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

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

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

[2] Université Paris-Saclay, Inserm U1204, Université d'Evry, 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 or inhibiting the interaction. Yet, structural data coming from experiments is very scarce for this family of proteins. To overcome this problem, we propose a computational approach for the prediction of alpha transmembrane protein multimeric 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 binding sites used 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 new binding sites and to thousands of new binary complexes between membrane proteins when allowing amino acid mutations to take place. We generate an online database of the annotated predicted interactions.

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

INTRODUCTION

The membranome may be considered as the set of bitopic (single-spanning) and polytopic (multiple-spanning) transmembrane proteins (TM) that span the cell membrane in its entirety, i.e. the membrane protein interactome. These integral membrane proteins represent around one third of the proteomes of organisms (Stevens and Arkin, 2000) and come mostly in the form of alpha helix 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. As proteins are the core of the cell machinery in organisms (activation, transport, degradation, stabilization, and participation in the production of other proteins), and 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 knowledge of the interfaces needed for complex formation, as molecular recognition is a fundamental phenomenon governing all processes of life. (Kastritis and Bonvin, 2013)

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 of these proteins. Indeed, deficient oligomerization is associated with known diseases. (Yamamoto *et al.*, 2017; Guidolin *et al.*, 2018; Pin *et al.*, 2007)

Estimates of the total number of human PPIs range from 130,000 to over 600,000, one order of magnitude bigger than the *D melanogaster* interactome. (Bork *et al.*, 2004; Stumpf *et al.*, 2008; Venkatesan *et al.*, 2009) 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 (Iyer *et al.*, 2005) are not well suited for identifying membrane protein interactions. The data covering PPIs is increasing exponentially -in year 2012, there were more than 235,000 binary interactions reported. (Licata *et al.*,

2012) Most protein interaction databases – bioGRID, (Chatr-aryamontri *et al.*, 2017) BIND, (Alfarano *et al.*, 2005) STRING, (Szklarczyk *et al.*, 2017) IntAct, (Orchard *et al.*, 2014) MINT, (Licata *et al.*, 2012), and others - offer general information about experimentally validated PPIs of all types. The IMEx Consortium groups all nonredundant protein interaction data in one interface. (Orchard *et al.*, 2012) Nevertheless, these databases 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 the rules that apply to interactions between soluble segments are not necessarily valid within the membrane. Thus the difficulty in the experimental determination of their 3D structures (Carpenter *et al.*, 2008) while embedded within intact membranes.

Proteome-wide maps of the human interactome network have been generated in the past. (Babu *et al.*, 2012; Mosca *et al.*, 2013; Rolland *et al.*, 2014) Nevertheless, these assays, just like the Y2H assay, are depleted of interactions among proteins containing TM helices. 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. (Babu *et al.*, 2012; Mosca *et al.*, 2013) 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. (Petschnigg *et al.*, 2014) But to the best of our knowledge, MaMTH has not been used as a systematic screening assay to map the human membrane protein interactome. Thus, only few databases are specific for TM proteins.

On the other hand, there are many methods for the prediction of PPIs from sequence alone (PIPE2, (Pitre *et al.*, 2008) SigProd, (Martin *et al.*, 2005) MULTIPROSPECTOR; (Lu *et al.*, 2002) machine learning (PCA-EELM, (You *et al.*, 2013; Hamp and Rost, 2015a)); and from template structures (Szilagyi and Zhang, 2014)). But, again, most of the approaches address soluble globular proteins.(Keskin *et al.*, 2016) Qi et al developed a random forest learning machine approach but limited to the human membrane receptor interactome, HMRI. (Qi *et al.*, 2009) 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 transmembrane contact residues issued from membrane protein multimers, or their portions thereof reported in the PDB (Berman *et al.*, 2000) and validated by the OPM database, (Lomize *et al.*, 2006) in the context of the lipid bilayer. Querying thereafter the PDBsum database, (de Beer *et al.*, 2014; Laskowski *et al.*, 1997) we obtain the atomic details of the protein-protein interfaces, i.e. the residues that are in contact at the interface of the complexes. We then gather those amino acids at the recognition interface to generate regular expressions that represent in a linear form the interaction motifs in space. With this information in hand, we proceed to find the obtained motifs in other TM proteins by allowing a certain degree of amino acid mutations in the sequences of the motifs. The allowed mutation rates of the interface residues allow us to explore local, spatial non-homologous motifs. Homologs of the starting motifs are expected to interact in an analogous manner, (Aloy *et al.*, 2003) as opposed to a global structural approach that would find structurally, physically homologous PDB interfaces, limiting the search to functionally-related partners without paying attention to the particular sequence at the interface or to biological sense. In all cases, it is reasonable to assume that the number of interface motifs is limited in nature. (Keskin and Nussinov, 2005) 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 of the found proteins, such as in template-based predictions. (Zhang *et al.*, 2012) 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 membranomes may vary between species, tissues, cell types, and developmental stages, we focus in this work on the plasma membrane of same-species eukaryotes, ensuring thus that we are probing proteins with the same sub-cellular localization.

RESULTS

Fully automated pipeline for MPPI predictions

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 77 PDBsum files. After checking which PDBsum files to consider, we ended up with only 57 distinctly different interactions, associated to 52 unique reviewed UniProt entries for all species, of which 52 are structural homomers and 5 structural heteromers (Table 1). Of these, besides *H. sapiens*, four species are represented (*M. musculus, R. norvegicus, B. Taurus, A. thaliana,*). On another hand, there are 21 structures of complexes between bitopic proteins, 32 between polytopic proteins, and 1 mixed bitopic-polytopic protein complex (PDB ID 2L35, Table 1). This is thus the number of experimentally solved structures of TM protein complexes that we used as the template database. 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, except PDB ID 4OR2 (G protein-coupled metabotropic glutamate receptor 1). The NMR complexes are not submitted to the test, since they are formed in solution and in the absence of a crystallographic lattice.

The protein-protein docking benchmark 5.0 (PPDB) (Vreven *et al.*, 2015) 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 the 57 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 11 sets of structures (complexes and their unbound components, Table 1, residues1 fingernail, column C or Table 1A).

Table 1A

Complex	Unbound PDB ID A	Unbound PDB ID 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 (Koukos *et al.*, 2018), 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 TM segments.

The total number of motifs found after removing redundancies due to different chains of the same structures interacting more than once, was 98 (Table 1, MotifAB fingernail), grouped into 86 clusters based on consensus motifs (Table 2).

Table 1 (Supplemental) Col. A, PDB code ID of validated structures of membrane protein complexes protein A/protein B; Col. B, biological EPPIC assembly probability (PRODIGY Predicted interface Biological vs Crystallographic); Col. C, PDB ID for the unbound structures of the components of complexes, defining a benchmark of membrane protein-protein interactions; Col. D, protein A Uniprot ID; Col. E, protein B Uniprot ID; Col. H, number of TM helices of protomer in complex; Col. I, protein A binding motif; Col. J, protein B binding motif; Col. O, specific sequence of binding motif of protein A; Col. P, specific sequence of binding motif of protein B.

Table 2 (Supplemental) Consensus motifs obtained after multiple alignment of closely related individual motifs.

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. (Mayol *et al.*, 2019) Fig. 2 shows, in the shape of a symmetric “heatmap” 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 same-residue pair, and Trp-Cys for a different-residue 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 database for alpha helical MPs. (Mayol *et al.*, 2019) These contacts imply correlated mutations.

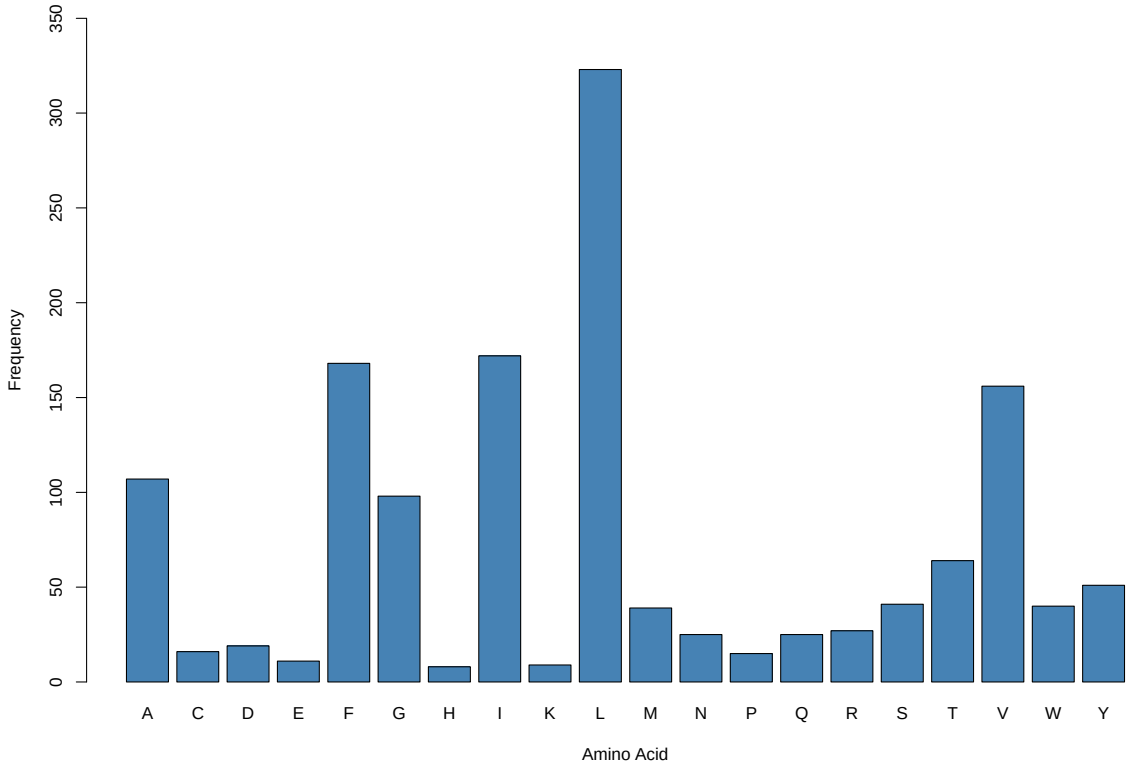


Fig. 1 Frequency of amino acids in the extracted motifs from the contact maps at the interfaces obtained from the PDBsum server.

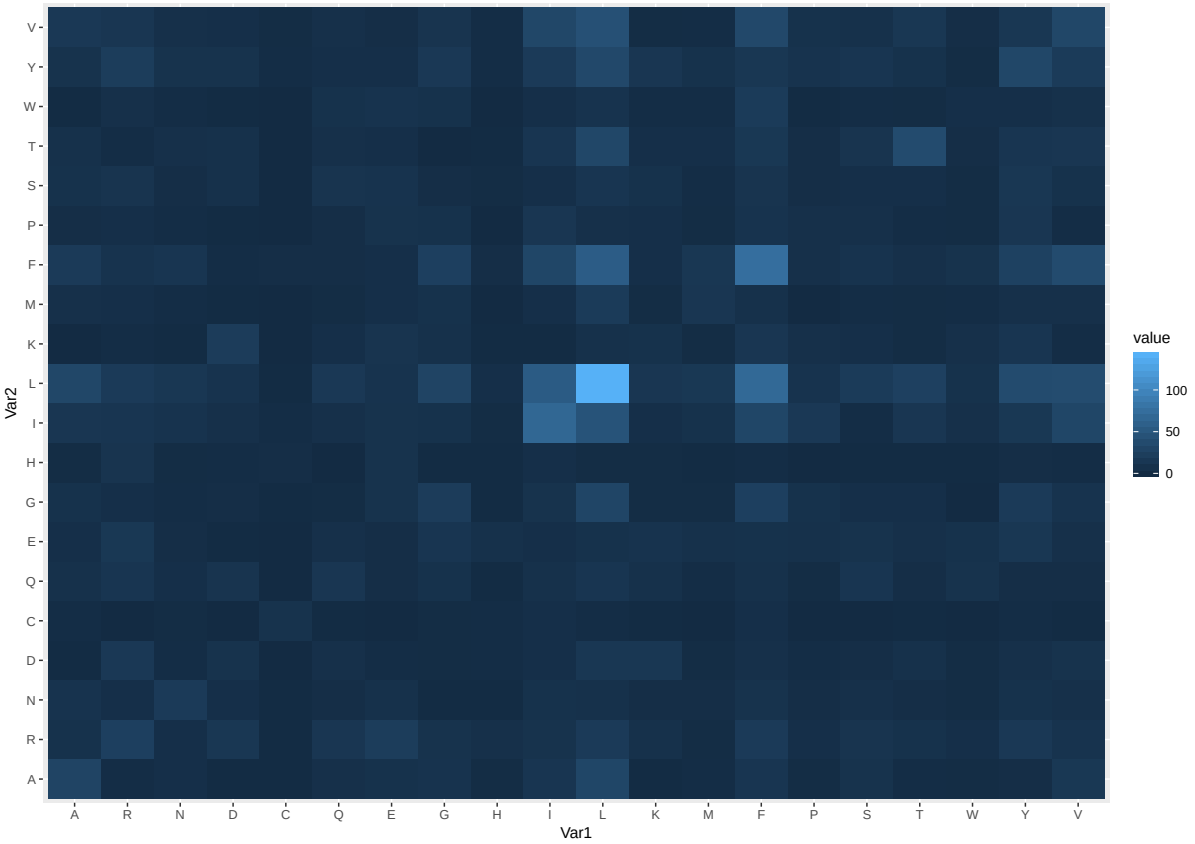
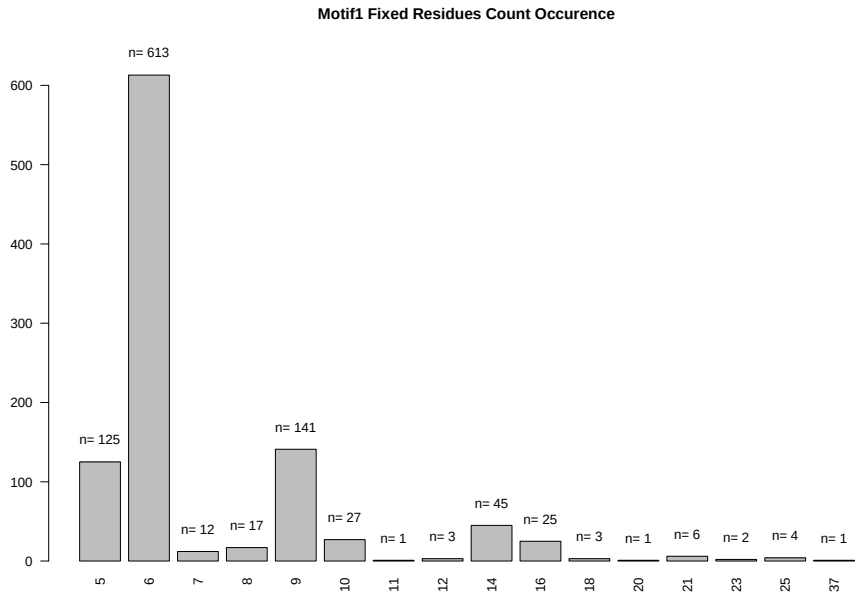


Fig. 2 Symmetrical heatmap of pairs of contact residues at the MP-MP interface of the template complexes. Amino acid residue names are represented by the one-letter code.

We then wanted to look at the number of motifs resulting for each number n of contact residues. As seen in Fig. 3, the count occurrence of contact residues is largest for n = 6, amounting to 613 for motifs A, and 610 for motifs B.

a)



b)

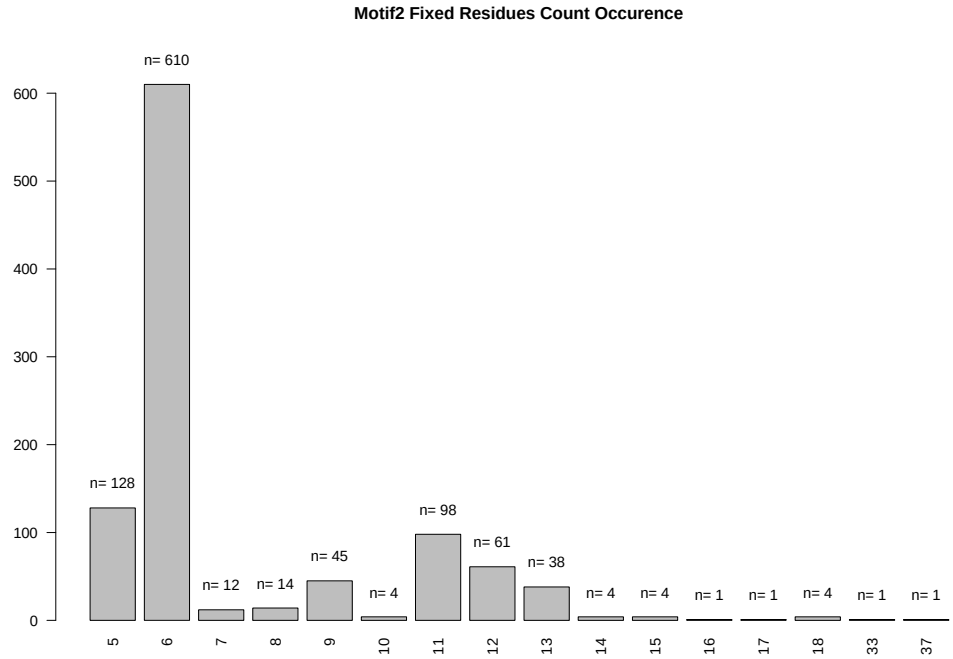


Fig. 3 Count occurrence of contact residues for five or more contact residues. a) motifs A; b) motifs B. Mutation rate ≤5%. Motifs with six contact residues occur the most.

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. We managed to predict many new sites for the corresponding motifs (Fig. 4). However, we noticed that more potential motifs were found when the original motifs had a lower number of contact residues (Fig. 5). Thus, we found small motifs with six to nine contact residues extremely abundant. As the number of residues increased, the number of hits decreased drastically until reaching merely 12 predictions at 39 contact residues, the longest motif. As a function of the mutation rate, TransINT predicts 1102 interactions across species for a mutation rate of 5%. For the mutations rates 10, 15 and 20%, the number of interactions is of 9966, 24003 and 2656, respectively.

As mentioned in the M&M section, we built a consensus sequence for all the matched sequences of a given binding site, considering the contact residues only. The purpose was to see if there was a conservation of amino acid residues or not. The most prevalent consensus motif found for a mutation rate >0% was C.{6}VV.{2}V.{4}.{2}W (Fig. 5c). Several least common motifs for mutation rates >0% are observed, among which C.{14}DL.{2}T.{3}AL.{2}Y.{3}R. Fig. 6 shows the consensus of a motif of length 10 of 47 different protein sequences. Out of the 10 positions, only two show significant variations. The rest showed high conservation in all sequences predicted.

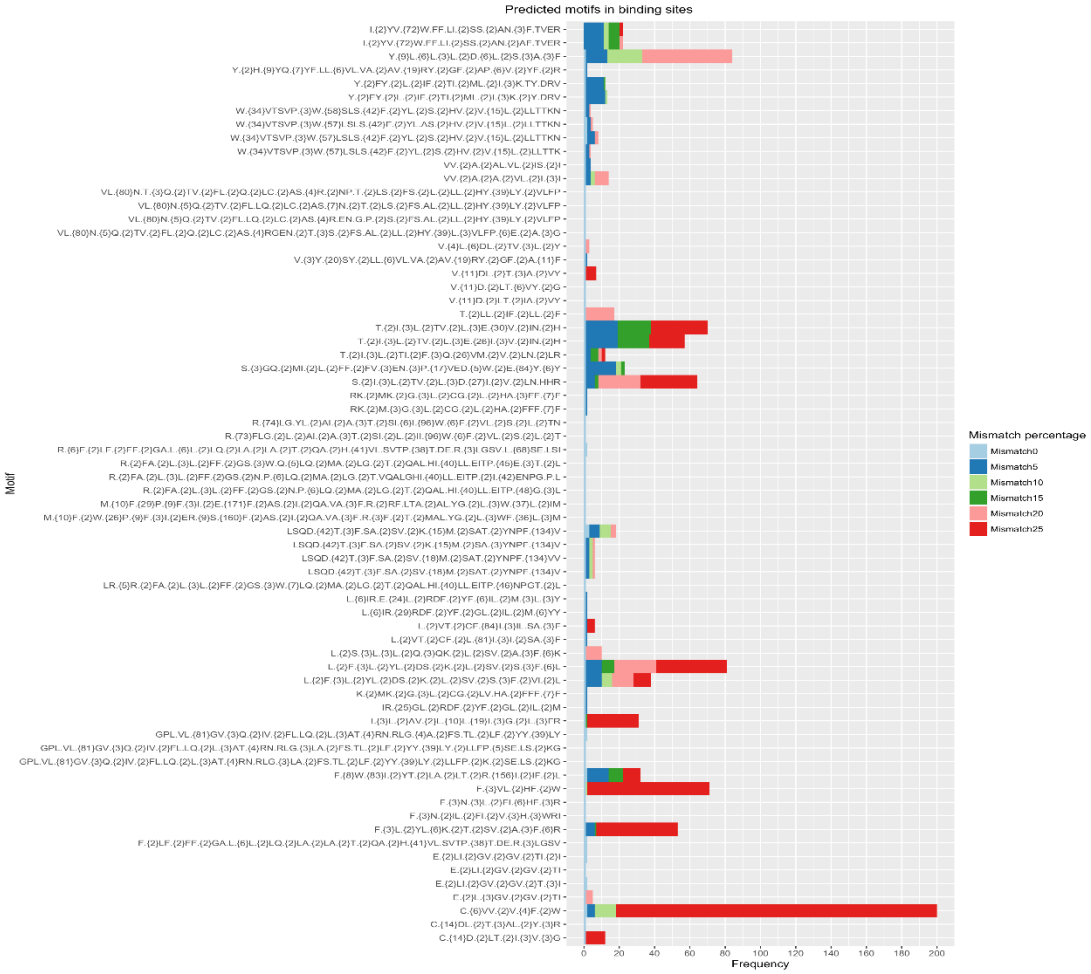
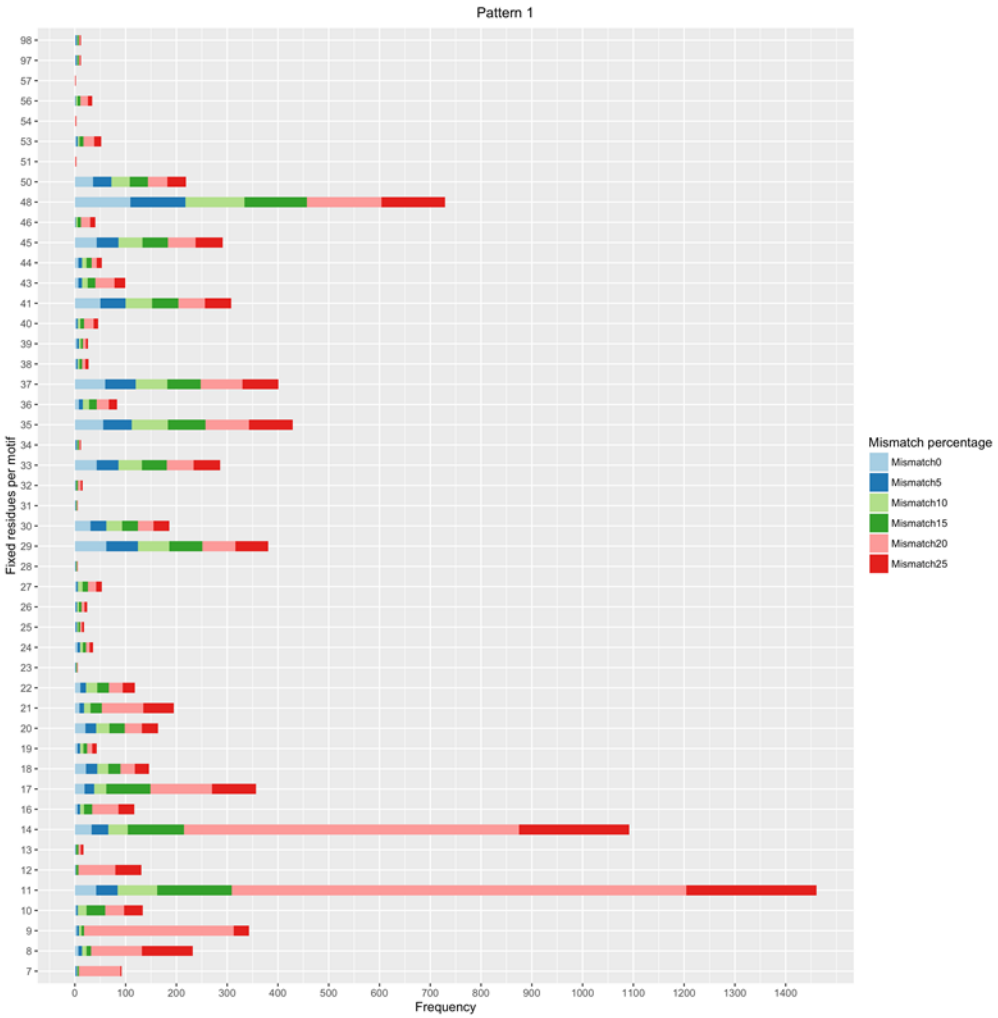
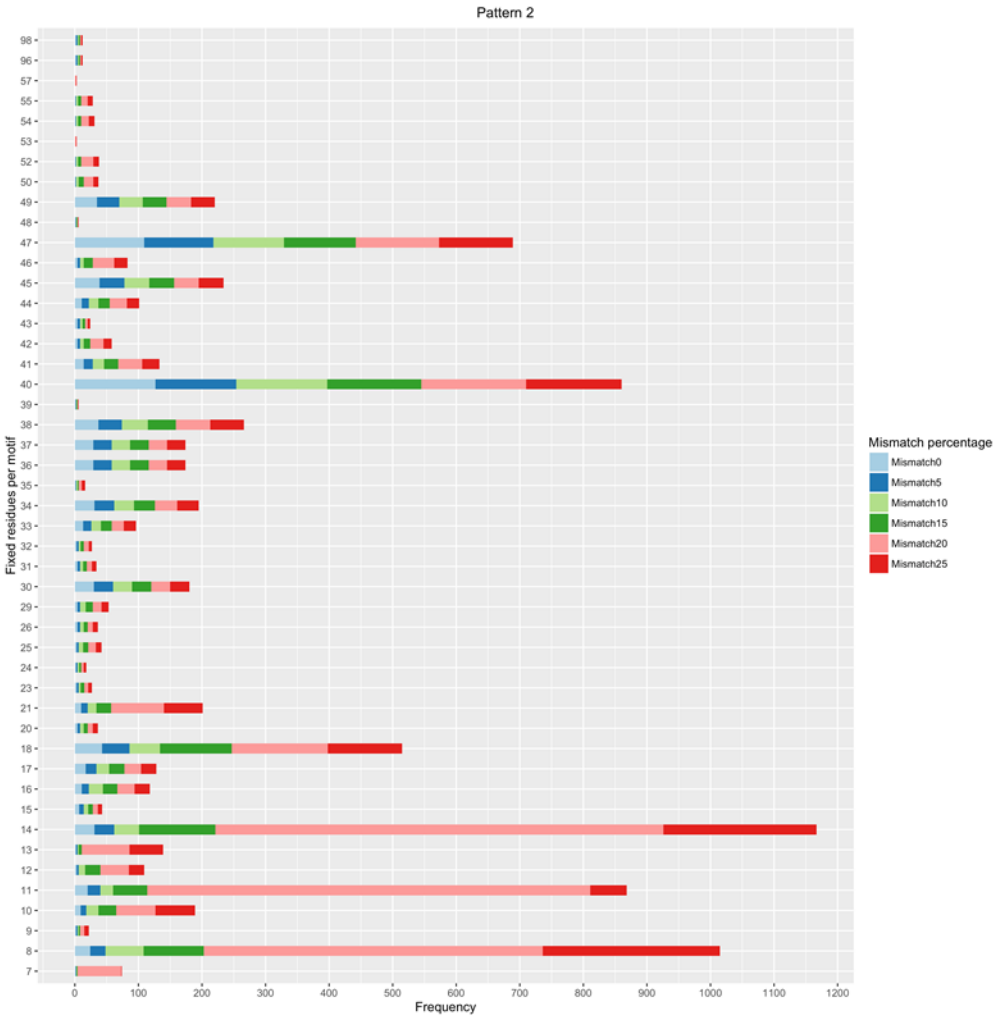


Fig. 4 Number of hits per motif.

a) Motif A



b) Motif B



C) All A and B motifs combined

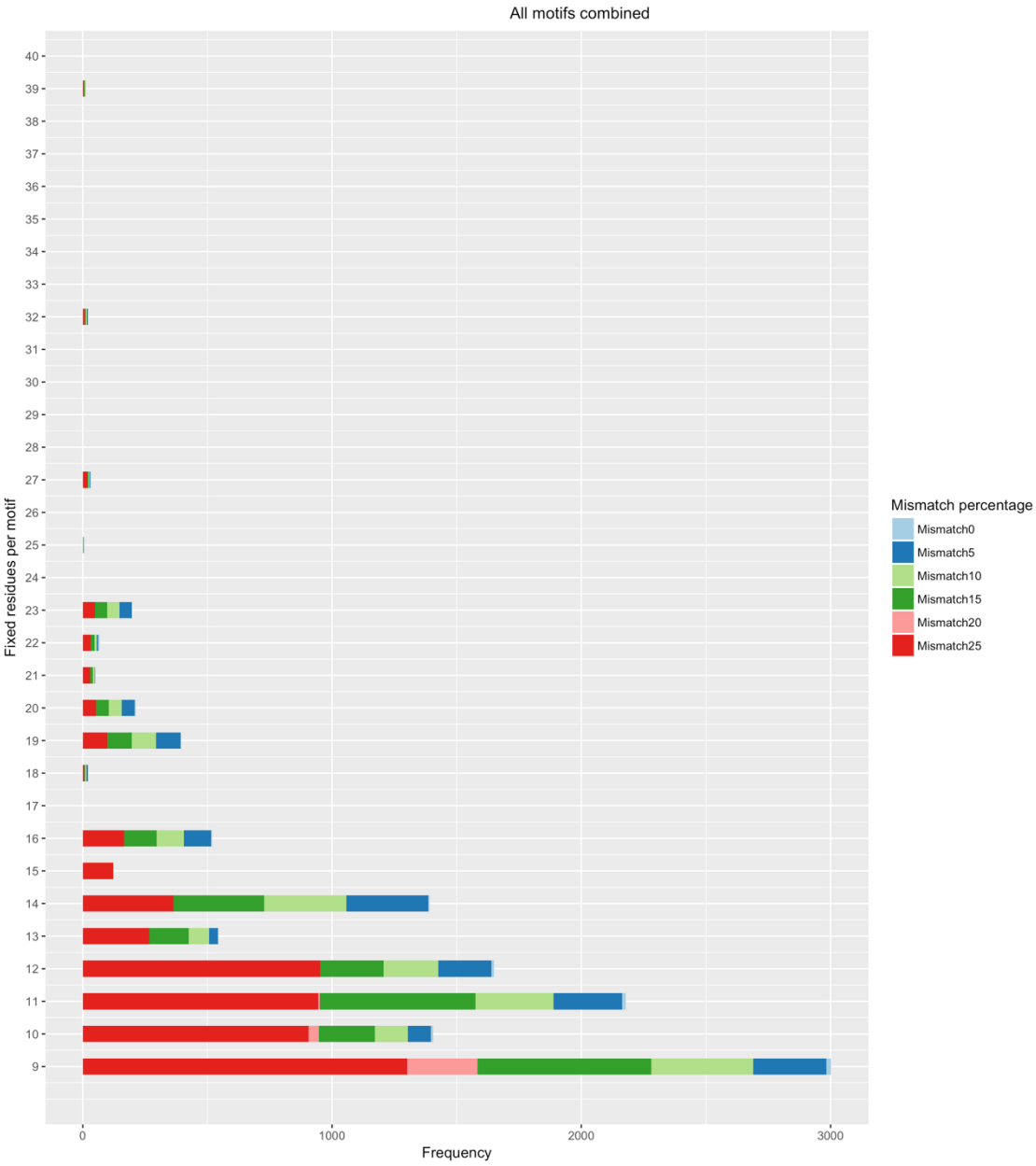


Fig. 5: Frequency of predicted proteins per number of contact residues for different mutation rates. In this graph, the mutation rates of protein A and of protein B have the same value. Number of contact residues 6, 7 and 8 were omitted due to the huge number of hits for each: 2,084,494, 159,135 and 8747, respectively.

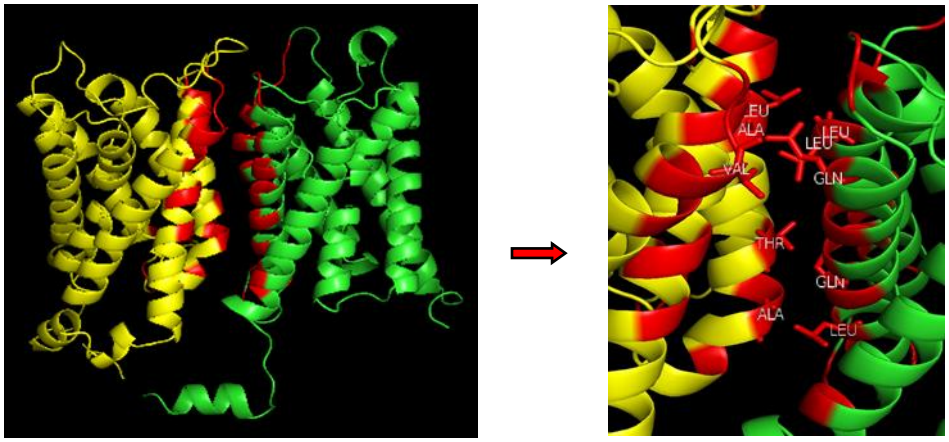
Fig. 6: Logo showing consensus for a motif of length 10. Not all sequences are shown as there are too many.

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 and AQP2 protomers and proceeded to a molecular docking as described in the M&M section. In

this case, the crystal structures existed for both -PDB 3D9S and 4NEF, respectively. Fig. 7a shows the resulting heterodimer with the interface regions. We then selected the predicted rat AQP2-ErbB-3 pair. Since the experimental structures are not available for either MP, we homology-modelled them using human AQP2 and human ErbB-3, respectively. Both aquaporins are 90% sequence identical, just like the ErB-3s; the resulting models are thus highly reliable. Fig 7b shows that the modeled rat AQP2-ErbB-3 heterodimer has the contact residues also at the interface and is thus feasible.

a



b

Fig. 7: Low-resolution cartoon structural model of predicted MP-MP complexes obtained by steered molecular docking. a) *H. sapiens'* protomers of aquaporin 5 (Uniprot P55064, PDB 3D9S, AQP5) and aquaporin 2 (P41181, 4NEF, AQP2) in complex, with interface residues in red. The figure to the right shows a zoom of the interface with the contact residues; b) *Rattus norvegicus'* protomers of aquaporin 2 (P34080, AQP2) and the receptor tyrosine-protein kinase erbB-3 (Q62799, Erbb3) in complex, with interface residues in purple and in red, respectively. In both complexes, the contact residues are indeed at the interface of the complex.

DISCUSSION

In this work, we developed a template sequence-based protocol destined to predict in large scale binary interactions among the alpha transmembrane segments of TM proteins. We illustrate the protocol in the case of the eukaryote plasma cell membrane. The highly matched motifs in different 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. Moreover, as we found, amino acids tend to interact the most with other amino acids sharing the same properties, such as hydrophobicity. (Jha *et al.*, 2010) 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. (Nath Jha *et al.*, 2011) This is the opposite to globular proteins, whose interaction interface is formed mainly of hydrophilic residues (the hydrophobic effect). 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, (Thanos *et al.*, 2006) as we did not perform any energy calculations nor estimate the binding affinities.

Some of the proteins, like the mouse glutamate receptor 3 (Gria3, Q9Z2W9) show more than one binding motif, suggesting a promiscuous binding behavior. (Levy, 2010) Thus, in this ligand-gated ion channel, we detect three distinctly different motifs: I.{2}YT.{177}VF.{2}L.{3}L.{2}AM, V.{2}V.{57}S.{3}V.{2}VW.{2}F.LI.{2}SS, and I.{2}A.{3}A.{6}V.{60}S.{3}V.{2}V.{2}FF.LI.{2}SS.TA that may explain the multiplicity of binding modes of many proteins, as well as a multifunctionality. For instance, Gria3 shows probable physical interaction with seven other proteins at the membrane level, including Disks large homolog 4, Disks large-associated protein 1, and Glutamate receptor 2 (IntAct Molecular Interaction Database https://www.ebi.ac.uk/intact/) (Orchard *et al.*, 2014), potentially using one or the other of the binding motifs generated by our work. Conversely, a given binding motif may bind to proteins bearing different

binding motifs. For example, I.{2}YT.{174}V.{2}VF.IL.{3}L.LAM.VA may bind to I.{2}A.{3}A.{3}V.{2}V.{2}LV.{56}IV.{2}V.{2}FF.{2}II and to A.{3}A.{2}GV.{2}V.{2}L.{54}S.{2}IV.{2}V.{5}L.{3}S. Also I.{3}L.{2}T.{6}R.{12}D.{2}MA.{6}F.{2}LL may bind to I.{2}YT.{177}VF.{2}L.{3}L.{2}AM and to GYY.IQ.{6}L.{2}IL.{2}V.{13}A.{3}T.{3}T.{3}Q. The interface recognition motifs produced here, while being mostly constituted of hydrophobic residues show a large diversity, yet accounting for a limited set of strictly different motifs.

The PPI template-based prediction algorithm and server PRISM2.0 (Baspinar *et al.*, 2014) (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 of TM interfaces. 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 water-soluble 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 (Qi *et al.*, 2009), 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 (Garcia-Garcia *et al.*, 2012) is based on sequence homology between proteins found in PPI databases 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 predictions at a 5% mutation rate for *H. sapiens* only to the FPClass dataset of genome-wide predicted human PPIs (Kotlyar *et al.*, 2015), a data mining-based method, we find several consistencies, listed in Table 3. Table 4 lists the correspondence between the 20% mutation rate predictions of TransINT for human proteins and those of FPClass.

UniProt AC A	Gene name B	UniProt AC B	FpClass total score
P05023	ATP1B1	P05026	0.8826
P42261	GRIA4	P48058	0.8826
P42261	GRIA2	P42262	0.8826
P42261	GRIA3	P42263	0.8826
P42262	GRIA3	P42263	0.8826
P42262	GRIA4	P48058	0.8826
P42263	GRIA4	P48058	0.8826
P50993	ATP1B1	P05026	0.4916

Table 3 Predicted interactions common to TransINT (5% mutation rate, *H. sapiens*) and FPClass.

UniProt AC A	Gene name B	UniProt AC B	FpClass total score
P18507	GABRB2	P47870	0.4029
P18507	GABRD	O14764	0.2902
P18507	GABRG3	Q99928	0.2873
P31644	GABRB1	P18505	0.5405
P31644	GABRD	O14764	0.2593
P34903	GABRA6	Q16445	0.2915
P34903	GABRB2	P47870	0.2803
P42261	GRIA2	P42262	0.8826
P42261	GRIA3	P42263	0.8826
P42261	GRIA4	P48058	0.8826
P42262	GRIA3	P42263	0.8826
P42262	GRIA4	P48058	0.8826
P42263	GRIA4	P48058	0.8826

P47869	GABRB2	P47870	0.2949
P47869	GABRD	O14764	0.2741
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
P50993	ATP1B1	P05026	0.4916
Q01814	ATP2B3	Q16720	0.3235
Q13683	FCER2	P06734	0.3151
Q16445	GABRB2	P47870	0.2837
Q16445	GABRA4	P48169	0.2786
Q16720	ATP2B2	Q01814	0.3235

Table 4 Predicted interactions common to TransINT (20% mutation rate, *H. sapiens*) and FPClass.

Although most datasets based on experimental approaches (non-exhaustive list: GPCR interactome (Sokolina *et al.*, 2017); IMEx consortium (http://www.imexconsortium.org/); IntAct (www.ebi.ac.uk/intact/); Human Protein Reference Database (hprd.org); HINT (http://hint.yulab.org/); CORUM (Ruepp *et al.*, 2010; Giurgiu *et al.*, 2019) (http://mips.helmholtz-muenchen.de/corum/); API (http://cicblade.dep.usal.es:8080/APID/init.action); HI-II-14 (Rolland *et al.*, 2014) (http://interactome.dfci.harvard.edu/H_sapiens/); HuRI (http://interactome.baderlab.org/); BioPlex (Huttlin *et al.*, 2015), (http://bioplex.hms.harvard.edu/); QUBIC (Hubner and Mann, 2011); IID (Kotlyar *et al.*, 2019)) 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 (Wodak *et al.*, 2013; Rao *et al.*, 2014)).

The experimental work of Rolland and his colleagues on the human interactome network found 13944 interactions, mapping 4303 unique Gene IDs, out of which 4188 were found to correspond to a total of 4198 proteins found in UniProt (Rolland *et al.*, 2014). 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 found. Moreover, amongst the interactions found, only 41 were between TM proteins. Twenty-eight of these proteins were found to be interacting in the IntAct database. Nevertheless, none of the interactions we got from the structural PDBsum database 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. Consequently, we did not find any of our predictions in their results. As for the BioPlex 3.0 database, using the TAP-MS technology for detecting PPI, the authors found more than 2,000 TM interactions out of ~73,000 stored in their database. TransINT predicts interactions between human Gamma-aminobutyric acid receptor subunit beta-3 (P28472) and Gamma-aminobutyric acid receptor subunit beta-2 (P47870 and between P28472 and Gamma-aminobutyric acid receptor subunit alpha-6 (Q16445). These interactions are confirmed in the BioPlex database (BioPlex p(interaction)>0.99 for both interactions). When assessing our data for a 5% mutation rate for *H. sapiens* with the IID database , we find that the predicted sodium/potassium-transporting ATPase subunit alpha-1 and beta-1 interaction is validated experimentally s are P05023 – P05026. With TransInt, we predict an interaction between Gamma-aminobutyric acid receptor subunit beta-3 (P28472) and Gamma-aminobutyric acid receptor subunit beta-2 (p47870), Gamma-aminobutyric acid receptor subunit gamma-3, Gamma-aminobutyric acid receptor subunit alpha-6, and Gamma-aminobutyric acid receptor subunit alpha-4. That these protomers are in the same complex is reported in HPRD.org and detected *in vivo*.

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. (Pitre *et al.*, 2008; Aloy and Russell, 2002) 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. (Mathivanan *et al.*, 2006)

The importance of recording negative results of PPI assays in interatomic databases, i.e. those indicating that the tested proteins do not interact, has been raised. (Alvarez-Ponce, 2016) 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 (Ben-Hur and Noble, 2006; Trabuco *et al.*, 2012). Looking for Negative interactions in the IntAct database for our 52 unique reviewed UniProt proteins for all species, we find only one negative interaction with another TM protein, that of Q9UNQ0 and isoform 1 of P11309 (a serine/threonine-protein kinase). Indeed, this negative interaction is absent from our positive interactions. In a later version of TransInt, we intend to go through the Negatome (Smialowski *et al.*, 2010; Blohm *et al.*, 2014), and Stelzl (Stelzl *et al.*, 2005) datasets that compile sets of protein pairs that are unlikely to engage in physical direct interactions in order to improve the ROC and MCC values. For the time being, we can observe that spanning the Negatome dataset with our predicted positive interactions for a 20% mutation rate (*H. sapiens*) results in no negative interactions. For example, Negatome lists a negative P08588-P24588 interaction. Indeed, according to TransINT, P24588 is absent among the interactions established by P08588 (Table 5).

	Negatome	TransINT
Protein A	Protein B	Protein B
P08588	P24588	Q8TAC9
P21917	Q53G59	Q9BZJ7, ∗
P22455	P09038	Q8TAC9
P23634	P27348	∗
	P63104	
P61073	P01303	Q8TAC9
	P10145	
	P55209	
Q9NPY3	P02745	∗

∗ Many other proteins, none of which is reported by Negatome.

Table 5 Comparison between the negative interactions in the Negatome dataset and the predicted positive interactions for a 20% mutation rate (*H. sapiens*) in the TransINT dataset.

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* crystal multimers reported in the PDB exist *in vivo* and are not a product of the crystal lattice or experimental conditions, being thus transient complexes; (Capitani *et al.*, 2016) 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, homomers 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 juxta-membrane regions of TM proteins, namely in the lumen. (Charollais and Van Der Goot, 2009) 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, a bottleneck in our *de novo* predictions exists due to the incompleteness of the membrane protein-protein complex 3D structure library, i.e. the PDB.

Within the limits of our assumptions, 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 database may be used as candidates to be validated experimentally.

CONCLUSION

The TransINT database 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 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. It is a database 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 membrane interactome in which the binding sites involve specifically TM regions are reported. Our membrane interactome (shiny) contains 43,059 entries for all species dealt with, and 15,226 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 databases 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, (Liu *et al.*, 2013; Hamp and Rost, 2015b; Sun *et al.*, 2017; Hopf *et al.*, 2014) our 1D-3D approach adds the spatial dimension to a given membrane interactome, 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 mutations in the motifs has allowed us to extend the initial set of template 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. In later work, we intend to apply machine-learning techniques to train a dataset with a diverse set of descriptors to increase the accuracy and extent of our MPPI predictions (amino acid residue type and conservation, hydrophobic/hydrophilic/polar character, H-bond formation capability, packing formers, molecular surface, etc.). Indeed, deficient or enhanced oligomerization is associated with diseases. (Murakami *et al.*, 2011; Watanabe *et al.*, 2006)

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. (Yin *et al.*, 2007; Caputo *et al.*, 2008; Stone and Deber, 2017; Yin and Flynn, 2016) Indeed, many of the TM proteins in the predicted complexes are involved in a variety of diseases (OMIM database https://www.omim.org/). In addition, our protein interaction data implies physical interactions and can lead to the construction of protein interaction networks. Thus, our results may help understand signaling pathways, the crosstalk between them and transactivation, may suggest potential drug targets, such as those targeting the MPPI interface, (Scott *et al.*, 2016; Corbi-Verge and Kim, 2016) and may aid in understanding the functional effects of mutations at the interface. (Jubb *et al.*, 2017)

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. In addition, its predictions can be refined, i.e. the number of false positives reduced by insuring the MPs belong to the same developmental stage, tissue, cell type, site of expression or cellular localization, reaction and metabolic pathways (https://reactome.org/), whose protomers in a complex do not show a gene distance of more than 20 and who are functionally similar.

MATERIALS AND METHODS

Sources of information retrieval and data filtering

We collected information from several publicly available databases. Fig. 8 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; (UniProt Consortium, 2015)) with cellular component annotations that matched the following GO (Gene Ontology Consortium, 2015) annotations: "integral component of membrane" (GO:0016021) and "plasma membrane" (GO:0005886), or "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 (Berman *et al.*, 2000) 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 3.5A or less and should 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, etc. 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 a natural variant sequence in the UniProt database 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 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 indicates. Of course, are not instanced 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, they do not interact, like in the human sigma-1 receptor (PDB 5HK2). Therefore, our approach includes in an implicit fashion the intra-molecular interactions that may be taking place among the protomers composing an oligomer.

Finally, to ensure that the oligomer structures that we took into account are biological quaternary structures, we used EPPIC, a protein-protein interface classifier (Capitani *et al.*, 2016), and PRODIGY, a classifier of biological interfaces in protein complexes (Elez *et al.*, 2018; Jiménez-García *et al.*, 2019) 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 database (Lomize *et al.*, 2006) 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. We double-checked the chosen PDB structures in the MPStruc database of membrane proteins of known 3D structure (White SH. Membrane proteins of known 3D structures. https://blanco.biomol.uci.edu/mpstruc/). PDBsum (Laskowski *et al.*, 1997) is a database 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 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 by a dot, and the curly braces for the consecutive occurrences of the dots, such as in the pattern E.{2}LI.{2}GV.{2}T.{3}I.

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, the cost is of 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 vary from 0 to 6.

To keep track of which pattern A motif is interacting with which pattern 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 (prop_cost), this time based on an amino acid substitution matrix for membrane proteins. (Jones *et al.*, 1994) updated by RB Russell (Betts and Russell, 2003) and corrected by our means. The higher the score, the more accurate our prediction was. These parameter's values oscillate between 10 and 70. (TransINT). 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 ones 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 database 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, (Dominguez *et al.*, 2003; de Vries *et al.*, 2010; van Zundert and Bonvin, 2014) ClusPro, (Kozakov *et al.*, 2017; Comeau *et al.*, 2004b, 2004a) and GRAMM-X. (Tovchigrechko and Vakser, 2006, 2005). 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 hand-filtered out non parallel, 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 high 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, (Viswanath *et al.*, 2014, 2015) nor to be concerned if the docking program was trained on sets composed primarily of soluble proteins.

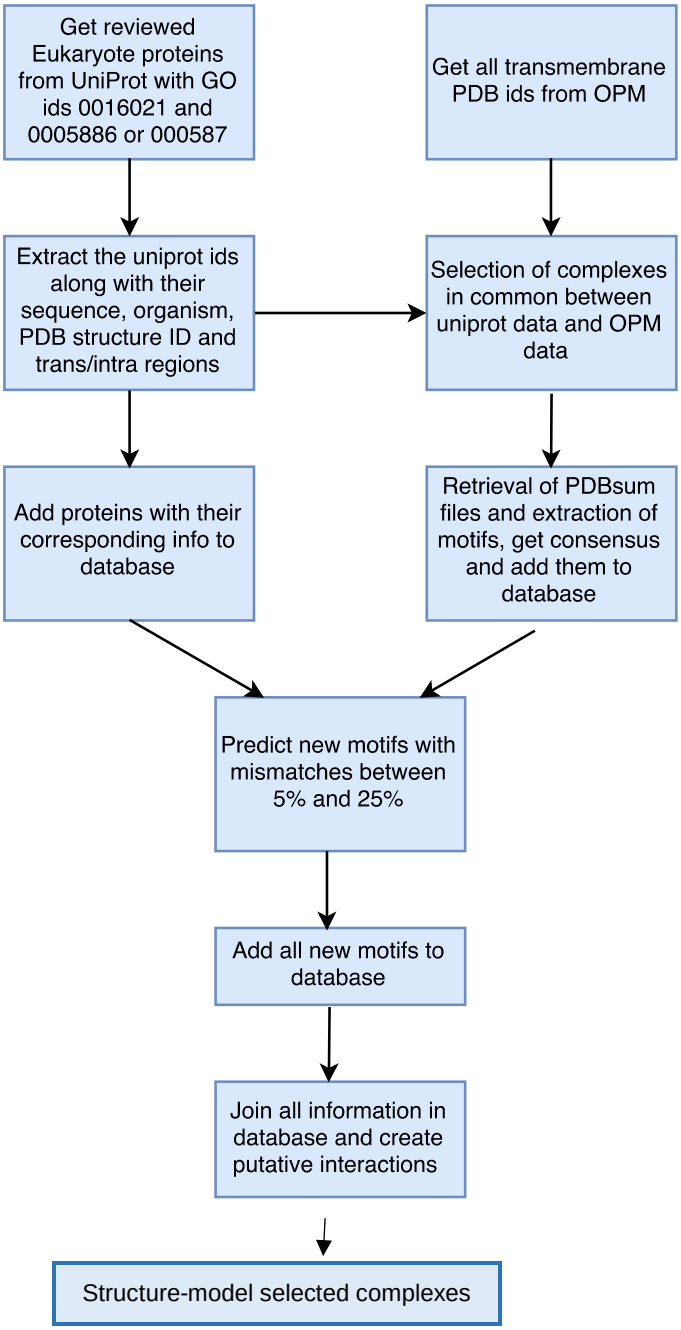


Fig. 8: Workflow illustrating the TransINT algorithm, from information retrieval to detection of recognition motifs to generation of putative interactions to 3D modeling of complexes.

The TransINT database

Thereafter, we built a homogeneous database named TransINT that contains all the found interactions. The database is the result of the implementation of a fully automated template-based recognition site search pipeline. We used a MySQL (version 5.0.11) database to keep all the information collected in an orderly manner. To access the database 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 database -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 database 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 database 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. French government fellowship.

REFERENCES

Alfarano,C. *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*.

Aloy,P. *et al.* (2003) The relationship between sequence and interaction divergence in proteins. *Journal of molecular biology*, **332**, 989–98.

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

Alvarez-Ponce,D. (2016) 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.

Babu,M. *et al.* (2012) Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. *Nature*, **489**, 585–9.

Baspinar,A. *et al.* (2014) PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. *Nucleic Acids Research*, **42**, W285–W289.

de Beer,T.A.P. *et al.* (2014) PDBsum additions. *Nucleic acids research*, **42**, D292-6.

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

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

Betts,M.J. and Russell,R.B. (2003) Amino Acid Properties and Consequences of Substitutions. In, *Bioinformatics for Geneticists*. John Wiley & Sons, Ltd, pp. 289–316.

Blohm,P. *et al.* (2014) 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.

Bork,P. *et al.* (2004) Protein interaction networks from yeast to human. *Current opinion in structural biology*, **14**, 292–9.

Capitani,G. *et al.* (2016) Understanding the fabric of protein crystals: computational classification of biological interfaces and crystal contacts. *Bioinformatics*, **32**, 481–489.

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

Carpenter,E.P. *et al.* (2008) Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, **18**, 581–586.

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

Chatr-aryamontri,A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, **45**, D369–D379.

Comeau,S.R. *et al.* (2004a) ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res.*, **32**, W96-99.

Comeau,S.R. *et al.* (2004b) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, **20**, 45–50.

Corbi-Verge,C. and Kim,P.M. (2016) Motif mediated protein-protein interactions as drug targets. *Cell Communication and Signaling*.

Dominguez,C. *et al.* (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**, 1731–1737.

Elez,K. *et al.* (2018) Distinguishing crystallographic from biological interfaces in protein complexes: role of intermolecular contacts and energetics for classification. *BMC Bioinformatics*, **19**, 438.

Garcia-Garcia,J. *et al.* (2012) BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic acids research*, **40**, W147-51.

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

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

Guidolin,D. *et al.* (2018) G protein-coupled receptor-receptor interactions give integrative dynamics to intercellular communication. *Reviews in the neurosciences*, **0**.

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

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

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

Hubner,N.C. and Mann,M. (2011) Extracting gene function from protein–protein interactions using Quantitative BAC InteraCtomics (QUBIC). *Methods*, **53**, 453–459.

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

Iyer,K. *et al.* (2005) Utilizing the Split-Ubiquitin Membrane Yeast Two-Hybrid System to Identify Protein-Protein Interactions of Integral Membrane Proteins. *Science Signaling*.

Jha,A.N. *et al.* (2010) Amino acid interaction preferences in proteins. *Protein Sci.*, **19**, 603–616.

Jiménez-García,B. *et al.* (2019) PRODIGY-crystal: a web-tool for classification of biological interfaces in protein complexes. *Bioinformatics*, **35**, 4821–4823.

Jones,D.T. *et al.* (1994) A mutation data matrix for transmembrane proteins. *FEBS letters*, **339**, 269–75.

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

Kastritis,P.L. and Bonvin,A.M.J.J. (2013) On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of the Royal Society, Interface*, **10**, 20120835.

Keskin,O. *et al.* (2016) Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. *Chemical Reviews*, **116**.

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

Kotlyar,M. *et al.* (2019) IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.*, **47**, D581–D589.

Kotlyar,M. *et al.* (2015) In silico prediction of physical protein interactions and characterization of interactome orphans. *Nature methods*, **12**, 79–84.

Koukos,P.I. *et al.* (2018) A Membrane Protein Complex Docking Benchmark. *J. Mol. Biol.*, **430**, 5246–5256.

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

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

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

Licata,L. *et al.* (2012) MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Research*.

Liu,C.H. *et al.* (2013) Human protein-protein interaction prediction by a novel sequence-based co-evolution method: Co-evolutionary divergence. *Bioinformatics*.

Lomize,M.A. *et al.* (2006) OPM: Orientations of Proteins in Membranes database. *Bioinformatics*, **22**, 623–625.

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

Martin,S. *et al.* (2005) Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**, 218–226.

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

Mayol,E. *et al.* (2019) Inter-residue interactions in alpha-helical transmembrane proteins. *Bioinformatics*, **35**, 2578–2584.

Mosca,R. *et al.* (2013) Towards a detailed atlas of protein–protein interactions. *Current Opinion in Structural Biology*, **23**, 929–940.

Murakami,K. *et al.* (2011) 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.

Nath Jha,A. *et al.* (2011) Amino acid interaction preferences in helical membrane proteins. *Protein Engineering Design and Selection*, **24**, 579–588.

Orchard,S. *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature methods*, **9**, 345–50.

Orchard,S. *et al.* (2014) The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*.

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

Pin,J.-P. *et al.* (2007) 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.

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

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

Rao,V.S. *et al.* (2014) Protein-protein interaction detection: methods and analysis. *Int J Proteomics*, **2014**, 147648.

Rolland,T. *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, **159**, 1212–26.

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

Scott,D.E. *et al.* (2016) Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery*.

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

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

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

Stevens,T.J. and Arkin,I.T. (2000) Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins: Structure, Function and Genetics*.

Stone,T.A. and Deber,C.M. (2017) Therapeutic design of peptide modulators of protein-protein interactions in membranes Elsevier.

Stumpf,M.P.H. *et al.* (2008) Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 6959–64.

Sun,T. *et al.* (2017) Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics*.

Szilagyi,A. and Zhang,Y. (2014) Template-based structure modeling of protein-protein interactions. *Current Opinion in Structural Biology*, **24**, 10–23.

Szklarczyk,D. *et al.* (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, **45**, D362–D368.

Thanos,C.D. *et al.* (2006) 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.

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

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

Trabuco,L.G. *et al.* (2012) Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. *Methods*, **58**, 343–348.

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

Venkatesan,K. *et al.* (2009) An empirical framework for binary interactome mapping. *Nature methods*, **6**, 83–90.

Viswanath,S. *et al.* (2014) DOCK/PIERR: web server for structure prediction of protein-protein complexes. *Methods Mol. Biol.*, **1137**, 199–207.

Viswanath,S. *et al.* (2015) Extension of a protein docking algorithm to membranes and applications to amyloid precursor protein dimerization. *Proteins*, **83**, 2170–2185.

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

de Vries,S.J. *et al.* (2010) The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc*, **5**, 883–897.

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

Wodak,S.J. *et al.* (2013) Protein-protein interaction networks: The puzzling riches.

Yamamoto,K. *et al.* (2017) 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.

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

Yin,H. and Flynn,A.D. (2016) Drugging Membrane Protein Interactions. *Annual review of biomedical engineering*, **18**, 51–76.

You,Z.-H. *et al.* (2013) Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC bioinformatics*, **14 Suppl 8**, S10.

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

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