TITLE

Predicting transmembrane protein-protein complexes: TransINT, a novel computational approach

Georges Khazen[1,*], Aram Gyulkhandanian[2], Tina Issa[1] and Rachid C. Maroun[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

*To whom correspondence should be addressed. Email: charbel.maroun@inserm.fr, georges.khazen@gmail.com

ABSTRACT

Because of their number and diversity, membrane proteins and their mutual complexes represent potential pharmacological targets par excellence for a variety of diseases. This has enormous implications for the design and discovery of new compounds that modulate the interactions between proteins. However, experimental structural data are very scarce and sequence data grows exponentially. To overcome this problem, we devised TransINT, a computational approach for the prediction of higher-order structures of α-helical transmembrane proteins. The approach involved data mining, filtering, and identification and characterization of the amino acid residues at the interface of the complexes as interaction motifs. These latter were expressed in a mathematical formalism used for motif search in databases, and prediction of membrane protein-membrane protein complexes. The number of predicted plasma membrane protein heteromers for the 39 species dealt with is 21,544 for 1,504 membrane proteins. Our template molecular interface-based approach using experimental recognition sites to formulate interaction motifs predicted thousands of binary complexes between plasma membrane proteins across species. The TransINT online database with the annotated predicted interactions is implemented as a web server and can be accessed at https://transint.shinyapps.io/transint/ or http://sertitan.imeve.univ-evry.fr:3838/transint/.

Contact: charbel.maroun@inserm.fr, georges.khazen@gmail.com

Supplementary information: Available online.

Keywords: protein-protein interactions; bioinformatics; biostatistics; integral membrane proteins; computational biology; prediction method; binary interactions; molecular recognition; protein complexes; data mining.

INTRODUCTION

Proteins are the core of cell machinery in organisms and membrane proteins (MP) encompass a broad variety of functions, including but not limited to activation, transport, degradation, stabilization, apoptosis, cell signaling and participation in the production of other proteins. Because protein complexes represent functional units of cells, it is fundamental to understand the interactions between their components. More specifically, knowledge of the 3D structure of the MPs involved and of the interfaces involved in complex formation remains a fundamental phenomenon governing all processes of life (1). Therefore, MPs represent ultimate potential pharmacological targets in human health and disease because they include many families involved in protein-protein interaction (PPI) networks, leading to different physiological processes: current estimates suggest that about one half of all drugs target MPs. Moreover, transmembrane (TM) helix-helix interactions, through the lipid-embedded domains, lead to oligomer formation and guide the assembly and function of many cell proteins and receptors (2). In addition, the assembly of MPs may lead to emergent properties, as a relationship exists between oligomerization and function (3, 4). Thus, the role of TM domains in protein function is crucial. These TM proteins span the cell membrane in its entirety and represent approximately one-third of the proteomes of organisms (5). In eukaryotes, they come mostly in the form of α-helical bundles that cross different types of cell membranes and are approximately 3×10^5 in number, excluding polymorphisms or rare mutations. The class of α-helical TM proteins has a higher diversity than their β-barrel counterparts. And the number of possible binary and multiple interactions between them is thus vastly larger (6). Estimates of the total number of human PPIs range from 130,000 to 600,000 (7–9), to 3,000,000 (10), several orders of magnitude larger than that of the *D. melanogaster* interactome. In this paper, the plasma membrane protein interactome is understood as the complete set of direct

interactions that take place in the membrane among the proteome of α-helical integral-to-membrane transmembrane proteins (MP).

High-throughput experimental and theoretical approaches are being used to build protein-protein interaction (PPI) networks. The data covering PPIs are increasing exponentially. Indeed, in the year 2012, more than 235,000 binary interactions were reported (11). Most protein interaction databases (DB) ─ bioGRID (12), BIND (13), STRING (14), IntAct (15), MINT (11), MENTHA (16), Clinically relevant GPCR PPI mapping (17), HPRD (18, 19), HINT (20), CORUM (21, 22), APID (23, 24), HI-II-14 (25), HI-III-20 (10), BioPlex (26), QUBIC (27), IID (28), HIPPIE (29), FPClass (30), BIPS (31), LMPID (32) – offer general information about experimentally validated PPIs of all types. The IMEx Consortium and the PSICQUIC service (http://www.ebi.ac.uk) group all data dealing with non-redundant protein interactions at one interface (33). Nevertheless, the DBs are mostly concerned with water-soluble globular proteins and the juxtamembrane interactions of MPs. However, unlike globular proteins, MPs are water-insoluble, their exterior is much more hydrophobic than the interior because of allosteric interaction with the lipid environment, and they lose their native structure when removed from their natural membrane environment. Their thorough investigation has thus lagged due to these technical difficulties (34).

Proteome-wide maps of the human interactome network have been generated in the past (25, 30, 35). Traditional experimental techniques like yeast two-hybrid (Y2H) assays (36) are not well suited for identifying MP interactions. Other assays, just like Y2H, are depleted of MPs or unreliable (37). A new biochemical technique using a split-ubiquitin membrane two-hybrid (MYTH) system has been developed; however, only a limited number of membrane complexes have been hitherto determined employing it (30, 35). This procedure has been significantly extended with a mammalian-membrane two-hybrid assay (MaMTH) technique (38). However, to the best of our knowledge, MaMTH has not been used as a systematic

screening assay to map the human MP interactome. Another approach for the identification of integral membrane PPIs in yeast involves a reverse Ras recruitment system using integral membrane proteins as baits (39). An alternative method advanced a novel MYTH yeast proteomics technology, allowing the characterization of interaction partners of full-length GPCRs in a drug-dependent manner (40).

On another hand, a variety of methods for the prediction of PPIs have flourished based on: i) machine learning and classifiers based on sequence alone: PIPE2 (41), SigProd (42), PCA-EELM (43), MCDPPI (44), amino acid composition (45), and deep learning (46); ii) a statistical integration of biological data together with expert feedback, limited to the human membrane receptor interactome, HMRI (47); and iii) template structures (48–50). Once more, most of the approaches are parameterized based on soluble globular proteins only, as in (51). *Ab-initio* prediction of MP interfaces is rendered difficult, as these include amino acid compositions that are not radically different from the rest of the protein surface, even though better conserved (52). Indeed, MP interfaces are decorated with hydrophobic residues on the membrane-exposed surface.

To circumvent this problem, we developed a knowledge-based approach to predict complexes between TM proteins. The approach is based on the detection of α-helical TM non-bonded contact residues between different chains in experimental 3D structures of MP multimers reported in the PDB DB (53), and validated by the OPM DB (54). Querying the PDBsum DB (55, 56) thereafter, we obtained the atomic details of the membrane protein-membrane protein interfaces, i.e., the contacts at the intermolecular 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. The regular expressions include wildcard regions between the interface contact residues representing the residues not in contact, even though they may be exposed to the membrane lipids. With this

information in hand, we proceeded to search in the UniProtKB DB of protein sequences the obtained motifs in other MPs. To extend our search, we allowed certain degrees of mutation in the membrane-embedded interface contact residues. Our assumption is that homologs of precisely the template interface motifs are expected to interact analogously (57), as opposed to a global approach that would limit the search to functionally related partners or to overall sequence homolog 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 (58). Thus, we do not focus on the overall sequence homology, such as in other template-based predictions (49, 59). The linear 1D motifs we obtain represent thus 3D epitopes implicitly and have the advantage of encompassing experimental 3D information. These linear motifs encode 3D binding sites and are useful for quick sequence searches. In this work, we focus on the plasma membrane of eukaryotes, ensuring thus that we are probing proteins within the same species and subcellular localization.

MATERIALS AND METHODS

Algorithm

Data mining and filtering

Figure 1 summarizes the steps followed to collecting, filtering, and processing the input data and to generating the output data. We started by obtaining a list of all eukaryote "reviewed" proteins from UniProtKB DB, a manually annotated and reviewed DB of proteins with experimental evidence of their existence (60). We mined proteins with cellular component annotations that matched the following GO annotations (61): "integral component of membrane" (GO:0016021) <u>and</u> "plasma membrane" (GO:0005886); <u>or</u> "integral component of plasma membrane" (GO:0005887). Thus, we took only proteins characterized as membrane

proteins. From the resulting list, we identified the subset of proteins that have an experimental 3D structure in the PDB DB, a resource containing the 3D shapes of biological macromolecules in the form of a complex spanning the TM region. We retained those with at least six buried interacting residues in each monomer (eight buried residues is typical of biological interfaces, (62)). This amount is the minimum for presenting an accessible surface area leading to quaternary structure. To have a high quality and homogeneous dataset, we adopted stringent selection conditions. Thus, for the PDB structure to be considered valid, it had to have a resolution of 3.5 Å or better if it was obtained by X-ray crystallography or cryoEM. We took in all different conformational states, regardless of the presence or absence of ligand(s), pH, symmetry group, apo or holo form, and allosteric effects, unless the differences modified the set of interface residues. We eliminated PDB MPs presenting engineered mutations, insertions, and deletions in the TM segment with respect to the wild-type or natural variant sequence in UniProtKB, just like chimeras in which the xenophobic part is TM. We also ignored redundant MPs, those whose 3D structures show no significant TM segments, and those pair interactions that are redundant due to symmetry. Manual curation excluded non-parallel configurations of the protomers, head-to-head or head-to-tail orientations (resulting most probably from crystal interfaces), out-of-membrane interactions only, and TM segments that do not interact, as dictated by the cell membrane. The intramolecular interactions within each protomer of an oligomer, as well as with the lipids and detergents environment are included implicitly. The PDBSum DB shows the PDB molecules that make up a complex and the interactions between them. This information delivers what interface residues of protomer A interacts with what interface residues of protomer or chain B, and what protomer A interacts precisely with what protomer B and not another.

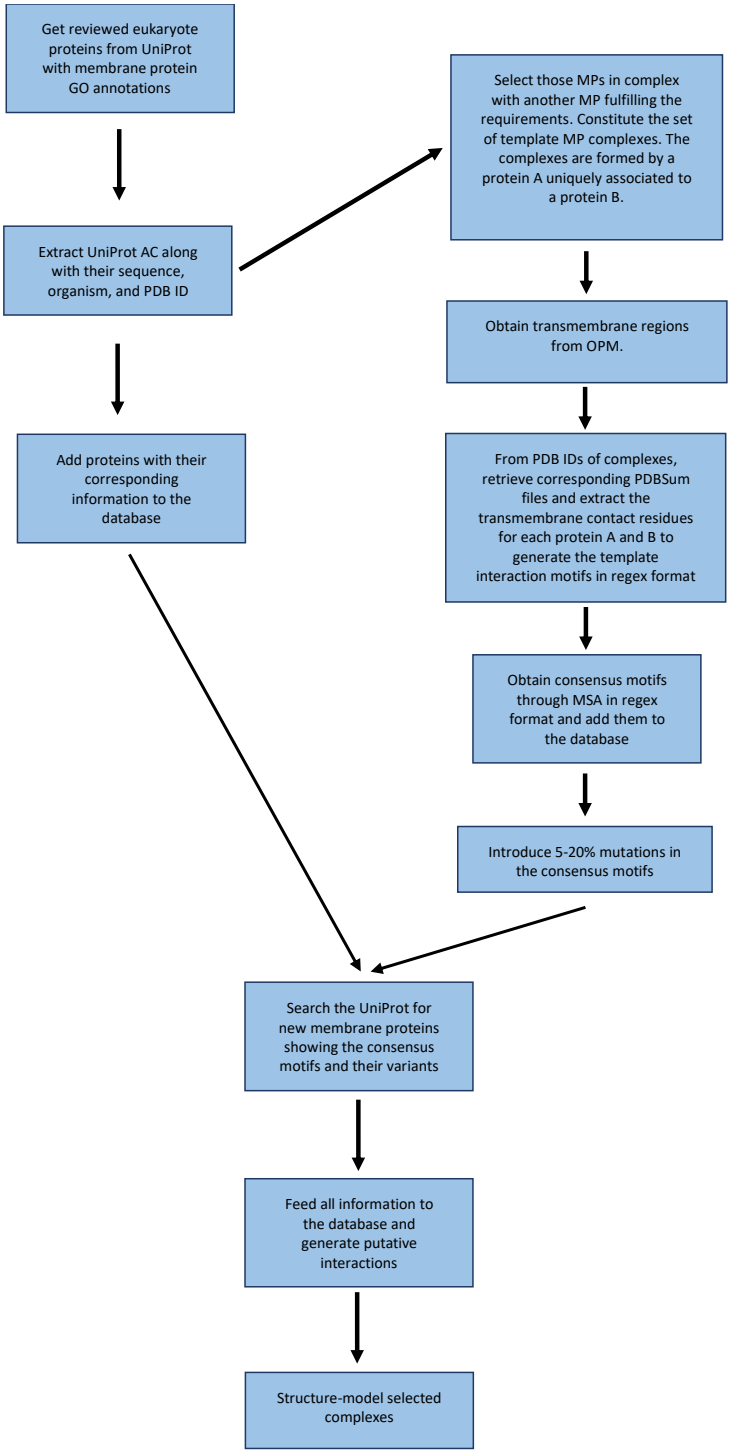


Figure 1 Workflow illustrating the TransINT algorithm, from information retrieval to detection of recognition motifs to generation of putative interactions, to 3D modeling of complexes.

Finally, to ensure that the oligomer structures we took into account are quaternary structures with biological sense, we used EPPIC, a protein-protein interface classifier (62–64),

and PRODIGY, a classifier of biological interfaces in protein complexes (65, 66) to distinguish between crystallographic and biological assemblies.

Motif extraction

To choose the PDB structures to work on, we referred to the OPM DB (67) that provides the orientation of known spatial arrangements of unique structures of representative MPs coming from the PDB with respect to the hydrocarbon core of the lipid bilayer. We chose all the PDB structures that mapped to the UniProtKB extracted MPs and pulled out all the available PDBsum files of these structures. We double-checked the chosen PDB structures with the MPStruc DB of MPs of known 3D structure (https://blanco.biomol.uci.edu/mpstruc/). PDBsum contains atomic non-bonded contacts between amino acid residues at the interface of molecules in a multimer complex. We used the information in PDBsum to extract the intermolecular contact residues. We filled in with the non-contact residues in the sequence, that is, those between the contact residues as wildcards. Thereafter, we defined the binding motifs by obtaining the corresponding linear sequences (motif instance). From the PDBsum file listing the contacts, we formulated two motifs, one corresponding to partner protein A and the other one to its partner protein B. Because we are only interested in the recognition site at the TM interface region, we ensured that each interacting residue belonged to the TM part of the sequence. We represented our motifs using the Regex format (https://www.regular-expressions.info/) and denoted the TM contact residues by their one-letter symbol, the residues in between (representing the wildcard regions) by a dot, and the curly braces for the consecutive occurrences of the wildcard. As an example, we take UniProtKB membrane protein P04626 (receptor tyrosine-protein kinase ErbB2). The spatial structure of the ErbB2 dimeric transmembrane domain in the PDB is 2N2A. Its PDBSum entry contains non-bonded contacts across the surface for each of the chains A and B. The residues in between constitute the

noncontact residues and are represented by curly braces. PDBSum lists chain A residues I12, V16, L20, V23, L24, V27, and L31 as establishing contacts with chain B. The result of the contact motif for chain A is $^{12}IX_3VX_3LX_2VLX_2VX_3L^{31}$ (Figure 2), equivalent in the regular expression notation to I.{3}V.{3}L.{2}VL.{2}V.{3}L. As this is a homodimer, the contact motif is the same for chain B. The wild cards {} may represent extramembrane loops or inner protein subdomains of different lengths that do not establish interface contacts. We do not focus on these variable segments; consequently, we do not look in our data mining at whether those segments are conserved or not. We focus only on the contact residues. If the resulting motif is a binding motif, the types of the variable residues should not be essential.

Lastly, to verify the consistency of our algorithm and the validity of the obtained binding motifs (that they will not match with many sequences just by random), we sent several BLAST (https://blast.ncbi.nlm.nih.gov) searches against the PDB of several motifs or their parts thereof (in case they were separated by long stretches of non-contact amino acid residues between the contact residues). We always recovered the corresponding template PDB files, as well as related PDB codes, among the maxima BLAST scores and with query covers of 100 (not shown).

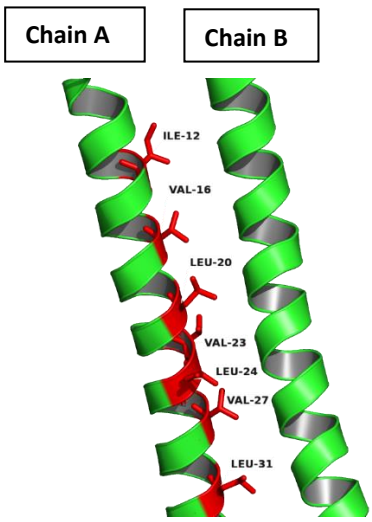


Figure 2 Contact interface residues of chain A of the ErbB2 dimeric transmembrane domain of sequence [12]**Ile**-Ser-Ala-Val-**Val**-Gly-Ile-Leu-**Leu**-Val-Val-**Val**-**Leu**-Gly-Val-**Val**-Phe-Gly-Ile-**Leu**[31]. Contact residues are in bold.

Searching for identified motifs in other protein sequences

To eliminate redundancy in our data, we grouped motifs derived from application of the mutation rates and built a consensus motif for each cluster using multiple sequence alignments (MSA). This allowed us to retain conserved interface residues, avoiding potentially deleterious effects on protein assembly. We generated new potential binding sites by applying mutation rates to the contact residues, ranging from 0% (exact match) to 20%, with increments of 5%. In this way, we wished to amplify our search, allowing us to find other protein sequences. The mutation procedure is applied independently to each protomer, on each motif. Most of the time, the two motifs coming from a homodimer are identical. However, in some cases, one of the monomers of a homodimer may reveal another interacting surface. In this case, the corresponding motif will be different. Subsequently, we queried our consensus motifs against the original UniProtKB dataset. We defined a COST parameter as the number of amino acid mutations allowed anywhere in the motif, depending on the number of contact residues it contains and the mutation rate for the run; we give a score of 0 to both insertions and deletions to ensure that no contact residue is lost. For instance, when generating new sites from a valid motif with eight contact residues, COST is two ($8 \times 0.20 = 1.6$ rounded up to 2) for a mutation rate of 20%. The values of COST_A and COST_B vary from 0 to 4, implying that there are motifs with up to 20 contact residues. In which case, the contact residues may be shared by several sub-motifs in TM α-helices.

To keep track of which A motif interacts with which motif B, we kept the motifs in separate pools. In other words, the rationale for keeping the two sets A and B of motifs distinct

is that a given motif A matches a given motif B and not another. Subsequently, we paired the predicted motifs from novel interactions based on the PDBsum validated interactions. In this way, we were sure that pattern A from the new protein A "binds" to its complementary pattern B of new protein B. Because motifs can be found fortuitously anywhere in the sequence, we considered only those motifs belonging to the TM region. We also checked which MPs had the motifs included in their PDB structures, if available.

Implementation

The TransINT database

Thereafter, we built a heterogeneous DB named TransINT, which contains all the found interactions. The DB is the result of the implementation of the fully automated template-based recognition site search pipeline. We used a MySQL DB to maintain all the information collected in a structured manner. To access the DB and search for our data, we built a web interface using PHP and HTML5, which allows the user to query the needed information. Users can query the DB for obtaining motifs by entering a UniProtKB AC, a gene name, a type of organism, a mutation percentage, or a motif of interest or part thereof using the Regex format. A link to the UniProtKB DB site for each UniProtKB AC is available. The user can choose more than one filter option when querying and will exclusively obtain interactions thought to occur in TM regions between plasma membrane proteins of the same species. The user can also adjust the values of the following parameters: Number of contact residues, Mutation rate, COST and Valid as an interval or as a fixed value. Homodimers of the TMs dealt with do not appear in the database, as all TransINT TM proteins are expected to form them. We update our DB following each release of the UniProtKB and OPM datasets and then regenerate all statistics.

Molecular docking

The predicted interface residues can be used as restraints to reconstruct the structure of dimers through docking. We consider successful docking as a partial validation procedure. To illustrate our approach, we generated 3D complexes derived from selected predicted pairs of MPs with the goal of looking at their internal architecture. To do so, we searched for a protein-protein docking program that would allow us to perform a steered docking simulation using the epitopes extracted from the molecular interface of the complex. We processed and analyzed a considerable number of experimental 3D MP structure files using several docking programs. As we know the docking interfaces, we did not perform *ab initio* molecular calculations. Neither were we concerned whether the docking program was trained on sets composed primarily of soluble proteins.

Even though several tested docking programs were sufficiently precise, we decided to use GRAMM-X (68, 69) for creating the novel protein-protein 3D complexes. GRAMM-X has an option in which the user is allowed to submit those residues that might form the interface between the "receptor" and the "ligand." The program also has options for listing the residues required to be in contact. To verify the performance of GRAMM-X for MPs, we benchmarked it against several MP complexes in the PDB. GRAMM-X was indeed able to reproduce many of the experimental MP complexes. For molecular docking simulation and identification of the TransINT-predicted complexes, we chose examples in which the 3D PDB structures of proteins were already available or represented very highly homologous templates. After docking, we manually curated the output by filtering out nonparallel, perpendicular, or oblique protomer pairs, regardless of the calculated energy. We also considered the topology of the MP in the membrane. We visualized the obtained 3D structures of the complexes using PyMol 2.4 (www.pymol.org).

RESULTS

Automated pipeline to predict and explore thousands of novel transmembrane protein-protein direct interactions (MPPI)

UniProtKB provided us with 13,352 MPs that include the GO annotations mentioned in the S&M section. Overall, these proteins mapped to 954 distinct oligomer MP PDB structures. As we focused on structures that satisfy the requirements signified in the S&M section, we were able to validate only ~50 PDB files of MP-MP complexes. After checking which corresponding PDBsum files to consider, we ended up with 53 non-redundant template interactions, associated with 48 unique reviewed UniProtKB entries across species. Fifty of these interactions are structural homomers and three are structural heteromers (Table S1, yellow highlighting). The set includes X-ray, solution NMR, and electron microscopy structures. The MPs include families like receptor tyrosine kinases, TLRs, ion channels, Cys-Loop and immune receptors, gap junctions, transporters, and GPCRs. In these, besides *H. sapiens*, a taxonomically diverse set of organisms across the eukaryotic kingdom of the tree of life is represented: *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 one mixed bitopic-polytopic protein complex in our template set. The oligomerization order is adequately 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 occurrences of the number of TM helices for each of the protomers of the reference MP complexes shows that bitopic single-span helix complexes are the most preponderant; these are followed by polytopic MPs composed of 6, 4, 7, 9, 2, 8, and 10 and 12 helices. The following EC numbers are represented in this set: EC 2 (10 transferases), EC 3 (one hydrolase), and EC 7 (one translocase). When verifying the

protein-protein interfaces in the complexes for the type of assembly they form (crystallographic or biological), we found that all the X-ray or electron microscopy complexes were classified as biological (Table S3). We did not submit the NMR-determined complexes to the test. Therefore, this is the number of experimentally solved structures of MP protein complexes that we used as the template set.

The TransINT pipeline results indicate that there are 1504 unique MPs involved in the predicted complexes, going from UniProtKB accession codes A0PJK1 to Q9Y6W8. Of these 417 are human (Table S4). 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).

Table 1 Non-redundant transmembrane α-helix TransINT recognition motifs

A.{2}RFE.{126}I.{2}EGL.AI.{2}V.{2}F.{115}L
A.{3}V.{3}I.{3}L.{6}S.{2}R.{19}L.{3}F.{2}L
AI.{2}AI.{2}T.{3}I.{5}I.{97}W.{6}F.{2}V.{2}LS.{2}L
AV.{2}G.{2}FV.{2}GP.{2}AL.{2}I.{112}NI.{19}FL.{3}F
AV.{2}GL.{2}GA.{2}LL
AY.{82}E.VA.{2}VF.LA.{2}WT.{124}L.{3}I.{2}M
C.{2}LD.{2}L.{2}YG.{2}LT.LF
D.{2}F.{2}FF.{2}VW.VA.{2}VA.EGL.{67}L.V.LLV.{2}LL.{2}NIL.{2}NL.IAM
DWN.{2}IA.{2}V.{2}LI.{24}F.{3}F.{2}A.{2}YLV.{2}FM
E.{2}TL.VGF.IL.{2}SL.{2}TY
EI.IL.{9}FW.LL.{2}I
F.{2}C.{2}VI.{2}GY.FC.WIV.{2}P.{39}LVW.F.Q.{3}Y.{2}I.{3}IYM.{2}S
F.{3}G.{2}F.{8}YITV.L.{2}WN.{2}VV.{2}I.{9}L
F.{3}N.{2}IL.{2}FI.{2}V.{3}H
F.{3}N.{3}L.{2}FI.{6}HF

F.{3}VL.{2}HF
FG.FG.{2}VM
G.{3}A.{2}WI.{2}P.{18}LW.{2}MM.SA.{2}I.{3}F.{45}V.{3}M.{18}LF.{2}M
G.M.{3}I.{3}I.{3}L.{2}F
GE.{2}AV.{2}GL.{2}GA.{2}L.{3}L
GYY.IQ.{6}L.{2}IL.{2}V.{13}A.{3}T.{3}T.{3}Q
I.{2}A.{3}T.{2}SI.{5}I.{97}W.{6}F.{2}V.{2}LS.{2}L
I.{2}LM.{2}G.{3}A.{2}IV.{2}L
I.{2}LV.IF.{2}L.{3}F
I.{3}L.{2}T.{3}S.{2}R.{19}L.{3}F.{2}L.{3}A
I.{3}L.{2}T.{6}R.{12}D.{2}MA.{6}F.{2}LL
I.{3}V.{3}L.{2}VL.{2}V.{3}L
I.{3}V.{3}L.{2}VL.{3}F.{2}L
IG.{2}A.{3}V.{2}L.{7}F
IQ.{6}LI.IL.{5}WI.{6}A.{2}AL.{2}T.{3}T.{64}R
IQ.{6}M.{2}IL.{2}V.{3}L.{6}A.{3}F.{2}T.{3}T.{130}R
IR.{3}L.{2}IF.{2}T.{3}L
K.{3}GF.{2}M.{2}II.{2}AY.QL.YL.{38}L.P.{2}F.{3}VF.{3}FI.{2}NM.LAI.N
K.{3}W.{2}CF.{5}L.{379}Q
K.{6}L.{6}FF.{2}GS.{11}LQ.{2}LA.{2}LA.{2}T.{2}QAL.PV.{40}I.{58}L
KD.{2}F.{2}FF.{2}VW.VA.{2}VA.EGI.{67}L.V.L.I.{2}LL.{2}NIL.{2}NL.I
KF.FI.{2}L.{3}AF.NG.NQL.F.{52}F.{2}FV.A.MFG.{2}NV.{2}L.{4}NM
KF.VL.{5}FAF.{2}G.FIL.S.{41}FI.{2}I.Y.{2}YG.{2}N.{7}LN
KW.{2}V.{3}F.{2}V.{2}YLI.GA.VF.{87}EGGKIFC.{2}YALLG.PL.{2}FL.{2}G.{2}DQ.{94}W.{13}L.{2}IG.{3}R
L.{2}I.{3}D.{2}LT.{2}I.{3}V.{16}I.{2}VA.{3}K.{2}V.{3}F
L.{2}I.{3}D.{3}T.{2}I.{3}V.{2}L
L.{2}L.{3}V.{2}YL.{2}GA.VF.{65}L.{2}AF.FSG.{2}I.{12}D.GR.FCI.Y.LVGIPL.{2}IL.{2}GV.{2}R.{58}L.{3}Y.{2}I.{3}T.{45} L

L.{2}L.{6}LY.{23}R.{10}L
L.{2}LF.{2}LI
L.{2}LL.LVL.YL.{2}GA.VFR.L.{62}L.{2}AF.FSG.{2}I.{14}GR.FCI.Y.{2}VGI.L.{2}IL.{2}G.{3}R.{58}L.{3}Y.{2}I.{3}T.{24}W.{9}F.{3}L.{2}IG.{2}LR
L.{2}VI.{2}IY.{14}G.{2}F.{115}F.{2}IF.GLLV
L.{3}V.{3}T.{2}W.{81}V.{2}MF
L.{4}F.{2}YF.{2}GL.{2}I.{3}M.{3}L.{2}Y
L.{45}T.{3}F.SA.{2}SV.{18}M.{2}SAT.{140}V
L.{86}D.{2}RT.{2}CL.FM.{2}TL.{2}L.{2}F.{19}V.{2}F.{105}LI
L.{86}Q.{3}V.{2}IL.{2}Q.{3}C.{2}AS.{14}S.{2}LS.TL.{2}LV.{44}L.{3}L
L.YGV.FF.{2}I.{6}TL
LA.FA.LFL
LA.FA.LFL.{5}V
LA.IV.{2}D.{2}LT.{2}I.{3}V
LA.IV.GD.{2}LT.{2}I.{3}V
LL.{3}VL.SI.{2}AF.{2}G
LL.{5}LY.{22}RR
LL.AS.{90}LV.EA.FAI.NI.{2}S.{2}L.{3}F.{19}I.{100}I
LQ.{6}LI.IL.{2}V.{2}WI.{6}A.{2}AL.{2}T.{10}H.{160}R
LSL.{43}F.{2}YL.AS.{2}HV.{2}V.{15}L.{2}L
M.{3}G.{3}L.{2}C.{3}LV.HA.{2}FF.{8}F
M.{3}L.{2}LL.{2}AL.{2}GL
M.{6}AY.QL.{2}LL.G.{32}NR.{3}P.{2}FV.{2}V
N.{2}A.{2}VF.{2}W.{3}F.{2}I.{27}F.{67}L.{3}L.{2}IN
P.{2}L.{3}V.{2}DL.{6}A.{3}Y
PL.{2}AG.{3}G.{2}IL.{2}V
R.{2}FA.{2}L.{3}L.{2}FF.{2}GS.{11}LQ.{2}MA.{2}LG.{2}T.VQALG.{42}LL.{57}L
R.{2}FV.{2}VF.{2}G.{2}TA.VT.{58}K.V.I.L.{5}I.{2}YIL.{2}NM.IAL

R.{3}V.{2}VF.FG.{2}TA.VTL.{59}V.{2}IL.L.{2}VI.{2}YILL.NM.{2}AL
R.{74}LW.{2}M.{2}MI.{2}A.{2}YAM.VGH.T.{3}Q
S.{2}LI.{2}FI.{2}L
S.{2}VL.{2}IV.{2}D.{2}L.{2}LI.{2}AV
S.{3}L.{87}L.RT.{2}CI.FM.{2}T.{2}LL.{2}F.{19}V.{2}FL.{104}LI.{2}FS
S.{3}V.{3}I.{3}L.{6}T.{19}M.{2}F.{3}F.{2}LL.{5}N
T.{2}AY.{85}A.{2}VF.L.{3}WT.{2}L.{18}MI.{3}L.{97}L.{2}LI.{2}M
T.{2}IS.{2}VG
T.{2}LI.{302}D.{2}SI.{2}GT.{2}L.{2}WV.VI.{2}L
T.{2}LL.{2}IF.{2}LL.{2}F
T.{2}VL.IP.{2}F
T.LV.{2}I.{2}V
V.{3}T.{3}V.{3}L.{6}I.{22}Y.{3}F.{2}L.{3}A
V.{3}Y.{24}LL.{6}VL.VA.{2}A.{24}GF.{2}A.{11}F
VI.{2}V.{3}F.WLV.LL.{2}SV.WF.{17}L.{6}SV.{2}Q.{44}IS.VFS.IN.{22}TS.{2}LT.{32}H.{3}S.{2}TF.N
VL.{86}Q.{2}TV.{2}FL.LQ.{2}LC.{2}AS.{14}S.{2}FS.AL.{2}LL.{43}LY
VT.{2}V.{3}V.{2}WF.{2}MF.{17}M.{2}NL.{2}A.{6}M.{12}T.{2}L.{23}V
VV.{2}A.{2}A.{2}VL.{2}I.{3}I
VV.{2}A.{2}AL.VL.{2}IS.{2}I
W.{2}V.{2}IF.{2}V.{2}YLI.GA.VFK.{87}GGK.FC.{2}YAL.GIPL.{2}FL.{2}G.{2}DQ.{108}L.{2}IG.{2}L
W.{2}V.{2}IF.{2}VV.YLI.GATVF.{88}GGKIFCI.YALLGIPL.{2}FL.AG.{2}DQ.{94}W.{3}L.{5}F.{3}L.{2}IG.WLR
W.{2}V.{2}IF.VV.{2}YL.{2}GATVF.{88}GGKIF.I.YALLGIPL.{2}FL.{2}G.{2}DQ.{94}W.{9}F.{3}L.{2}IG.{2}LR
W.{2}V.{3}F.{2}M.{3}V.{3}EV.{143}E.{2}VF.{2}F
W.{3}II.{2}I.{3}T.{3}F
W.{3}L.{2}W.{3}FQ
Y.{2}FV.{3}T.{2}IQ
Y.{2}H.{19}F.LL.{6}VL.VA.{2}A.{24}GF.{2}AP.{6}V.{2}YF.{2}R

Y.{2}I.{7}L.{2}IL.{16}GL.{2}T.{2}LT.{2}T.{2}SG.R
Y.{2}T.{317}N.{2}A.{2}LF.{2}W.{3}F

TransINT binding motifs containing the G.{3}G packing sub-motif are in bold orange; A.{2}T.{2}QA or A.{2}VF.LA in bold green; Aromatic.{2}Aromatic in bold blue; and L.{6}L heptad in bold gold.

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 abundant, followed by Ala and Gly. This residue distribution has been found previously (70). Leu was found more than 300 times, making about one fifth of all contact residues, like reported before (71) (Figure S1). As expected, the physicochemical properties of MPPI-binding sites are different from the exposed sites of soluble proteins. The amino acid residue abundances for helix interactions we found in the motifs match those in the literature (72, 73). Figure S2 shows, as a "heat map," couples of contact residues at the interface for our template set, as they come out from PDBsum. These residues are in the surface of the TM protein 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 pair. The least observed interactions include His-His for a homotypic pair and Trp-Cys for a heterotypic pair. This outcome suggests that residues tend to contact other residues sharing the same physicochemical properties, and agrees with the statistics obtained for inter-residue interactions in the MP Bundles DB for α-helical MPs (72). The statistical trends in the contacts imply specificity at the interface of the TM helices, as well as correlated mutations of coupled contact residues and paired motifs.

The number of TM atomic non-bonded contacts per complex covers a wide range: 11-87 for single-span TM proteins, and 8-186 for multiple-span TM complexes (Figure 3). As expected, the latter show in average a larger number of non-bonded contacts.

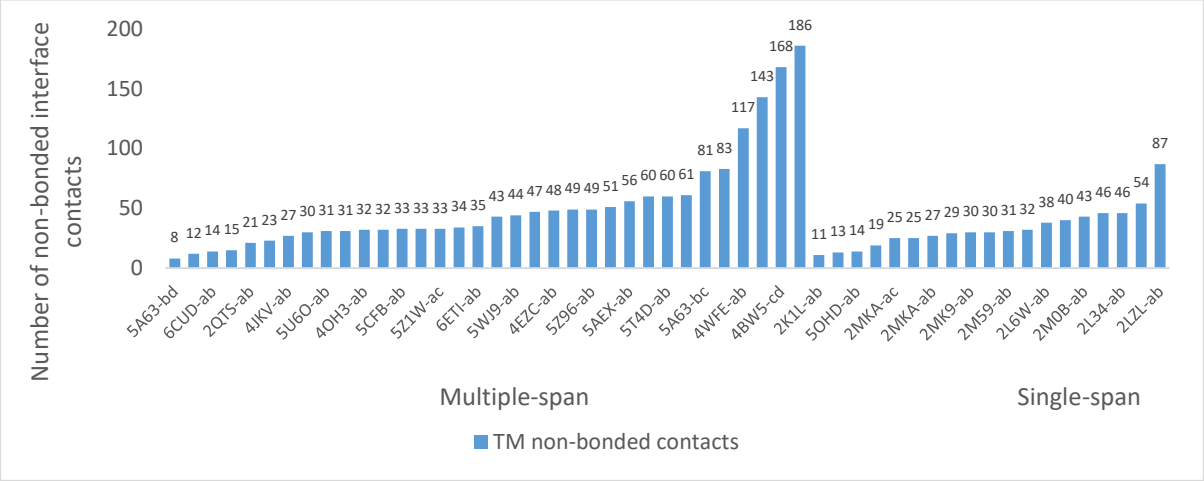


Figure 3 Number of PDBsum non-bonded interface contacts for each of the experimental template complex PDB structures of Table S1.

Finally, we looked at the Pfam-A set of protein families (74) to which the 63 different proteins in the template complexes of Table S1 belong. We found that the most populated families were PF07714 (Pfam family PK_Tyr_Ser-Thr) and PF00520 (Pfam family Ion_trans), with occurrences of 10/110 and 6/110, respectively. Forty-two proteins had only one Pfam occurrence each (not shown).

α-helix packing motifs

Walters and DeGrado (75) have examined helix-packing motifs in membrane proteins involved in the folding of a helical membrane protein. As we are dealing with inter-helical interactions between different proteins embedded in the membrane, helix-pairing motifs are not directly comparable. However, we can look at the pairing motifs of TM proteins composed

of a single TM α-helical domain, such as glycophorin A, presenting a G.{3}G packing sub-motif (76–78), in which Gly or other small residue space four residues that mediate parallel inter-helix protein interactions. Indeed, we find this small motif and its variants as part of the extracted binding motifs from our set of template MP complexes (Table S1). The structures include parallel single α-helix homodimers (Table S1, PDB 2L2T and 2LOH), multiple α-helix homo-oligomers with a different motif in each component (Table S1; PDB 5AEX, chain A; PDB 5CTG, chain A). The G.{3}G sub-motif appears as well in the non-redundant TransINT motifs (Table 1, in bold orange).

An antiparallel coiled-coil sub-motif of α-helices with Ala or another small residue in every seventh position has been termed Alacoil (79). It can be found in the binding motif of human aquaporin 5 (PDB 3D9S) as A.{2}T.{2}QA, and as A.{2}VF.LA in the TRPV1 ion channel in complex with double-knot toxin and resiniferatoxin (PDB 5IRX) of our template protein set (Table S1; Table 1 in bold green).

Another specific pattern is the Aromatic.{2}Aromatic motif (80), like in F.{2}F of connexin-26 (PDB 2ZW3), of the mouse TRPC4 ion channel (PDB 5Z96), and of TRPM4 (PDB 6BCO, 6BQV). In our template set, this sub-motif appears as part of the TrkA transmembrane domain (PDB 2N90). In the SWEET transporter (PDB 5CTG), it takes the form F.{2}Y; in polycystic kidney disease protein 2 (PKD2) (PDB 5T4D) and polycystic kidney disease-like channel PKD2L1 (PDB 5Z1W), it becomes F.{2}W (Table S1; Table 1 bold blue).

Another sub-motif that appears very frequently is the L.{6}L heptad, analogous to the Leu zipper that mediates protein complex formation in water-soluble proteins. As this pattern is important, it is shown in Table S2 and Table 1 (bold gold).

TransINT α-helix packing motifs

TransINT reveals original packing motifs through its non-redundant consensus motifs, such as the frequent motif A.{2}(A/L/V)(L/F) (Table 1). The following motifs with Ile are also well populated: for position i-1, (F/L/V/W)I; for positions i-2 and i-1, (F/L)CI; for positions i+3 and following, I.{2}(A/L/V), I.{3}(L/V), I.{5}I, and I.{6}A. TransINT also uncovers Ile composite motifs, like I.{2}(A/F)I.{2}(T/L) and I.{3}I.{3}L. The most frequent motifs are those beginning with Leu: L.{2}(A/F/G/I/L/N/T/V)(A/G/F/I/L/V), and L.{3}(F/V). Leu composite motifs include L.{2}(F/I)L.{2}G, L.{2}IF.{2}LL.{2}F, L.{3}L.{2}FF, and L.{3}L.{2}INP.{2}L.{3}V. The occurrence of these patterns is reflected in the inter-helical contacts between proteins of the Ile-Ile, Leu-Leu and Ile-Leu residue pairs (Figure S2). The composition of these residues shows that hydrophobic residues are enriched at the interface between MPs.

A statistical analysis of amino acid patterns in TM helices (78, 81) results in 30 over-represented ($p <<<$, odds ratio >1), and 30 under-represented ($p <<<$, odds ratio <1) pairs. Thus, G.{3}G (GG4 in Senes et al. notation) is the most significant pair among over-represented pairs. Indeed, this and other significant over-represented pairs (21 out of 30), such as I.{3}I, G.{3}A, IG, I.G, V.G, I.{3}V, IP, V.{3}V, V.{3}I, AV, and G.{2}L, have their equivalents in non-redundant TransINT motifs (Table 1). On the other hand, several under-represented pairs (10 out of 30) are in fact absent of TransINT motifs in Table 1, such as I.I, F.{3}I, and I.{3}G. Moreover, several most significant over-represented triplets of Senes et al. appear in our motifs. This lends support to our template-based approach.

We then wished to look at the number of UniProtKB ACs resulting for each number n of contact residues for the biological species recorded in the PDB and UniProtKB. As seen in Figure 4, the count occurrence of contact residues in the motifs is largest for n = 7 for both protein A and protein B of the interacting pair. The quasi-periodicity of the interface

recognition residues reflects a spatial arrangement corresponding to a heptad α-helical pattern. This is true also for motifs separated by values of the linker residues ≥4. For example, the long motif DWN.{2}IA.{2}V.{2}LI.{24}F.{3}F.{2}A.{2}YLV.{2}FM (Table 1) contains two motifs of 14 and 17 residues separated by 24 linker residues. A motif may attain a total of 20 residues, like in E.{2}TL.VGF.IL.{2}SL.{2}TY (Table 1), possessing enough length to span a bilayer. Our motifs therefore may be rather long, implying residue correlations beyond the pair level. Moreover, an additional information is given by the TransINT motifs: the residues that compose them are not only on the same face of the helix but are interfacial residues. The corresponding statistics for *H. sapiens* show similar trends: small motifs with six to seven buried contact residues are the most abundant (Peak at n = 7, combined number of redundant proteins A and B, 9655). The range of contact residues in the motifs was 6-32. As the number of contact residues increased in the motifs, the number of hits decreased drastically, with only one to four predictions for both proteins in the 18-33 contact residue range (Figure 4). These findings suggest a limited set of binding motifs in nature.

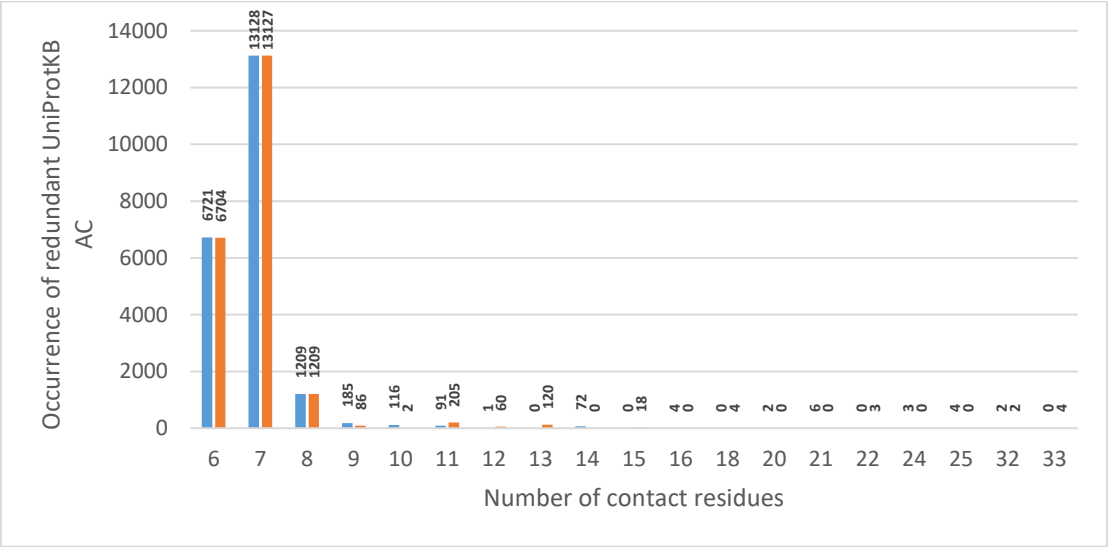


Figure 4 Occurrence of (redundant) UniProtKB ACs as a function of interface contact residues for Proteins A (blue) and corresponding Proteins B (orange). Mutation rate 0-20%. Organisms extracted from the PDB and UniProtKB.

The protein-protein docking benchmark 5.0 PPDB (82) assembles non-redundant high-resolution complex structures, 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 MPs. On our side, from our 53 non-redundant template structures of MP-MP complexes, we could extract a subset of them along with the unbound structures of their components to define a benchmark. This benchmark is made up of ten sets of structures (Table 2).

Table 2 Benchmark of experimental structures of membrane protein-membrane protein complexes and their unbound components.

Complex	**Unbound**	**Unbound**	**Type**
PDB code	PDB code protein A	PDB code protein B	Homo- (Hm) or Hetero- (Ht) mer
2LOH	2LLM	2LLM	Hm
2QTS	3IJ4 4FZ1 4NTW 4NTX 4NTY 4NYK	3IJ4 4FZ1 4NTW 4NTX 4NTY 4NYK	Hm
2ZW3	5ER7 5ERA	5ER7 5ERA	Hm

3SYQ	3SYA 3SYC 3SYO 3SYP 4KFM	3SYA 3SYC 3SYO 3SYP 4KFM	Hm
4JKV	4N4W 4O9R 4QIM 4QIN	4N4W 4O9R 4QIM 4QIN	Hm
4NEF	4OJ2	4OJ2	Hm
4WO1	2L34 4WOL	2L34 4WOL	Hm
4X5T	2M6B	2M6B	Hm
5A63	2N7Q 2N7R	2N7Q (for Q92542) 2N7R (for Q92542)	Ht
5O9H	6C1Q 6C1R	6C1Q 6C1R	Hm

The first column shows MP complexes of known structure. The 2nd and 3rd columns list the experimental structures of the isolated chains that compose them. The 4th column gives the type of oligomer (homo- or heteromer). PDB structure 5A63 represents the sole heteromer of the benchmark. All the other complexes are homomers. Note: This table is derived from Table S3, Column C.

When comparing our benchmark to the “Dimeric complexes” table of the Membrane Protein Complex Docking Benchmark, MemCplxDB (83), we only recover the PDB 5A63

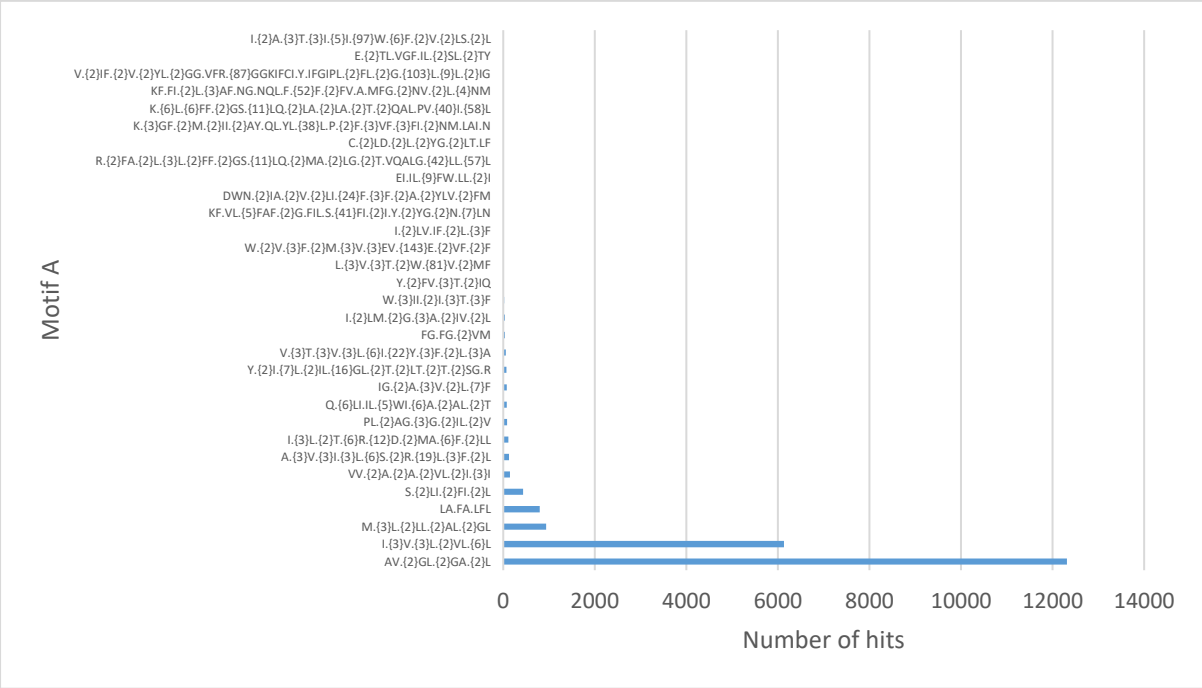
complex. The reason is that MemCplxDB shows many interactions between MP and non-MP, soluble proteins (antibodies, peripheral proteins, etc.) which we do not deal with. MemCplxDB includes interactions as well between oligomers within a multimer complex, and prokaryotic MPs of β-barrel structure. Our benchmark represents thus a gold-standard set of positives for integral-to-membrane proteins interacting through their α-helical TM segments. As the 3D structures of more TM complexes appear, the benchmark could serve for a machine learning approach of membrane PPIs.

Predicted interactions

The number of TransINT-predicted membrane heteromer interactions for the 39 species dealt with is 21,544 among 1,504 MPs. The homodimers are thus 1,504 in number and represent 6.5% of all complexes. Of the total heteromer interactions, 9,797 among 417 MPs correspond to *H. sapiens*, the homodimers representing 4.1%. We identified many potential recognition sites and novel binary complexes at different mutation rates. As mentioned in the S&M section, we built a consensus amino acid motif for all the matched sequences of a given binding site in the 0-20 % mutation range considering the contact residues only. This led us to detect conserved amino acid residues among the contact residues of the binding motifs. The most prevalent consensus motif A found was AV.{2}GL.{2}GA.{2}L, illustrated by the instance sequence ^{423}AVFSGLICVAMYL^{436} of the sodium-dependent proline transporter (UniProtKB Q99884) at a 15% mutation allowance (Figure 5a). For consensus motif B, the same motif was the most frequent (Figure 5b). The least frequent motifs are characterized by hefty linker segments between the contact residues and thus by substantial sequence lengths. 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 (UniProtKB Q9UBN4).

Incidentally, as contact residues, the amino acids composing our binding motifs do not necessarily represent hot-spots, i.e., those that contribute to the binding free energy (84). Although their degeneracy in hydrophobic amino acid type suggests that the TM interface residues contribute equally to the free energy of binding and are part of conserved sequences of TM regions.

a)



b)

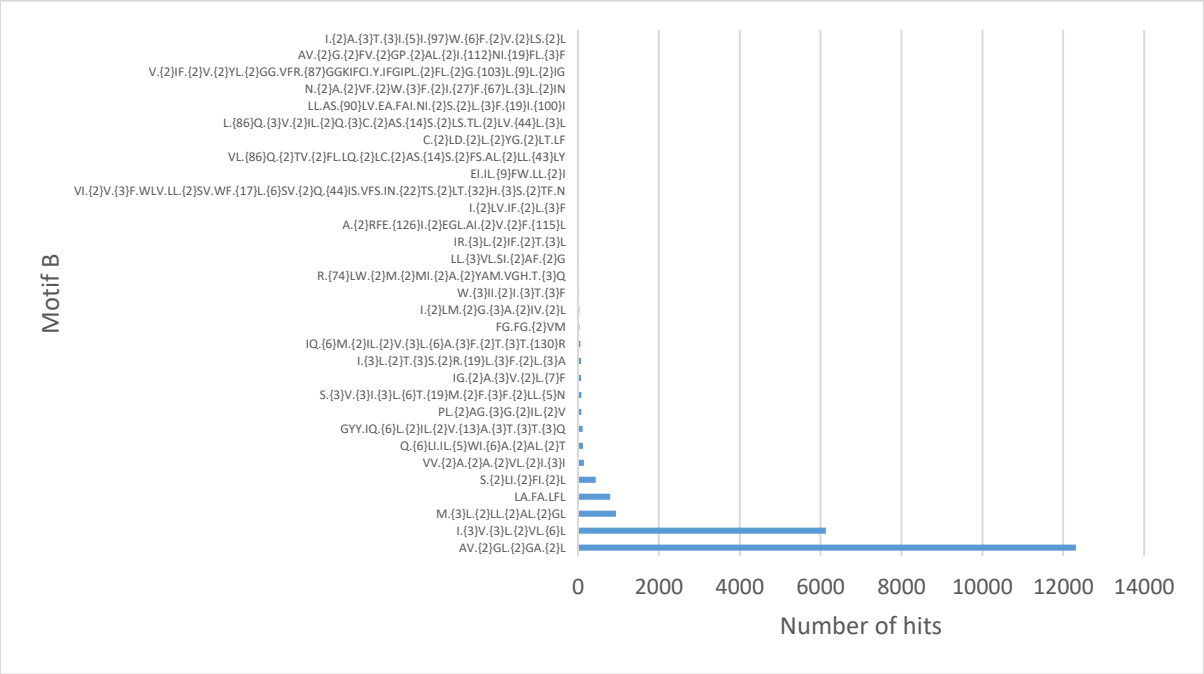


Figure 5 Number of hits per motif for all species. a) motif A, b) motif B. Mutation rate 0-20%.

Mutation rates of 0 % for motif A and motif B result in those proteins whose contact residue sequences conform exactly to the consensus motifs. The TransINT outcome gives 79 entries of hetero-oligomer complexes across species, of which 25 are *H. sapiens'*. Retaining motif A only, TransINT predicts 346 interactions for *H. sapiens* at a mutation rate of 0%. For the mutations rates 5, 10, 15, and 20%, the number of interactions is 16, 4439, 4272, and 724, respectively.

As mentioned, we derived a consensus sequence from multiple alignments of the contact residues of all sequences found for a given motif. Thus, for example, the consensus sequence resulting from up to a 20% mutation rate of the VV.{2}A.{2}A.{2}VL.{2}I.{3}I motif of length 19 is shown in Figure 6, leading to a fingerprint of the binding motif — $VVX_2AX_2AX_2VL$.

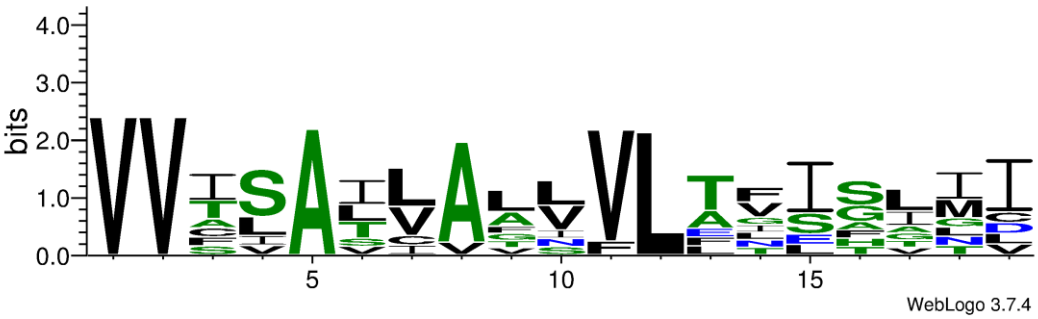


Figure 6 Multiple sequence alignment WebLogo representation (https://weblogo.berkeley.edu/) showing the consensus sequence for the VV.{2}A.{2}A.{2}VL.{2}I.{3}I binding motif of length 19. Black denotes hydrophobic residues; green, polar residues.

Some of the proteins, like the ligand-gated chloride channel human glycine receptor subunit α-3 (O75311), show several distinctly different interface motifs, suggesting a promiscuous binding behavior: 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, and I.{3}L.{2}T.{6}R.{12}D.{2}MA.{6}F.{2}LL, This property is thus linked to the multiplicity of the binding modes of proteins and to their multi-functionality.

TransINT and other datasets

The PPI template-based prediction algorithm and server PRISM2.0 (85) also uses the 3D structural similarity of interfaces with respect to templates to propose other protein complexes. PRISM is not specific for MPs and requires the 3D structure of the targets to propose an interaction. Thus, when having an interface template corresponding to an MP, it may propose not only TM protein complexes but also globular protein complexes. Therefore, many of our MP template interfaces are absent in the PRISM dataset.

Although most datasets based on experimental approaches cover the entire human proteome, again, MPs are under-represented (for an overview of the major high-throughput experimental methods used to detect PPIs, see (86, 87)). Thus, the experimental work of Rolland and colleagues on the human interactome network (25) found only 41 interactions between MPs. Twenty-eight of these proteins were found to interact in the IntAct DB. Nevertheless, none of the interactions we extracted from the structural PDBsum DB were found among the 41 interactions above. Perhaps some of these interactions may bear between the juxtamembrane regions of the MPs reported by Rolland et al. for MPs. We did not find either any of our predictions in their results. In the HI-III-20/HuRI updated dataset (10), none of the high-quality binary PPIs are MPs ($\log_2$ odds ratio < 0, Extended Data Figure 7a of Luck et al.)!

When relating our TransINT predictions at a 0-20% mutation rate for human proteins to several datasets like FPClass, BioPlex 3.0, MENTHA, IID, HMRI, and BIPS, we find several consistencies, listed in Table 3.

Table 3 Predicted interactions common to TransINT (0-20% mutation rate, *H. sapiens*) and several datasets

- FPClass, dataset of genome-wide predicted human PPIs.

Common interactions

UniProtKB AC A	**UniProtKB AC B**	**FpClass total score**
P18507	P47870	0.4029
P34903	Q16445	0.2915
P47869	Q16445	0.2738
P47870	P48169	0.2901
P47870	Q16445	0.2837
P48169	P47870	0.2901
P48169	Q16445	0.2786

Q01814	Q16720	0.3235
Q16445	P47870	0.2837
Q16445	P48169	0.2786
Q16720	Q01814	0.3235

- BioPlex 3.0 DB, which uses the TAP-MS technology for detecting PPIs, finds more than 2,000 MP interactions.

Common interactions

UniProtKB AC A	**UniProtKB AC B**	**BioPlex Probability of Interaction**
Q9Y5G3	Q9Y5G2	1.0
Q9UN71	Q9Y5G0	1.0
P47870	P28472	0.99
P51795	P51790	0.99
Q01814	P23634	0.99
Q13507	Q9HCX4	0.99
Q16445	P28472	0.99
Q9H313	O95180	0.99
Q9H313	O95427	0.99
Q9H313	Q8IYS0	0.99
Q9H313	Q96D96	0.99
Q9Y5G2	Q9UN71	0.99
Q9Y5G2	Q9Y5F8	0.99
Q9Y5G2	Q9Y5F9	0.99
Q9Y5G2	Q9Y5G0	0.99
Q9Y5G2	Q9Y5G1	0.99
Q9Y5G3	Q9UN71	0.99

Q9Y5G3	Q9Y5G0	0.99
Q9Y5G3	Q9Y5G1	0.99
P34903	P31644	0.99
Q01814	Q16720	0.99
P36021	Q8IYS0	0.99
Q9H313	Q9UP95	0.99
Q9UN71	Q9Y5G1	0.99
Q9UNQ0	O15126	0.97
Q9Y5G0	Q9Y5G1	0.89
Q8NGN1	Q9UP95	0.88
P25090	O95427	0.86
Q9UBK5	Q96D53	0.86
Q96FT7	Q68CP4	0.84
P47871	Q96D53	0.83
O75311	P28472	0.78
O75311	P28472	0.78
O75311	P28472	0.78
Q15758	Q8TCT6	0.78

For example, TransINT predicts the interaction between isoform 2 of the human γ-aminobutyric acid receptor subunit β-3 (GABRB3, P28472-2) and subunit α-6 (GABRA6, Q16445). In the BioPlex 3.0 dataset, the interaction is detected with a probability >0.99. The ComplexPortal DB reports the same interaction as Complex AC: CPX-2164 and CPX-2951. If we take 2,000 BioPlex MP interactions, then the rate of the 35 TransINT interactions is ~ 2%.

In the following, we illustrate our data for *H. sapiens* with other datasets:

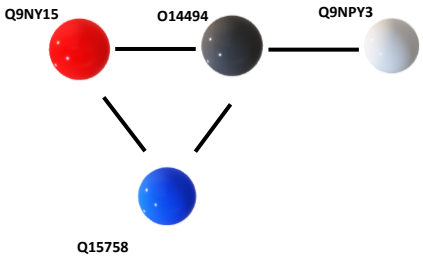
- The MENTHA experimentally-determined direct protein interactions DB presents also the P28472-Q16445, as well as the P28472-P47870 (GABRB2) interaction.

- The IID DB validates experimentally the Vascular endothelial growth factor receptor 2 (P35968) – receptor 3 (P35916) interaction.
- The HMRI DB, which seems to list only heteromers and for which not all the interactions are between MPs, shows a correspondence for the heterotypic pair TYROBP- KLRC2 (p value = 0,034341).
- BIPS (88) predicts putative interactions and is based on sequence homology between proteins found in PPI DBs and templates. We find several correlations between BIPS and TransINT. For instance, we propose an interaction between T-cell surface glycoprotein CD3 ζ chain (P20963) and high immunity immunoglobulin ε receptor subunit γ (P30273). BIPS predicts a similar pair between T-cell surface glycoprotein CD3 ζ chain (P20963) and low-affinity immunoglobulin γ Fc region receptor III-A (P08637).

Finally, as the template set is so small, we did not evaluate quantitively interface predictions using standard criteria (ROC plot, TP/FP, etc.).

Interaction networks

Networks link overlapping pairs of proteins, from which it is possible to propose multimer complexes if the binding sites are independent and non-overlapping. The architecture of a network reflects biological processes and molecular functions. Thus, from the predictions, *de novo* connections can be found, linking the network to a disease pathway, and proposing innovative possible cellular roles for some of the complexes. We illustrate below a subnetwork of TransINT-predicted MPPIs for *H. sapiens*:



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

In support of the proposed subnetwork, we found that all four MPs were present in two tissues of *H. sapiens'* — adipose tissue (major) and breast (minor) (Protein Atlas). Furthermore, a *CD93* - *STAB1* interaction in humans has been reported in the String DB of PPI networks, and *CD93* and *PLPP1* are co-expressed in *G. gallus* (89).

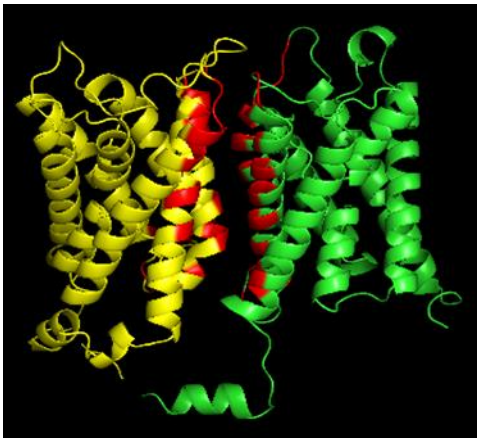
Negative interactions

The importance of recording negative results of PPI assays in interatomic DBs, i.e., those indicating the tested proteins do not interact, has been raised (90). However, their identification is less straightforward. This feature should lead us to define a set of true negative interactions for training a predictor of MPPIs as the sampling of negatives is crucial for optimal performance (91, 92). Looking for negative interactions in the IntAct DB for our TransINT proteins, we find two negative interactions of two TransINT proteins, but with two non-MPs. Analogously, the Negatome datasets (93, 94) and Stelzl (95) compile sets of protein pairs that are unlikely to engage in direct physical interactions. We observed that spanning the Negatome dataset with our predicted positive interactions for *H. sapiens* gives no results. In other words, the Negatome DB does not report any of our complexes as a negative interaction. Conversely, several MPs in the Negatome set are absent from TransINT. As this information is not conclusive, we will continue to probe our dataset as more negative interaction data become available.

Molecular docking simulations

To support the predictions from a structural point of view, we selected, for molecular docking simulations, several MP pairs for which TransINT predicted an interaction based on the interface epitopes. To begin with, we took human AQP5 (P55064) and AQP2 (P41181) proteins using the predicted pattern AV.{2}GL.{2}GA.{2}L as the binding site (S&M section). In this case, the crystal structures existed for both MPs -PDB 3D9S and 4NEF, respectively. Figure 7a shows that a heterodimer is feasible involving the corresponding membrane-exposed interface regions. On another hand, we selected the predicted rat AQP2-ErbB-3 pair, interacting through the same pattern present on both protein surfaces. Since the experimental structures are unavailable for either MP, we homology-modeled them using the 3D structures of human AQP4 (PDB 2D57) and human ErbB-2 (PDB 2JWA), respectively. The sequences of both aquaporins are more than 30% identical in the TM region, just like among ErbB-2 and ErbB-3; the resulting individual 3D models are thus highly reliable. Figure 7b shows that the modeled rat AQP2-ErbB-3 heterodimer respects the query interface, suggesting that the complex is viable.

a)



b)

Figure 7 Low-resolution cartoon structural model of predicted TransINT MP-MP complexes obtained by directed molecular docking. a) *H. sapiens'* protomers of aquaporin 5 (UniProtKB P55064, PDB 3D9S, yellow) and aquaporin 2 (P41181, PDB 4NEF, green) in complex, with TransINT interface residues in red.; b) 3D model of the *R. norvegicus'* protomers of aquaporin 2 (P34080) and the receptor tyrosine-protein kinase ErbB-3 (Q62799) in complex, with TransINT interface residues A, V, G, and L interacting through the AV.{2}GL.{2}GA.{2}L binding motif present on both protein surfaces. AQP2 interface residues are in purple and ErbB-3 residues in red. In both complexes, the contact residues are indeed at the interface of the complex and lead to its formation.

Lastly, in the TransINT user interface, when Valid=2, the 3D structures of each of the isolated protomers making up a putative complex are available. It is therefore possible to perform the directed docking implemented above to obtain 3D structures of any one of the corresponding 67 complexes (40 for *H. sapiens*). The PDB IDs can be found in the UniProtKB DB through the UniProtKB link in our database.

DISCUSSION

In this work, we developed TransINT, a template interface residue-based protocol destined to predict at large scale MP complexes through binary interactions among their α-helical TM segments. TransINT is a model-driven biological discovery tool to be queried for the discovery of verified and potential interactions, and for obtaining varied types of necessary information about them. It contains TM oligomerization recognition sites based on the assumption that homologous structural interacting motifs always interact in similar orientations and with equivalent interfaces. In this work, we report exclusively interactions in the eukaryotic plasma membrane interactome in which the binding sites specifically involve α-helical TM regions. TransINT predicts thus an α-helix - α-helix TM protein MP interactome with thousands of *de novo* interactions, including multiple recognition sites, i.e., MPs with more than one interface, important for multimer formation. The obtained sequence motifs identify homo- and heterodimer interface amino acid residues that represent the first step to generating higher-order macromolecular edifices.

The resulting MP interactome represents "first draft" predictions and contains 21,544 unique entries for all species dealt with, of which 9,798 for *H. sapiens*. The considerable number of protein partners we uncover suggests that even distantly related TM proteins frequently make use of regions of their surface with similar sequences and 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. This shows that complex formation is feasible as the interacting surfaces of the individual proteins manage to face each other in the docked complex and provides a partial validation of the method.

The comparison between our predicted set of TM complexes and other datasets is only qualitative. Disagreements occur because MPs are not well probed in experiments that employ

altogether diverse approaches. Complementary to the sequence-based co-evolution PPI prediction methods (6, 96–98), our 3D-to-1D approach encodes 3D into 1D and thus adds the spatial dimension to a given MP interactome. This may lead to pioneering biological hypotheses concerning PPIs at the membrane level, to genotype-phenotype relationships, to investigation of the effect of pathological mutations on the interaction between MPs, and to propose molecular mechanisms of action. TransINT predicts interactions for about 420 human genes, including many disease genes.

Because of their number and diversity, the higher-order structures and interfaces generated in this work represent potential pharmacological targets for the discovery of modulators targeting the intramembrane protein-protein interface. These modulators include membrane insertable, metabolically stable, non-toxic small-molecule active MPPI inhibitors, stabilizers and allosteric binders that influence *in vivo* the assembly of intact MPs involved in various diseases. Examples of these modulators are exogenous peptides and peptidomimetics (99–104). Indeed, many of the MPs in the predicted complexes are involved in a variety of diseases as deficient or enhanced oligomerization is associated with diseases (105, 106). Additionally, our protein interaction data imply direct interactions and can lead to the construction of protein interaction networks. Thus, our results may help understand biochemical pathways and the crosstalk between them, including transactivation; they may suggest as well potential drug targets, and may aid in understanding the functional effects of mutations at the interface (107). The TransINT approach should be useful as well for the design of MP structure.

Incidentally, TransINT can be extended to other cell membranes (mitochondria, nucleus, endoplasmic reticulum, Golgi apparatus) and across the tree of life, as the 3D structures of protein complexes integral to these membranes become available. Our methodology can equally be extended by introducing side-chain and main-chain h-bond and

electrostatic information at the interface, important for MPPIs (108). Other applications of our approach include homologous protein networks in other organisms and PPIs of soluble proteins.

By applying machine learning methods, the TransINT method can be improved by insuring that the MPs belong to the same developmental stage, tissue, cell type, site of expression, reaction and metabolic pathways, display functional similarity, and do not show a gene distance of more than 20 (109).

The experimental validation of the predicted MP-MP interactions could be tested with techniques such as FRET (110), or by our exclusive Microtubule Bench approach (111).

Data availability

Figure S3 shows a screen capture of TransINT's first page. The corresponding source code and instructions for the prediction algorithm has been deposited at github.com/transint. The database with the annotated predicted interactions is implemented as a web application that supports sorting and filtering. The output data can be downloaded as a .csv file and the predictions can be accessed at https://transint.shinyapps.io/transint/ or http://sertitan.imeve.univ-evry.fr:3838/transint/.

Several pertinent notes are to be found in the Supplemental Material.

Funding

This work was supported 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 innovative projects in the framework of the evolution of scientific policy.

- French Embassy in Armenia. Bourse du Gouvernement Français.

Acknowledgments

We thank Prof. D. Pastré for critical reading of the manuscript.

Conflict of Interest Disclosure

The authors declare no competing interests exist.

References

1. 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.

2. Bocharov,E.V., Mineev,K.S., Pavlov,K.V., Akimov,S.A., Kuznetsov,A.S., Efremov,R.G. and Arseniev,A.S. (2017) Helix-helix interactions in membrane domains of bitopic proteins: Specificity and role of lipid environment. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, **1859**, 561–576.

3. Yamamoto,K., Caporini,M.A., Im,S.-C., Waskell,L. and Ramamoorthy,A. (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.

4. Guidolin,D., Marcoli,M., Tortorella,C., Maura,G. and Agnati,L.F. (2018) G protein-coupled receptor-receptor interactions give integrative dynamics to intercellular communication. *Reviews in the neurosciences*, **0**.

5. 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*, 10.1002/(SICI)1097-0134(20000601)39:4<417::AID-PROT140>3.0.CO;2-Y.

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

7. Bork,P., Jensen,L.J., von Mering,C., Ramani,A.K., Lee,I. and Marcotte,E.M. (2004) Protein interaction networks from yeast to human. *Current opinion in structural biology*, **14**, 292–9.

8. Stumpf,M.P.H., Thorne,T., de Silva,E., Stewart,R., An,H.J., Lappe,M. and Wiuf,C. (2008) Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 6959–64.

9. Venkatesan,K., Rual,J.-F., Vazquez,A., Stelzl,U., Lemmens,I., Hirozane-Kishikawa,T., Hao,T., Zenkner,M., Xin,X., Goh,K.-I., *et al.* (2009) An empirical framework for binary interactome mapping. *Nature methods*, **6**, 83–90.

10. Luck,K., Kim,D.-K., Lambourne,L., Spirohn,K., Begg,B.E., Bian,W., Brignall,R., Cafarelli,T., Campos-Laborie,F.J., Charloteaux,B., *et al.* (2020) A reference map of the human binary protein interactome. *Nature*, **580**, 402–408.

11. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E., *et al.* (2012) MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Research*, 10.1093/nar/gkr930.

12. Chatr-aryamontri,A., Oughtred,R., Boucher,L., Rust,J., Chang,C., Kolas,N.K., O'Donnell,L., Oster,S., Theesfeld,C., Sellam,A., *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, **45**, D369–D379.

13. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E., *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*.

14. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A., Bork,P., *et al.* (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, **45**, D362–D368.

15. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., Del-Toro,N., *et al.* (2014) The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 10.1093/nar/gkt1115.

16. Calderone,A., Castagnoli,L. and Cesareni,G. (2013) mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods*, **10**, 690–691.

17. Sokolina,K., Kittanakom,S., Snider,J., Kotlyar,M., Maurice,P., Gandía,J., Benleulmi-Chaachoua,A., Tadagaki,K., Oishi,A., Wong,V., *et al.* (2017) Systematic protein-protein interaction mapping for clinically relevant human GPCRs. *Mol. Syst. Biol.*, **13**, 918.

18. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K.B., Gronborg,M., *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.

19. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A., *et al.* (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Res.*, **37**, D767-772.

20. Das,J. and Yu,H. (2012) HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol*, **6**, 1–12.

21. Ruepp,A., Waegele,B., Lechner,M., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Mewes,H.-W. (2010) CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic acids research*, **38**, D497-501.

22. Giurgiu,M., Reinhard,J., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Ruepp,A. (2019) CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.*, **47**, D559–D563.

23. Alonso-López,D., Campos-Laborie,F.J., Gutiérrez,M.A., Lambourne,L., Calderwood,M.A., Vidal,M. and De Las Rivas,J. (2019) APID database: redefining protein-protein interaction experimental evidences and binary interactomes. *Database (Oxford)*, **2019**.

24. Alonso-López,D., Gutiérrez,M.A., Lopes,K.P., Prieto,C., Santamaría,R. and De Las Rivas,J. (2016) APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res.*, **44**, W529-535.

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

26. Huttlin,E.L., Ting,L., Bruckner,R.J., Gebreab,F., Gygi,M.P., Szpyt,J., Tam,S., Zarraga,G., Colby,G., Baltier,K., *et al.* (2015) The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, **162**, 425–440.

27. Huang,S.-Y. (2015) 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.

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

29. Alanis-Lobato,G., Andrade-Navarro,M.A. and Schaefer,M.H. (2017) HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res*, **45**, D408–D414.

30. Miller,J.P., Lo,R.S., Ben-Hur,A., Desmarais,C., Stagljar,I., Noble,W.S. and Fields,S. (2005) Large-scale identification of yeast integral membrane protein interactions. *Proceedings of the National Academy of Sciences*, **102**, 12123–12128.

31. Garcia-Garcia,J., Schleker,S., Klein-Seetharaman,J. and Oliva,B. (2012) BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic Acids Res*, **40**, W147-151.

32. Sarkar,D., Jana,T. and Saha,S. (2015) LMPID: a manually curated database of linear motifs mediating protein-protein interactions. *Database (Oxford)*, **2015**.

33. Orchard,S., Kerrien,S., Abbani,S., Aranda,B., Bhate,J., Bidwell,S., Bridge,A., Briganti,L., Brinkman,F.S.L., Brinkman,F., *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature methods*, **9**, 345–50.

34. Carpenter,E.P., Beis,K., Cameron,A.D. and Iwata,S. (2008) Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, **18**, 581–586.

35. Mosca,R., Pons,T., Céol,A., Valencia,A. and Aloy,P. (2013) Towards a detailed atlas of protein–protein interactions. *Current Opinion in Structural Biology*, **23**, 929–940.

36. Iyer,K., Burkle,L., Auerbach,D., Thaminy,S., Dinkel,M., Engels,K. and Stagljar,I. (2005) Utilizing the Split-Ubiquitin Membrane Yeast Two-Hybrid System to Identify Protein-Protein Interactions of Integral Membrane Proteins. *Science Signaling*, 10.1126/stke.2752005pl3.

37. Rual,J.-F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N., *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.

38. Petschnigg,J., Groisman,B., Kotlyar,M., Taipale,M., Zheng,Y., Kurat,C.F., Sayad,A., Sierra,J.R., Mattiazzi Usaj,M., Snider,J., *et al.* (2014) The mammalian-membrane two-hybrid assay (MaMTH) for probing membrane-protein interactions in human cells. *Nature methods*, **11**, 585–92.

39. Hubsman,M., Yudkovsky,G. and Aronheim,A. (2001) A novel approach for the identification of protein–protein interaction with integral membrane proteins. *Nucleic Acids Research*, **29**, e18–e18.

40. Kittanakom,S., Barrios-Rodiles,M., Petschnigg,J., Arnoldo,A., Wong,V., Kotlyar,M., Heisler,L.E., Jurisica,I., Wrana,J.L., Nislow,C., *et al.* (2014) CHIP-MYTH: a novel interactive proteomics method for the assessment of agonist-dependent interactions of the human β_2-adrenergic receptor. *Biochem Biophys Res Commun*, **445**, 746–756.

41. Pitre,S., Alamgir,M., Green,J.R., Dumontier,M., Dehne,F. and Golshani,A. (2008) Computational Methods For Predicting Protein–Protein Interactions. In *Advances in biochemical engineering/biotechnology*.Vol. 110, pp. 247–267.

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

43. You,Z.-H., Lei,Y.-K., Zhu,L., Xia,J. and Wang,B. (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.

44. You,Z.-H., Zhu,L., Zheng,C.-H., Yu,H.-J., Deng,S.-P. and Ji,Z. (2014) 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.

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

46. Hashemifar,S., Neyshabur,B., Khan,A.A. and Xu,J. (2018) Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, **34**, i802–i810.

47. Qi,Y., Dhiman,H.K., Bhola,N., Budyak,I., Kar,S., Man,D., Dutta,A., Tirupula,K., Carr,B.I., Grandis,J., *et al.* (2009) Systematic prediction of human membrane receptor interactions. *Proteomics*, **9**, 5243–5255.

48. Lu,L., Lu,H. and Skolnick,J. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**, 350–364.

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

50. Maheshwari,S. and Brylinski,M. (2016) Template-based identification of protein–protein interfaces using eFindSitePPI. *Methods*, **93**, 64–71.

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

52. Li,B., Mendenhall,J. and Meiler,J. (2019) Interfaces Between Alpha-helical Integral Membrane Proteins: Characterization, Prediction, and Docking. *Computational and Structural Biotechnology Journal*, **17**, 699–711.

53. Burley,S.K., Berman,H.M., Kleywegt,G.J., Markley,J.L., Nakamura,H. and Velankar,S. (2017) Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol. Biol.*, **1607**, 627–641.

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

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

56. Laskowski,R.A., Hutchinson,E.G., Michie,A.D., Wallace,A.C., Jones,M.L. and Thornton,J.M. (1997) PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends in biochemical sciences*, **22**, 488–90.

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

58. 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.

59. Zhang,Q.C., Petrey,D., Deng,L., Qiang,L., Shi,Y., Thu,C.A., Bisikirska,B., Lefebvre,C., Accili,D., Hunter,T., *et al.* (2012) Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, **490**, 556–560.

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

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

62. Duarte,J.M., Srebniak,A., Schärer,M.A. and Capitani,G. (2012) Protein interface classification by evolutionary analysis. *BMC Bioinformatics*, **13**, 334.

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

64. Duarte,J.M., Biyani,N., Baskaran,K. and Capitani,G. (2013) An analysis of oligomerization interfaces in transmembrane proteins. *BMC Structural Biology*, **13**, 21.

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

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

67. Lomize,A.L. and Pogozheva,I.D. (2013) Solvation models and computational prediction of orientations of peptides and proteins in membranes. *Methods Mol. Biol.*, **1063**, 125–142.

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

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

70. Ulmschneider,M.B. and Sansom,M.S.P. (2001) Amino acid distributions in integral membrane protein structures. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, **1512**, 1–14.

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

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

73. Eilers,M., Patel,A.B., Liu,W. and Smith,S.O. (2002) Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys J*, **82**, 2720–2736.

74. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Research*, **49**, D412–D419.

75. Walters,R.F.S. and DeGrado,W.F. (2006) Helix-packing motifs in membrane proteins. *Proc Natl Acad Sci U S A*, **103**, 13658–13663.

76. Lemmon,M.A., Flanagan,J.M., Treutlein,H.R., Zhang,J. and Engelman,D.M. (1992) Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry*, **31**, 12719–12725.

77. Russ,W.P. and Engelman,D.M. (2000) The GxxxG motif: a framework for transmembrane helix-helix association. *J Mol Biol*, **296**, 911–919.

78. Senes,A., Gerstein,M. and Engelman,D.M. (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol*, **296**, 921–936.

79. Gernert,K.M., Surles,M.C., Labean,T.H., Richardson,J.S. and Richardson,D.C. (1995) The Alacoil: A very tight, antiparallel coiled-coil of helices. *Protein Science*, **4**, 2252–2260.

80. Sal-Man,N., Gerber,D., Bloch,I. and Shai,Y. (2007) Specificity in Transmembrane Helix-Helix Interactions Mediated by Aromatic Residues *. *Journal of Biological Chemistry*, **282**, 19753–19761.

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

82. Vreven,T., Moal,I.H., Vangone,A., Pierce,B.G., Kastritis,P.L., Torchala,M., Chaleil,R., Jiménez-García,B., Bates,P.A., Fernandez-Recio,J., *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**.

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

84. Thanos,C.D., DeLano,W.L. and Wells,J.A. (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.

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

86. Wodak,S.J., Vlasblom,J., Turinsky,A.L. and Pu,S. (2013) Protein-protein interaction networks: The puzzling riches.

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

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

89. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P., *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.

90. 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.

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

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

93. Smialowski,P., Pagel,P., Wong,P., Brauner,B., Dunger,I., Fobo,G., Frishman,G., Montrone,C., Rattei,T., Frishman,D., *et al.* (2010) The Negatome database: a reference set of non-interacting protein pairs. *Nucleic acids research*, **38**, D540-4.

94. Blohm,P., Frishman,G., Smialowski,P., Goebels,F., Wachinger,B., Ruepp,A. and Frishman,D. (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.

95. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S., *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–68.

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

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

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

99. Yin,H., Slusky,J.S., Berger,B.W., Walters,R.S., Vilaire,G., Litvinov,R.I., Lear,J.D., Caputo,G.A., Bennett,J.S. and DeGrado,W.F. (2007) Computational design of peptides that target transmembrane helices. *Science*, **315**, 1817–1822.

100. Caputo,G.A., Litvinov,R.I., Li,W., Bennett,J.S., Degrado,W.F. and Yin,H. (2008) Computationally designed peptide inhibitors of protein-protein interactions in membranes. *Biochemistry*, **47**, 8600–8606.

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

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

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

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

105. Murakami,K., Murata,N., Noda,Y., Tahara,S., Kaneko,T., Kinoshita,N., Hatsuta,H., Murayama,S., Barnham,K.J., Irie,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.

106. Watanabe,T., Kitani,A., Murray,P.J., Wakatsuki,Y., Fuss,I.J. and Strober,W. (2006) Nucleotide binding oligomerization domain 2 deficiency leads to dysregulated TLR2 signaling and induction of antigen-specific colitis. *Immunity*, **25**, 473–485.

107. Jubb,H.C., Pandurangan,A.P., Turner,M.A., Ochoa-Montaño,B., Blundell,T.L. and Ascher,D.B. (2017) Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in Biophysics and Molecular Biology*, 10.1016/j.pbiomolbio.2016.10.002.

108. Bowie,J.U. (2011) Membrane protein folding: how important are hydrogen bonds? *Curr Opin Struct Biol*, **21**, 42–49.

109. Hopf,T.A., Schärfe,C.P.I., Rodrigues,J.P.G.L.M., Green,A.G., Kohlbacher,O., Sander,C., Bonvin,A.M.J.J. and Marks,D.S. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*, **3**.

110. Chen,T., He,B., Tao,J., He,Y., Deng,H., Wang,X. and Zheng,Y. (2019) Application of Förster Resonance Energy Transfer (FRET) technique to elucidate intracellular and In Vivo biofate of nanomedicines. *Adv Drug Deliv Rev*, **143**, 177–205.

111. Boca,M., Kretov,D.A., Desforges,B., Mephon-Gaspard,A., Curmi,P.A. and Pastré,D. (2015) Probing protein interactions in living mammalian cells on a microtubule bench. *Sci Rep*, **5**, 17304.

Table and Figure Legends

Table 1 Non-redundant α-helix TransINT recognition motifs

Table 2 Benchmark of experimental structures of membrane protein-membrane protein complexes and their unbound components.

Table 3 Predicted interactions common to TransINT (0-20% mutation rate, H. sapiens) and several datasets

Figure 1 Workflow illustrating the TransINT algorithm, from information retrieval to detection of recognition motifs to generation of putative interactions, to 3D modeling of complexes.

Figure 2 Contact interface residues of chain A of the ErbB2 dimeric transmembrane domain of sequence [12]**Ile**-Ser-Ala-Val-**Val**-Gly-Ile-Leu-**Leu**-Val-Val-**Val**-**Leu**-Gly-Val-**Val**-Phe-Gly-Ile-**Leu**[31]. Contact residues are in bold.

Figure 3 Number of PDBsum non-bonded interface contacts for each of the experimental template complex PDB structures of Table S1.

Figure 4 Occurrence of (redundant) UniProtKB ACs as a function of interface contact residues for Proteins A (blue) and corresponding Proteins B (orange). Mutation rate 0-20%. Organisms extracted from the PDB and UniProtKB.

Figure 5 Number of hits per motif for all species. a) motif A, b) motif B. Mutation rate 0-20%.

Figure 6 Multiple sequence alignment WebLogo representation (https://weblogo.berkeley.edu/) showing the consensus sequence for the VV.{2}A.{2}A.{2}VL.{2}I.{3}I binding motif of length 19. Black denotes hydrophobic residues; green, polar residues.

Figure 7 Low-resolution cartoon structural model of predicted TransINT MP-MP complexes obtained by directed molecular docking. a) *H. sapiens'* protomers of aquaporin 5 (UniProtKB P55064, PDB 3D9S, yellow) and aquaporin 2 (P41181, PDB 4NEF, green) in complex, with TransINT interface residues in red.; b) 3D model of the *R. norvegicus'* protomers of aquaporin 2 (P34080) and the receptor tyrosine-protein kinase ErbB-3 (Q62799) in complex, with TransINT interface residues A, V, G, and L interacting through the AV.{2}GL.{2}GA.{2}L binding motif present on both protein surfaces. AQP2 interface residues are in purple and ErbB-

3 residues in red. In both complexes, the contact residues are indeed at the interface of the complex and lead to its formation.