4N4W	4N4W	Hm

	4O9R	4O9R	Hm
	4QIM	4QIM	Hm
	4QIN	4QIN	Hm
4NEF	4OJ2	4OJ2	Hm
4WO1	2L34	2L34	Hm
	4WOL	4WOL	Hm
4X5T	2M6B	2M6B	Hm
5A63	2N7Q	2N7Q (for Q92542)	Ht
	2N7R	2N7R (for Q92542)	Ht
5O9H	6C1Q	6C1Q	Hm
	6C1R	6C1R	Hm

When comparing to the "Dimeric complexes" table of the Membrane Protein Complex Docking Benchmark, MemCplxDB [43], we only recover the PDB ID 5A63 complex, given that MemCplxDB shows many interactions between TM proteins and non TM, soluble proteins (antibodies, peripheral proteins, etc), which we do not deal with; MemCplxDB includes as well interactions between oligomers within a multimer complex, and prokaryote membrane proteins (beta-barrel). Our benchmark represents thus a gold-standard set of positives for integral-to-membrane proteins interacting through their alpha-helical TM segments.

Predicted interactions

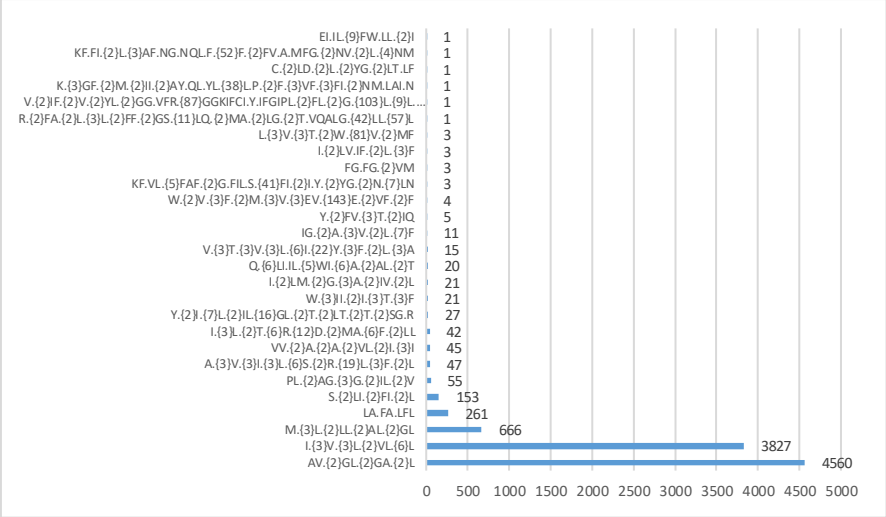
We identified many potential recognition sites at different mutation rates. The number of predicted motifs in TM regions varied depending on the mutation rate. Thanks to this

variable, we managed to predict many new sites for the corresponding motifs and many new binary complexes (Fig. 4). However, we noticed that more potential motifs were found when the original motifs had a lower number of contact residues (Fig. 5). The range of contact residues in the motifs is 6-32. Thus, we found small motifs with six to nine contact residues extremely abundant. As the number of contact residues increased, the number of hits decreased drastically, with only three predictions for both motifs at 32 contact residues, the longest motif. As a function of the mutation rate, TransINT predicts 346 interactions for the *H. sapiens* species for a mutation rate of 0% (motif A). For the mutations rates 5, 10, 15 and 20%, the number of interactions is of 16, 4440, 4274 and 724, respectively (Fig. 6).

As mentioned in the M&M section, through multiple sequence alignments, we built a consensus sequence motif for all the matched sequences of a given binding site in the 0-20 % mutation range, considering the contact residues only. This led to detection of conserved amino acid residues among the contact residues. The most prevalent consensus motif A found was AV.{2}GL.{2}GA.{2}L, illustrated by the specific sequence ^{423}AVFSGLICVAMYL^{436} of the sodium-dependent proline transporter (UniProt Q99884) at 15% mutation allowance (Fig. 4a). For consensus motif B, two motifs are highly populated - I.{2}LV.IF.{2}L.{3}F and LL.{3}VL.SI.{2}AF.{2}G (Fig. 4b). The least common motifs are characterized by important sequence lengths and long segments in between the contact residues. An example is motif B LL.AS.{90}LV.EA.FAI.NI.{2}S.{2}L.{3}F.{19}I.{100}I, found in the short transient receptor potential channel 4 (UniProt Q9UBN4).

Mutation rates of 0 % for protein A and protein B result in those proteins whose binding motifs conform exactly to the consensus binding motifs. This produces 246 entries of complexes across species and includes homo- and hetero-oligomers.

a)



b)

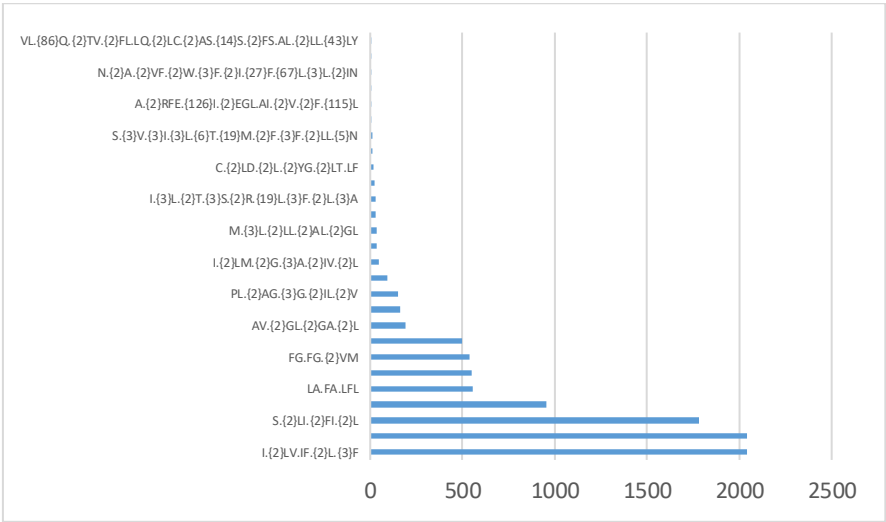
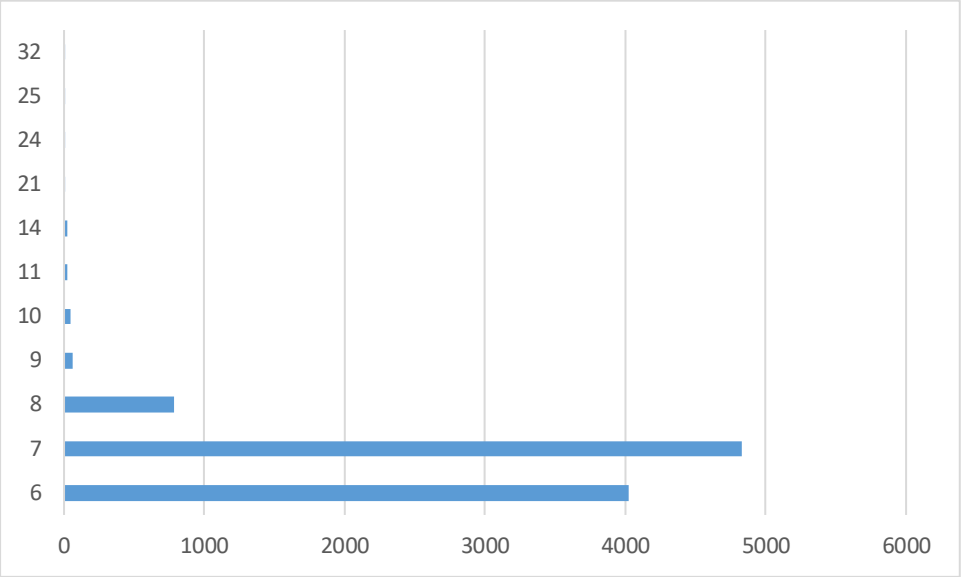
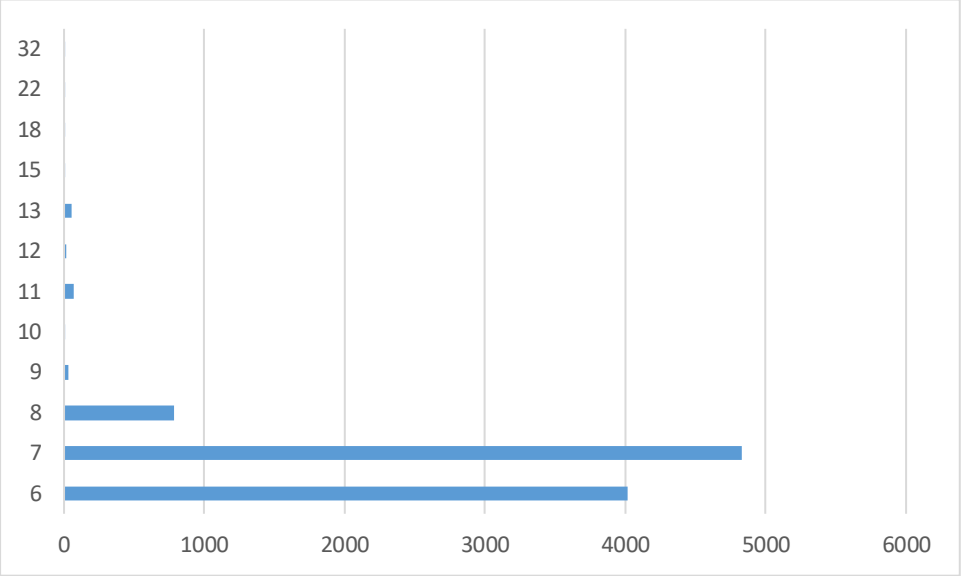


Fig. 4 Number of hits per motif. a) motif A, b) motif B. Mutation rate 0-20%

a)


32
25
24
21
14
11
10
9
8
7
6
0
1000
2000
3000
4000
5000
6000

b)


32
22
18
15
13
12
11
10
9
8
7
6
0
1000
2000
3000
4000
5000
6000

c)

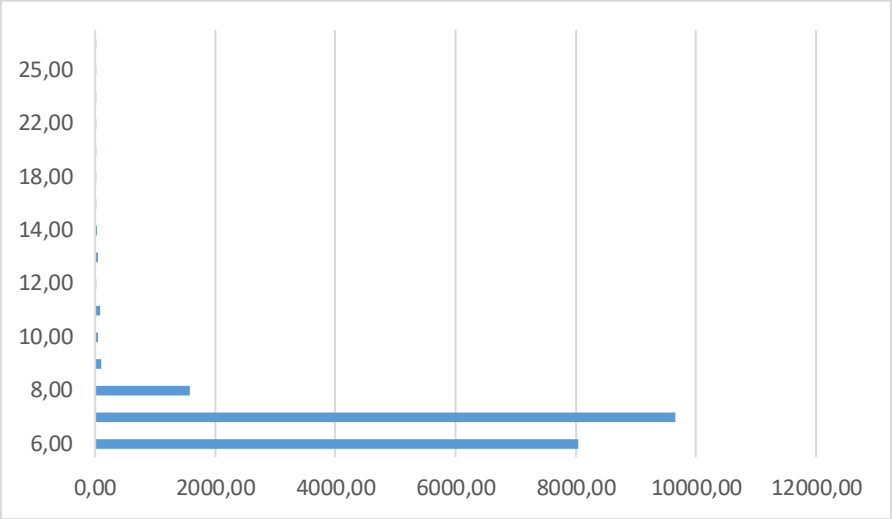


Fig. 5 Frequency of predicted proteins per number of contact residues for different mutation rates. *H. sapiens*, mutation rates 0-20%. a) Motif A; b) Motif B; c) All A and B motifs combined.

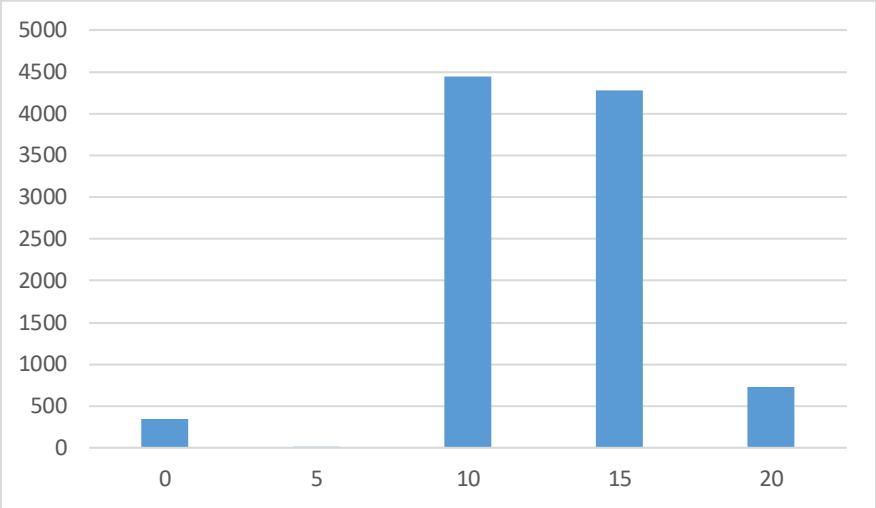


Fig. 6 Number of interactions as a function of mutation rate for *H sapiens*.

A multiple alignment of several different protein sequences results in a consensus sequence. Thus, for example, with the VV.{2}A.{2}A.{2}VL.{2}I.{3}I motif of length 19 and up to 20% mutation rate is shown in Fig. 7. Out of the 19 positions, several show significant variations. The consensus sequence leads to a fingerprint of that binding motif, in this case $VVX_2AX_2AX_2VL$.

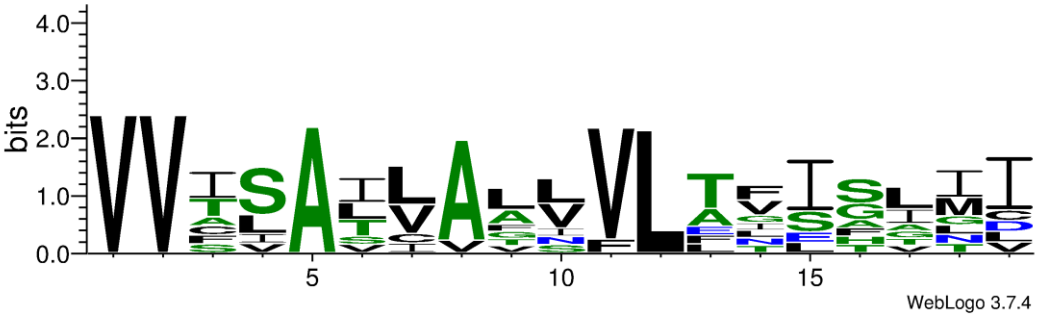
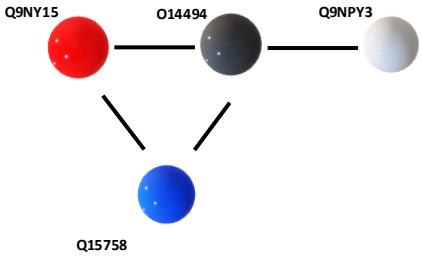


Fig. 7 Multiple sequence alignment WebLogo (https://weblogo.berkeley.edu/) showing consensus for a motif of length 19. Black denotes hydrophobic residues; green polar residues.

Interaction networks

Networks link overlapping pairs of proteins, from which one can propose multimer complexes if the binding sites are different and non-overlapping, and if the interactions are simultaneous. From the predictions, unexpected connections can be found, and new possible cellular roles can be proposed for some of the complexes, including linking the network to a disease pathway. The architecture of a network reflects biological processes and molecular function.

We illustrate below a subnetwork of predicted membrane MPPIs for *H. sapiens*:



Where:

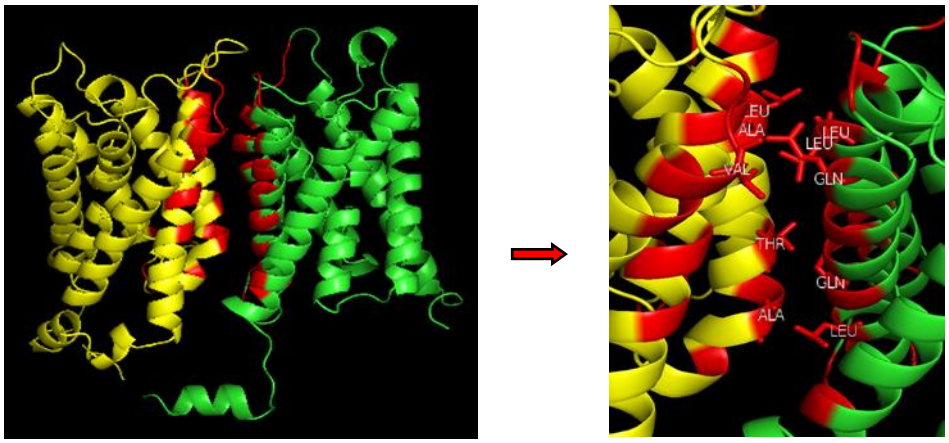
Q9NY15 (*STAB1*) Stabilin 1, O14494 (*PLPP1*) Phospholipid phosphatase 1, Q15758 (*SLC1A5*) Solute carrier family 1 member 5 or Neutral amino acid transporter B(0), and Q9NPY3 (*CD93*) Complement component C1q receptor.

In support of the proposed subnetwork, we found that all four MPs are present in two tissues of *H. sapiens'* -adipose tissue (major) and breast (minor) (Protein Atlas). In addition, a *CD93* - *STAB1* interaction in human is reported in the String DB of PPI networks, and *CD93* and *PLPP1* are co-expressed in *G. gallus*. [44]

Molecular docking simulations

We then selected several MP pairs for which TransINT predicted an interaction based on the interface epitopes. To begin with, we took the human AQP5 (P55064) and AQP2 (P41181) protomers and proceeded to a molecular docking as described in the M&M section. In this case, the crystal structures existed for both MPs -PDB 3D9S and 4NEF, respectively. Fig. 8a shows the resulting heterodimer with the interface regions. We then selected the predicted rat AQP2-ErbB-3 pair, interacting through the AV.{2}GL.{2}GA.{2}L pattern present on both protein surfaces. Since the experimental structures are not available for either MP, we homology-modelled them using the 3D structures of human AQP4 (PDB code 2D57) and human ErbB-2 (PDB code 2JWA), respectively. The sequences of both aquaporins are more than 30% identical, just like those of ErbB in the TM region; the resulting models are thus highly reliable. Fig. 8b shows that the modeled rat AQP2-ErbB-3 heterodimer has the contact residues also at the interface and is thus viable. The interface residues are generally located on the membrane-exposed surface of the alpha-helical to which they belong.

a)



b)

Fig. 8: Low-resolution cartoon structural model of predicted MP-MP complexes obtained by steered molecular docking. a) *H. sapiens'* protomers of aquaporin 5 (Uniprot P55064, 3D9S, yellow) and aquaporin 2 (P41181, 4NEF, green) in complex, with interface residues in red. The figure to the right shows a zoom of the interface with the contact residues; b) 3D model of the *Rattus norvegicus'* protomers of aquaporin 2 (P34080) and the receptor tyrosine-protein kinase ErbB-3 (Q62799) in complex, with interface residues in purple and in red, respectively. In both complexes a) and b). The contact residues are indeed at the interface of the complex and allow its formation.

DISCUSSION

In this work, we developed an interface residue-based protocol destined to predict in large scale binary interactions among the alpha-helical transmembrane segments of TM proteins. We illustrate the protocol in the case of the eukaryote plasma cell membrane. The highly matched motifs in the different alpha-helical membrane proteins we found belong all to TM alpha-helices as no 3D structure of a complex of an integral-to-membrane beta-barrel protein has yet been obtained for eukaryotes. Our motifs are most likely part of conserved domains of TM regions, especially in the case of receptors of the same family, like the GPCRs. Supplementary studies on these specific domains, mainly the ones with six to eight contact residues can help confirm this hypothesis as well as help further investigation in other organisms. As we found, amino acids tend to interact the most with other amino acids sharing the same properties, such as hydrophobicity. [45] Membrane proteins being mostly hydrophobic, it seems thus logical that Leu, Val, Ile and Phe have the highest frequencies of occurrence in the found patterns. [46] This is the opposite to globular proteins, whose non-uniform interaction interface is characterized by two main regions -core (buried in the interface, rich in hydrophobic residues) and rim (containing solvent-accessible hydrophilic residues). Further analysis of these interactions would be to study if the proteins still bind when substituted with amino acids of different chemical properties and their complementary counterparts. Being contact residues, the amino acids composing the binding motifs we report do not represent necessarily so-called hot-spots, i.e. those that contribute to the binding free energy, [47] as we did not perform any energy calculations nor estimate the binding affinities. But, because of their degeneracy in amino acid type (Leu, Phe, Ile, i.e. hydrophobic; Fig. 2), it may be that all TM interface residues contribute equally to the free energy of binding as interface core and rim regions are absent in TM proteins.

Some of the proteins, like the human glycine receptor subunit alpha-3 (O75311), show more than one binding motif, suggesting a promiscuous binding behavior. [48] Thus, in this ligand-gated chloride channel, we detect several distinctly different interface motifs - A.{3}V.{3}I.{3}L.{6}S.{2}R.{19}L.{3}F.{2}L, Y.{2}I.{7}L.{2}IL.{16}GL.{2}T.{2}LT.{2}T.{2}SG.R, Q.{6}LI.IL.{5}WI.{6}A.{2}AL.{2}T, I.{3}L.{2}T.{6}R.{12}D.{2}MA.{6}F.{2}LL, and Q.{6}LI.IL.{5}WI.{6}A.{2}AL.{2}T – found by TransINT as direct interactions, furnishing an explanation of the multiplicity of binding modes of proteins and their multifunctionality. The interface recognition motifs evidenced in this work, while being mostly constituted of hydrophobic residues, show a diversity in their linear sequence, a spatial arrangement corresponding to an alpha-helical pattern, and account for a limited set of strictly different motifs.

The PPI template-based prediction algorithm and server PRISM2.0 [49] (http://cosbi.ku.edu.tr/prism) uses also 3D structural similarity of interfaces with respect to templates to propose other protein complexes. In TransINT, we look for sequence similarity at the TM interface level. PRISM is not specific for TM proteins and requires the 3D structure of the targets in order to propose an interaction. Thus, when having an interface template corresponding to a TM protein, it may propose not only TM protein complexes, but also globular protein complexes. Many of our TM template interfaces are absent in the PRISM dataset. Our TransINT dataset is formed of plasma eukaryote membrane TM proteins, thus belonging to the same cell compartment and to the same species. When comparing our dataset of interactions with that of the machine learning approach applied to the human membrane receptor interactome HMRI [32], which seems to list only heteromers and for which not all the interactions are between TM proteins, we find a correspondence for the heteromer pair of genes TYROBP- KLRC2 (p value = 0,034341), for example. The prediction of putative interactions

by the BIPS approach [50] is based on sequence homology between proteins found in PPI DBs and templates, all this based on sequence alignments. We find several correlations between BIPS and TransINT. For instance, we propose an interaction between T-cell surface glycoprotein CD3 zeta chain (P20963, gene name *CD247*) and high immunity immunoglobulin epsilon receptor subunit gamma (P30273, gene name *FCER1G*). BIPS predicts a similar pair P20963-P08637 (low affinity immunoglobulin gamma Fc region receptor III-A, gene name *FCGR3A*). When relating our TransINT predictions at a 0-20% mutation rate for human proteins to the FPClass dataset of genome-wide predicted human PPIs [19], a data mining-based method, we find several consistencies, listed in Table 2.

Table 2 Predicted interactions common to TransINT (20% mutation rate, *H. sapiens*) and FPClass.

UniProt AC A	Gene name B	UniProt AC B	FpClass total score
P18507	*GABRB2*	P47870	0.4029
P34903	*GABRA6*	Q16445	0.2915
P47869	*GABRA6*	Q16445	0.2738
P47870	*GABRA4*	P48169	0.2901
P47870	*GABRA6*	Q16445	0.2837
P48169	*GABRB2*	P47870	0.2901
P48169	*GABRA6*	Q16445	0.2786
Q01814	*ATP2B3*	Q16720	0.3235
Q16445	*GABRB2*	P47870	0.2837
Q16445	*GABRA4*	P48169	0.2786
Q16720	*ATP2B2*	Q01814	0.3235

Although most datasets based on experimental approaches (non-exhaustive list: Clinically relevant GPCR PPI mapping [51]; IMEx consortium (http://www.imexconsortium.org/); IntAct (www.ebi.ac.uk/intact/); Human Protein Reference Database (hprd.org); HINT (http://hint.yulab.org/); CORUM [52,53] (http://mips.helmholtz-muenchen.de/corum/); API (http://cicblade.dep.usal.es:8080/APID/init.action); HI-II-14 [21] (http://interactome.dfci.harvard.edu/H_sapiens/); HI-III-20 [9] (http://interactome.baderlab.org/); BioPlex [54], (http://bioplex.hms.harvard.edu/); QUBIC [55]; IID [56]) cover the entire human proteome, membrane proteins are under-represented (for an overview of the major high-throughput experimental methods used to detect interactions, see [57,58]). Thus, the experimental work of Rolland and colleagues on the human interactome network [21] found 13,944 high-quality binary protein-protein interactions, mapping 4303 unique Gene IDs, out of which 4188 were found to correspond to a total of 4198 proteins found in UniProt. Among the 4198 proteins, the authors found 179 TM proteins interacting as protein A and 141 TM ones interacting as protein B, for a total of 253 unique TM proteins. Moreover, amongst the interactions found, only 41 were between TM proteins. Twenty-eight of these proteins were found interacting in the IntAct DB. Nevertheless, none of the interactions we extracted from the structural PDBsum DB and that served us as template structures for TM complexes were found among the 41 interactions. Perhaps some of these interactions are between the juxtamembrane regions of the TM proteins reported by Rolland et al. for TM proteins. We did not find either any of our predictions in their results. In the HI-III-20/HuRI updated dataset [9], 52,569 high-quality binary protein-protein interactions involving 8,275 proteins are reported, none of which are MPs ($\log_2$ odds ratio < 0, Extended Data Fig. 6a of Luck et al.). As for the BioPlex 3.0 DB, using the TAP-MS technology for detecting PPIs, the authors found more than 2,000 TM interactions out of ~73,000 stored in their DB. TransINT predicts interactions between human Gamma-aminobutyric acid receptor subunit beta-3

(P28472) and Gamma-aminobutyric acid receptor subunit beta-2 (P47870), and between the former and Gamma-aminobutyric acid receptor subunit alpha-6 (Q16445). The two interactions are listed in the MENTHA experimentally determined direct protein interactions DB [16], and the latter is confirmed in the BioPlex (p(interaction)>0.99) and ComplexPortal DBs (Complex AC: CPX-2164 and CPX-2951). In addition, when assessing our data for *H. sapiens* with the IID DB, we find, for example, that the predicted P35968 – P35916 (Vascular endothelial growth factor receptor 2 - Vascular endothelial growth factor receptor 3) is validated experimentally.

Note that there are large discrepancies and dramatic differences in the content between experimental PPI data collected by the same or different techniques, making our attempt to compare our predictions to experimental data more haphazard. Indeed, the intersections between various interaction maps are very small. [24,59] In addition, the presence of orthologs makes the research more cumbersome, given that sometimes a gene name may belong to a species different from the searched one. In a few words, there are a number of caveats of any analysis comparing PPIs. [60]

The importance of recording negative results of PPI assays in interatomic DBs, i.e. those indicating that the tested proteins do not interact, has been raised. [61] But identification of them is less straightforward. This stage should lead us to define a set of true negative interactions for training a predictor of MPPIs, since sampling of negatives is crucial for optimal performance [62,63]. Looking for negative interactions in the IntAct DB for our unique UniProt reviewed TransINT proteins in the 0-20% mutation rate range, we find two negative interactions of two TransINT proteins with two other proteins, i.e. TransINT Q9UNQ0 (ATP-binding cassette transporter ABCG2) with P11309-1 (a serine/threonine-protein kinase); and TransINT P04626 (receptor tyrosine-protein kinase erbB-2) with O75367-2 (core histone macro-H2A.1). But P11309-1 and O75367-2 are not MPs! Analogously, the Negatome [64,65]

and Stelzl [66] datasets compile sets of protein pairs that are unlikely to engage in physical direct interactions. We observed that spanning the Negatome dataset with our predicted positive interactions for *H. sapiens* results in no negative interactions, but several MPs in the Negatome set are absent from TransINT.

Our approach has some limitations as, on one hand, the inaccuracies in the original data will likely propagate; on the other hand, the predictions do not consider explicitly several aspects. Firstly, we are assuming the membrane proteins to be rigid bodies in the sense that the implicit 3D relationship between the amino acid residues that is coded in the regular expression giving the interaction sites is not largely modified in going from the reference protein to the working protein. i.e. the complex is formed with no major conformational changes. Indeed, conformational changes upon quaternary structure formation based on experiments and computer simulations indicate that the changes are not major for the TM domains. On the other hand, the non-interacting surfaces may also play important roles in binding affinity through allosteric mechanisms. Also, there is no guarantee that the *in vitro* multimers reported in the PDB exist *in vivo* and are not a product of the crystal lattice or other experimental conditions, being thus transient complexes; [67] yet, the complexes could exist and be obligate since there are no enzyme-peptide complexes, and the more there is contact between the subunits, the more the complex will be structurally and thermodynamically stable. Moreover, homotypic interactions are obligate assemblies in general. In addition, the lipid composition of the surrounding membrane, absent in the experimentally determined structures, may modulate the oligomerization interface and thus, the activities or functions of the complex. There may be as well ligand-induced binding effects not considered in this work. In addition, situations in which the TM interaction might be driven by extra-membrane interactions are not considered in this work, as we are looking only at in-membrane interactions between TM fragments. Moreover, several experimentally determined structures that enter in our initial dataset originate from TM

fragments of parent proteins and not of the entire protein. On another hand, the possible control of membrane PPIs by post-translational modifications, such as palmitoylation is not considered, given that these modifications take place in the juxtamembrane regions of TM proteins, namely in the lumen. [68] Splice variants of many genes code for multiple protein isoforms. We examine only the isoforms of the PDB template structures and those recovered from UniProt canonical sequences. Also, the oligomerization state of the experimental structure of TM proteins or their fragments thereof is most of the times assessed in a detergent-solubilized state, with nonpolar solvents and/or denaturing conditions; protein association modes in the cell might thus be different. Finally, our *de novo* predictions are limited in number due to the incompleteness of the membrane protein-protein complex 3D structure library, i.e. the PDB DB.

Within the limits of our assumptions and of our approach, the proposed amino acid residue recognition sites and corresponding molecular complexes in this work can address the possible effects of natural or artificial mutations on PPIs, as well as protein function and disease associations. Thus, the potential interacting membrane proteins identified in our DB may be used as candidates to be validated experimentally.

Finally, because the MPPI data set is scarce and noisy, we intend to apply machine-learning techniques to build a better predictive model to estimate complexes between one MP and its interacting partners. For this, we will train heterogeneous features of the interface amino acid residues in the binding motifs (type, frequency and conservation, physicochemical characteristics, patterns of interactions between interface residues) from different data sources using the extracted motifs. On another hand, we will train protein-related attributes (molecular function, biological process, metabolic pathway, protein length, gene and protein expression, tissue and cell type), leading to an increased accuracy, specificity and coverage of our MPPI predictions.

CONCLUSION

The TransINT DB is a model-driven biological discovery tool. It contains TM protein recognition sites representing interactions as binary states without specification of the affinity between interacting units. Based on the assumption that close homolog structural interacting motifs interact always in similar orientations and with equivalent interfaces, TransINT predicts a TM protein interactome with thousands of *de novo* interactions, including multiple recognition sites for the same couple of partners, i.e. identifying multiple-interface proteins and binding sites for which different proteins may compete. The obtained sequence motifs identify homodimeric interaction interfaces.

TransINT is a DB to be queried for the discovery of verified and potential interactions, and for obtaining different types of necessary information about them. Only interactions in the plasma membrane interactome in which the binding sites involve specifically TM regions are reported in this paper. Our membrane interactome represents "first draft" predictions and contains 42,746 entries for all species dealt with (19,101 for *H. sapiens*). The large number of protein partners we predict suggest that even distantly related proteins often use regions of their surface with similar arrangements to bind to other proteins. The predicted interaction partners allowed us to generate low-resolution 3D structures for a selected number of predicted complexes, showing that complex formation is feasible as the interacting surfaces of the individual proteins manage to face each other in the docked complex.

Comparing our predictions with the predictions of DBs and approaches based on water-soluble globular PPIs and benchmarked on (different versions) of the PPDB from those approaches is not recommended as it will not give an accurate picture of the TransINT approach. Complementary to the sequence-based co-evolution PPI prediction methods, [30,69–71] our 3D-to-1D approach encodes 3D into 1D and adds the spatial dimension to a given

membrane interactome. It may lead to new biological hypotheses concerning PPIs at the membrane level, to genotype – phenotype relationships, to investigate the effect of pathological mutations on the interaction between MPs and to propose molecular mechanisms of action. Allowing for sequence variation in the interface residues of the motifs has allowed us to extend the initial set of template-derived TM PPIs to other families and family members and to detect remote homologs of the starting template complexes that use regions of their surface with similar arrangements of tertiary structure elements for binding to other proteins. In addition, TM proteins showing more than one interface may lead to multimer formation. In fact, binary protein complexes could be the first step to generate higher order macromolecular edifices.

Because of their number and diversity, the higher-order structures and interfaces generated in this work, once validated, represent potential pharmacological targets for the discovery of modulators of the protein-protein interface, such as membrane–insertable, metabolically stable, non-toxic small-molecule active MPPI inhibitors or stabilizers, for example exogenous peptides and peptidomimetics, influencing in vivo the assembly of intact membrane proteins involved in various diseases. [72–75] Indeed, many of the TM proteins in the predicted complexes are involved in a variety of diseases (OMIM, An Online Catalog of Human Genes and Genetic Disorders, https://www.omim.org/) as deficient or enhanced oligomerization is associated with diseases. [76,77] In addition, our protein interaction data implies direct interactions and can lead to the construction of protein interaction networks. Thus, our results may help understand biochemical pathways, the crosstalk between them and transactivation, may suggest potential drug targets, such as those targeting the MPPI interface, [78,79] and may aid in understanding the functional effects of mutations at the interface. [80]

Finally, the TransINT approach can be extended to other cell membranes (mitochondria, nucleus, endoplasmic reticulum, Golgi apparatus) and across the tree of life, as the 3D structures of proteins integral to these membranes become available, and by using

homologous protein networks in different organisms. In addition, its predictions can be refined, i.e. the number of false positives and the noise reduced by insuring the MPs belong to the same developmental stage, tissue, cell type, site of expression, reaction and metabolic pathways, and whose protomers in a complex do not show a gene distance of more than 20 and who are functionally similar [81].

MATERIALS AND METHODS

Sources of information retrieval and data filtering

We collected information from several publicly available DBs. Fig. 9 summarizes the steps followed to collect, filter, and process the input data, and to generate the resulting data. We started by obtaining a list of all eukaryote "reviewed" proteins from UniProt (release 2017_06[82]) with cellular component annotations that matched the following GO [83] annotations: "integral component of membrane" (GO:0016021) <u>and</u> "plasma membrane" (GO:0005886), <u>or</u> "integral component of plasma membrane" (GO:0005887). From the list, we then identified the subset of proteins that have an experimental 3D structure in the PDB [33] in the form of a complex spanning the TM region with at least six interacting residues in each monomer, enough to present an accessible surface area leading to quaternary structure. For the found PDB structure to be considered as valid, it had to have a resolution of 3.5 Å or less and not be a theoretically modeled structure. We took in consideration all different conformational states of receptors (active, inactive), channels (resting, open, closed, desensitized, etc.), and transporters, regardless of pH, symmetry group, apo or holo form, and allosteric effects, unless the differences modified the set of interface residues. PDB membrane proteins presenting engineered mutations, insertions and deletions in the TM segment with respect to the wild-type or natural variant sequence in the UniProt DB were eliminated (for ex. 3UKM, 3RHW, 4U2P and 4PIR), just like chimeras of which the xenophobic part is not TM (for ex. 2L35). Also, we

did not discriminate whether the structure of the TM protein was in the presence or absence of ligand(s), such as agonists and antagonists. We also ignored redundant membrane proteins and those whose 3D structure show no significant TM segments; excluded also are those pair interactions that are redundant due to the symmetry of a given complex (for ex. 4X5T). As the structures chosen were based on OPM [34] and included also manual curation, we excluded non-parallel or perpendicular configurations between the subunits of a homodimer, as biological knowledge of their relative orientation of TM proteins in the membrane indicates. Of course, those proteins that do not form oligomers, show head-to-head or head-to-tail orientations of the protomers, have only out-of-membrane interactions, or have TM segments that are so far from each other in the oligomer are not instanced; in an analogous fashion are excluded those complexes whose TM segments do not interact, like in the human sigma-1 receptor (PDB 5HK2). As a corollary, our approach includes implicitly the intra-molecular interactions that take place among the protomers composing an oligomer.

Finally, to ensure that the oligomer structures we took into account are quaternary structures biologically speaking, we used EPPIC, a protein-protein interface classifier [67], and PRODIGY, a classifier of biological interfaces in protein complexes [84,85] to distinguish between crystallographic and biological assemblies.

Motif extraction

We then had to choose the PDB structures to work on. For this, we referred to the OPM DB [34] that provides the orientation of known spatial arrangements of unique structures of representative MP coming from the PDB with respect to the hydrocarbon core of the lipid bilayer. We chose all the PDB structures that map to the TM proteins we extracted from UniProt and we extracted all the available PDBsum files of these structures. PDBsum [36] is a DB that, among other things, shows schematic diagrams of the non-bonded contacts between

amino acid residues at the interface of molecules in a multimer complex (non-bonded contacts are defined as any contacts between ligand and protein involving either a carbon or a sulfur atom, where the interaction distance is ≤3.9Å; R. A. Laskowski, personal communication). We double-checked the chosen PDB structures in the MPStruc DB of membrane proteins of known 3D structure (White SH. Membrane proteins of known 3D structures. https://blanco.biomol.uci.edu/mpstruc/). We then used the information in PDBsum to extract and identify the motifs at the binding sites by obtaining the linear sequence of the contact residues and the residues in between. From the PDBsum file listing the contacts between the couple of interacting we formulated thus two motifs, one corresponding to partner protein A, the other one corresponding to partner protein B. Since we are only interested in the recognition site in the TM region, we ensured that each interacting residue belonged to the TM part of the sequence. In addition, we kept binding sequences made up of at least six TM residues. We represented our motifs using the regular expression format. We denoted the TM contact residues by their one letter symbol, the residues in between (representing the wildcard regions) by a dot (the wildcard that matches any character), and the curly braces for the consecutive occurrences of the wildcard, such as in the pattern I.{3}V.{3}L.{2}VL, which matches any string that contains an I, then three other characters, then a V, then three other characters, followed by an L, then two other characters followed by VL.

Searching for identified motifs in other protein sequences

For eliminating redundancy in our data, we grouped similar motifs together and built a consensus for each cluster using multiple sequence alignment. The similarity threshold we used was 20% of the number of contact residues in the sequence. We generated new potential binding sites by applying mutation rates of the contact residues ranging from 0% (exact match) to 20%, with increments of 5%. Subsequently, we queried our consensus motifs against the

original UniProt dataset. We defined a Cost parameter as the percent of mutations allowed per motif, depending on the number of contact residues it contains and the mutation rate for the run. Cost was given as the score for substitution, while both insertions and deletions were given a score of 0 to ensure no contact residue is lost. For instance, when generating new sites from a valid motif with eight contact residues, Cost is two for a mutation rate of 20%, such as in the sequence PRRAAVAIAGCWILSLV, derived from the PL.{2}AG.{3}G.{2}IL.{2}V motif (originating from PDBsum 2L2T of UniProt ID Q15303, human receptor tyrosine-protein kinase erbB-4) containing eight contact residues, in which the two underlined contact residues represent the mutations. The values of Cost_A vary from 0-5, and those of Cost_B from 0-8.

To keep track of which A motif is interacting with which motif B, we made sure we kept the motifs in separate pools. After collecting all predicted motifs with their corresponding matched sequence for each mutation rate, we calculated a second score, this time based on an amino acid substitution matrix for membrane proteins, [21] updated by Betts and Rusell [86] and corrected by our means. The higher the score, the more accurate our prediction was. The values oscillate between 3 and 140. Afterwards, we associated the predicted motifs from new interactions based on the PDBsum-validated interactions. This way, we were sure that pattern A from protein A bound to its corresponding pattern B of protein B. Since motifs can be found anywhere in the sequence, we checked which ones were in the TM region and considered only these. We also checked which TMs had the motifs included in their PDB structures when available.

Molecular docking

To illustrate and partially validate our approach, we proceeded to generate model molecular complexes. For that purpose, we selected several predicted pairs of MPs and proceeded to search for a protein-protein docking program that would allow us to perform a

steered docking simulation using the epitopes we extracted, i.e., those contact residues obtained from the PDBsum DB at the molecular interface of the complex. Thus, we processed and analyzed large amounts of experimental 3D TM structure files using three docking programs - HADDOCK, [87–89] ClusPro, [90–92] and GRAMM-X. [93,94]. Even though all three docking programs are sufficiently precise, in this work we decided to use GRAMM-X for creating the new protein-protein 3D complexes since it has an "Interface residue constraints" option in which the user is allowed to submit parameters like "Potential receptor interface residues" and "Potential ligand interface residues," i.e. those residues that might form the interface between the "receptor" and the "ligand". The program has also options like "Receptor residues required" and "Ligand residues required" in which the residues required to be in contact are listed. The "Receptor-ligand residue pairs required" option takes receptor-ligand residue pairs from the lists of the potential interface residues that are required to be in contact with each other and allows thus to perform steered molecular docking using the binding residues. Finally, we proceeded to a manual curation by filtering out nonparallel, perpendicular or oblique protomers, regardless of the calculated energy. In our research, these options fit perfectly our task. We also considered the topology of the MP in the membrane, i.e. its orientation with respect to the membrane plane. To verify the performance of GRAMM-X for MPs, we benchmarked it against several MP complexes in the PDB (not shown). GRAMM-X was indeed able to reproduce many of the experimental MP complexes. For the molecular docking and the identification of the TransINT predicted protein complexes, we chose examples in which the 3D PDB structures of proteins were already available or represented very highly homologous templates. We studied proteins from several organisms, including *H. sapiens*. The obtained 3D structures of the MP-MP complexes were visualized via PyMol program (www.pymol.org). Since we have the docking interfaces, we did not have to do an *ab initio* molecular calculation in the membrane, such as done by other programs like DOCK/PIERR,

[95,96] nor to be concerned if the docking program was trained on sets composed primarily of soluble proteins.

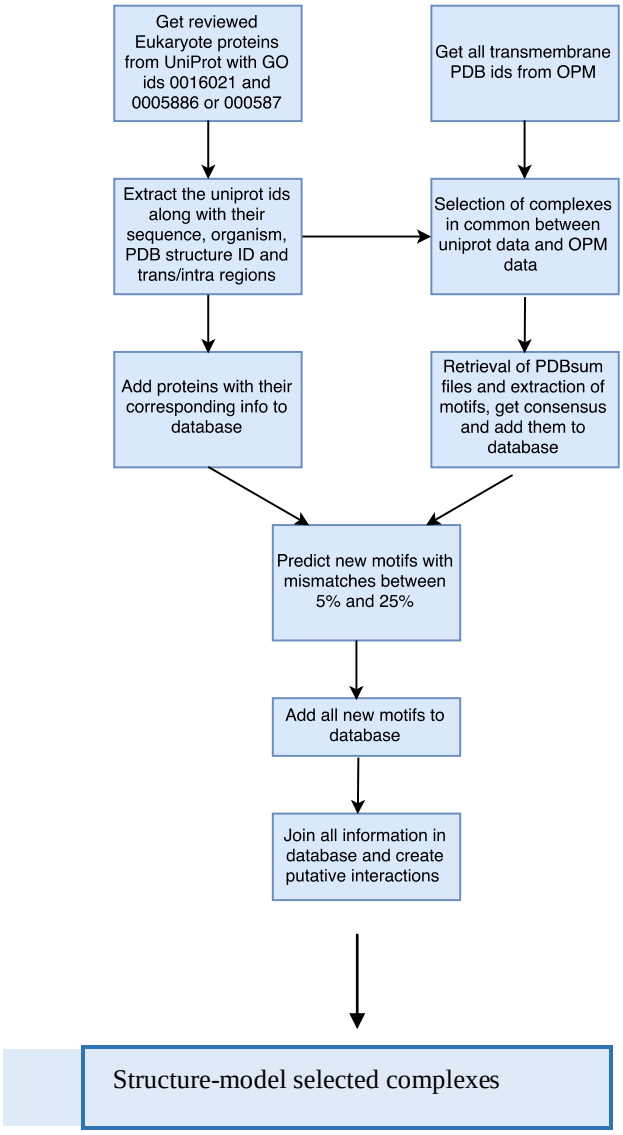


Fig. 9: Workflow illustrating the TransINT algorithm, from information retrieval to detection of recognition motifs to generation of putative interactions to 3D modeling of complexes.

The TransINT database

Thereafter, we built a heterogeneous DB named TransINT that contains all the found interactions. The DB is the result of the implementation of a fully automated template-based

recognition site search pipeline. We used a MySQL DB (version 5.0.11) to keep all the information collected in an orderly manner. To access the DB and search for our data, a web interface was built using PHP (version 5.6.14) and HTML5, which allows the user to query the needed information. Users cannot update the DB -they can query it for obtaining motifs by entering a UniProt ID, a type of organism, a mutation percent (to get all proteins with a mutation score below or equal to this rate), or to get the motifs with at least a certain number of contact residues. Another way to query the DB is by entering a motif of interest or a part thereof, using the regular expression form, retrieving thus the predicted interactions we managed to generate. The user can choose more than one filter option when querying and will only obtain interactions thought to occur in TM regions and between plasma membrane proteins of the same species. Our DB is updated following each release of UniProt and OPM datasets. All statistics are then regenerated.

Acknowledgments

Financial support supplied by:

- Hubert Curien CEDRE program, grant 13 Santé/L2.

- Fonds pour le Rayonnement de la Recherche financing program 2016, 2017, 2018. Université d'Evry-Val-d'Essonne, Action 3 – Incoming and outgoing mobilities.

- Fonds pour le Rayonnement de la Recherche financing program 2016, Université d'Evry-Val-d'Essonne, Action 1 – Financing for supporting the emergence of innovating projects in the framework of the evolution of scientific policy.

- French Embassy in Armenia. Bourse du Gouvernement Français.

REFERENCES

1. Kastritis, P. L. & Bonvin, A. M. J. J. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of the Royal Society, Interface* **10**, 20120835 (2013).
2. Stevens, T. J. & Arkin, I. T. Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins: Structure, Function and Genetics* (2000) doi:10.1002/(SICI)1097-0134(20000601)39:4<417::AID-PROT140>3.0.CO;2-Y.
3. Yamamoto, K., Caporini, M. A., Im, S.-C., Waskell, L. & Ramamoorthy, A. Transmembrane Interactions of Full-length Mammalian Bitopic Cytochrome-P450-Cytochrome-b5 Complex in Lipid Bilayers Revealed by Sensitivity-Enhanced Dynamic Nuclear Polarization Solid-state NMR Spectroscopy. *Scientific Reports* **7**, 4116 (2017).
4. Guidolin, D., Marcoli, M., Tortorella, C., Maura, G. & Agnati, L. F. G protein-coupled receptor-receptor interactions give integrative dynamics to intercellular communication. *Reviews in the neurosciences* **0**, (2018).
5. Pin, J.-P. *et al.* International Union of Basic and Clinical Pharmacology. LXVII. Recommendations for the recognition and nomenclature of G protein-coupled receptor heteromultimers. *Pharmacological reviews* **59**, 5–13 (2007).
6. Bork, P. *et al.* Protein interaction networks from yeast to human. *Current opinion in structural biology* **14**, 292–9 (2004).
7. Stumpf, M. P. H. *et al.* Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 6959–64 (2008).
8. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nature methods* **6**, 83–90 (2009).
9. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).

10. Iyer, K. *et al.* Utilizing the Split-Ubiquitin Membrane Yeast Two-Hybrid System to Identify Protein-Protein Interactions of Integral Membrane Proteins. *Science Signaling* (2005) doi:10.1126/stke.2752005pl3.
11. Licata, L. *et al.* MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Research* (2012) doi:10.1093/nar/gkr930.
12. Chatr-aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Research* **45**, D369–D379 (2017).
13. Alfarano, C. *et al.* The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* (2005).
14. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research* **45**, D362–D368 (2017).
15. Orchard, S. *et al.* The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* (2014) doi:10.1093/nar/gkt1115.
16. Calderone, A., Castagnoli, L. & Cesareni, G. mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods* **10**, 690–691 (2013).
17. Orchard, S. *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature methods* **9**, 345–50 (2012).
18. Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology* **18**, 581–586 (2008).
19. Miller, J. P. *et al.* Large-scale identification of yeast integral membrane protein interactions. *Proceedings of the National Academy of Sciences* **102**, 12123–12128 (2005).
20. Mosca, R., Pons, T., Céol, A., Valencia, A. & Aloy, P. Towards a detailed atlas of protein–protein interactions. *Current Opinion in Structural Biology* **23**, 929–940 (2013).

21. Hwang, H., Petrey, D. & Honig, B. A hybrid method for protein-protein interface prediction. *Protein Science* **25**, 159–165 (2016).

22. Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).

23. Petschnigg, J. *et al.* The mammalian-membrane two-hybrid assay (MaMTH) for probing membrane-protein interactions in human cells. *Nature methods* **11**, 585–92 (2014).

24. Pitre, S. *et al.* Computational Methods For Predicting Protein–Protein Interactions. in *Advances in biochemical engineering/biotechnology* vol. 110 247–267 (2008).

25. Martin, S., Roe, D. & Faulon, J.-L. Predicting protein-protein interactions using signature products. *Bioinformatics* **21**, 218–226 (2005).

26. Lu, L., Lu, H. & Skolnick, J. MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins: Structure, Function, and Genetics* **49**, 350–364 (2002).

27. You, Z.-H., Lei, Y.-K., Zhu, L., Xia, J. & Wang, B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinformatics* **14 Suppl 8**, S10 (2013).

28. You, Z.-H. *et al.* Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinformatics* **15 Suppl 15**, S9 (2014).

29. Roy, S., Martinez, D., Platero, H., Lane, T. & Werner-Washburne, M. Exploiting Amino Acid Composition for Predicting Protein-Protein Interactions. *PLOS ONE* **4**, e7813 (2009).

30. Lage, K. Protein–protein interactions and genetic diseases: The interactome. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1842**, 1971–1980 (2014).

31. Keskin, O., Tuncbag, N. & Gursoy, A. Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. *Chemical Reviews* vol. 116 (2016).

32. Qi, Y. *et al.* Systematic prediction of human membrane receptor interactions. *Proteomics* **9**, 5243–5255 (2009).

33. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, 235–42 (2000).

34. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: Orientations of Proteins in Membranes database. *Bioinformatics* **22**, 623–625 (2006).

35. de Beer, T. A. P., Berka, K., Thornton, J. M. & Laskowski, R. A. PDBsum additions. *Nucleic acids research* **42**, D292-6 (2014).

36. Laskowski, R. A. *et al.* PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends in biochemical sciences* **22**, 488–90 (1997).

37. Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. The relationship between sequence and interaction divergence in proteins. *Journal of molecular biology* **332**, 989–98 (2003).

38. Keskin, O. & Nussinov, R. Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Engineering Design and Selection* **18**, 11–24 (2005).

39. Zhang, Q. C. *et al.* Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* **490**, 556–560 (2012).

40. Szilagyi, A. & Zhang, Y. Template-based structure modeling of protein-protein interactions. *Curr. Opin. Struct. Biol.* **24**, 10–23 (2014).

41. Mayol, E., Campillo, M., Cordomí, A. & Olivella, M. Inter-residue interactions in alpha-helical transmembrane proteins. *Bioinformatics* **35**, 2578–2584 (2019).

42. Vreven, T. *et al.* Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology* **427**, (2015).

43. Koukos, P. I., Faro, I., van Noort, C. W. & Bonvin, A. M. J. J. A Membrane Protein Complex Docking Benchmark. *J. Mol. Biol.* **430**, 5246–5256 (2018).

44. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

45. Jha, A. N., Vishveshwara, S. & Banavar, J. R. Amino acid interaction preferences in proteins. *Protein Sci.* **19**, 603–616 (2010).

46. Ye, L., Van Eps, N., Zimmer, M., Ernst, O. P. & Scott Prosser, R. Activation of the A2A adenosine G-protein-coupled receptor by conformational selection. *Nature* **533**, 265–268 (2016).

47. Thanos, C. D., DeLano, W. L. & Wells, J. A. Hot-spot mimicry of a cytokine receptor by a small molecule. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 15422–7 (2006).

48. Levy, E. D. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. *Journal of Molecular Biology* **403**, 660–670 (2010).

49. Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O. & Gursoy, A. PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. *Nucleic Acids Research* **42**, W285–W289 (2014).

50. Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J. & Oliva, B. BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic acids research* **40**, W147-51 (2012).

51. Sokolina, K. *et al.* Systematic protein-protein interaction mapping for clinically relevant human GPCRs. *Mol. Syst. Biol.* **13**, 918 (2017).

52. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic acids research* **38**, D497–501 (2010).

53. Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).

54. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).

55. Huang, S.-Y. Exploring the potential of global protein–protein docking: an overview and critical assessment of current programs for automatic ab initio docking. *Drug Discovery Today* **20**, 969–977 (2015).

56. Kotlyar, M., Pastrello, C., Malik, Z. & Jurisica, I. IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.* **47**, D581–D589 (2019).

57. Wodak, S. J., Vlasblom, J., Turinsky, A. L. & Pu, S. *Protein-protein interaction networks: The puzzling riches*. vol. 23 (2013).

58. Rao, V. S., Srinivas, K., Sujini, G. N. & Kumar, G. N. S. Protein-protein interaction detection: methods and analysis. *Int J Proteomics* **2014**, 147648 (2014).

59. Aloy, P. & Russell, R. B. The third dimension for protein interactions and complexes. *Trends Biochem. Sci.* **27**, 633–638 (2002).

60. Mathivanan, S. *et al.* An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* **7**, S19 (2006).

61. Alvarez-Ponce, D. Recording negative results of protein–protein interaction assays: an easy way to deal with the biases and errors of interactomic data sets. *Briefings in Bioinformatics* **18**, bbw075 (2016).

62. Ben-Hur, A. & Noble, W. S. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics* **7 Suppl 1**, S2 (2006).

63. Trabuco, L. G., Betts, M. J. & Russell, R. B. Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. *Methods* **58**, 343–348 (2012).

64. Smialowski, P. *et al.* The Negatome database: a reference set of non-interacting protein pairs. *Nucleic acids research* **38**, D540–4 (2010).

65. Blohm, P. *et al.* Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic acids research* **42**, D396–400 (2014).

66. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–68 (2005).

67. Capitani, G., Duarte, J. M., Baskaran, K., Bliven, S. & Somody, J. C. Understanding the fabric of protein crystals: computational classification of biological interfaces and crystal contacts. *Bioinformatics* **32**, 481–489 (2016).

68. Charollais, J. & Van Der Goot, F. G. Palmitoylation of membrane proteins (Review). *Molecular membrane biology* **26**, 55–66 (2009).

69. Liu, C. H., Li, K. C. & Yuan, S. Human protein-protein interaction prediction by a novel sequence-based co-evolution method: Co-evolutionary divergence. *Bioinformatics* (2013) doi:10.1093/bioinformatics/bts620.

70. Hamp, T. & Rost, B. More challenges for machine learning protein interactions. *Bioinformatics* **2**, 1–5 (2015).

71. Sun, T., Zhou, B., Lai, L. & Pei, J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* (2017) doi:10.1186/s12859-017-1700-2.

72. Yin, H. *et al.* Computational design of peptides that target transmembrane helices. *Science* **315**, 1817–1822 (2007).

73. Caputo, G. A. *et al.* Computationally designed peptide inhibitors of protein-protein interactions in membranes. *Biochemistry* **47**, 8600–8606 (2008).

74. Yin, H. & Flynn, A. D. Drugging Membrane Protein Interactions. *Annu Rev Biomed Eng* **18**, 51–76 (2016).

75. Stone, T. A. & Deber, C. M. Therapeutic design of peptide modulators of protein-protein interactions in membranes. *Biochim Biophys Acta Biomembr* **1859**, 577–585 (2017).

76. Murakami, K. *et al.* SOD1 (copper/zinc superoxide dismutase) deficiency drives amyloid β protein oligomerization and memory loss in mouse model of Alzheimer disease. *J. Biol. Chem.* **286**, 44557–44568 (2011).

77. Watanabe, T. *et al.* Nucleotide binding oligomerization domain 2 deficiency leads to dysregulated TLR2 signaling and induction of antigen-specific colitis. *Immunity* **25**, 473–485 (2006).

78. Scott, D. E., Bayly, A. R., Abell, C. & Skidmore, J. Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery* (2016) doi:10.1038/nrd.2016.29.

79. Corbi-Verge, C. & Kim, P. M. Motif mediated protein-protein interactions as drug targets. *Cell Communication and Signaling* (2016) doi:10.1186/s12964-016-0131-4.

80. Jubb, H. C. *et al.* Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in Biophysics and Molecular Biology* (2017) doi:10.1016/j.pbiomolbio.2016.10.002.

81. Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, (2014).

82. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204-212 (2015).

83. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049-1056 (2015).

84. Elez, K., Bonvin, A. M. J. J. & Vangone, A. Distinguishing crystallographic from biological interfaces in protein complexes: role of intermolecular contacts and energetics for classification. *BMC Bioinformatics* **19**, 438 (2018).

85. Jiménez-García, B., Elez, K., Koukos, P. I., Bonvin, A. M. & Vangone, A. PRODIGY-crystal: a web-tool for classification of biological interfaces in protein complexes. *Bioinformatics* **35**, 4821–4823 (2019).

86. Betts, M. J. & Russell, R. B. Amino Acid Properties and Consequences of Substitutions. in *Bioinformatics for Geneticists* 289–316 (John Wiley & Sons, Ltd, 2003). doi:10.1002/0470867302.ch14.

87. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).

88. de Vries, S. J., van Dijk, M. & Bonvin, A. M. J. J. The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc* **5**, 883–897 (2010).

89. van Zundert, G. C. P. & Bonvin, A. M. J. J. Modeling protein-protein complexes using the HADDOCK webserver 'modeling protein complexes with HADDOCK'. *Methods Mol. Biol.* **1137**, 163–179 (2014).

90. Kozakov, D. *et al.* The ClusPro web server for protein-protein docking. *Nat Protoc* **12**, 255–278 (2017).

91. Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* **20**, 45–50 (2004).

92. Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res.* **32**, W96-99 (2004).

93. Tovchigrechko, A. & Vakser, I. A. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.* **34**, W310-314 (2006).

94. Tovchigrechko, A. & Vakser, I. A. Development and testing of an automated approach to protein docking. *Proteins* **60**, 296–301 (2005).

95. Viswanath, S., Ravikant, D. V. S. & Elber, R. DOCK/PIERR: web server for structure prediction of protein-protein complexes. *Methods Mol. Biol.* **1137**, 199–207 (2014).

96. Viswanath, S., Dominguez, L., Foster, L. S., Straub, J. E. & Elber, R. Extension of a protein docking algorithm to membranes and applications to amyloid precursor protein dimerization. *Proteins* **83**, 2170–2185 (2015).