TITLE

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

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

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

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

Structure-Activité des Biomolécules Normales et Pathologiques

91025 Evry

France.

ABSTRACT

Motivation: Because of their number and diversity, membrane proteins and their complexes represent potential pharmacological targets par excellence for a variety of diseases, with very important implications for the design and discovery of new compounds modulating the interaction. However, experimental structural data are very scarce. To overcome this problem, we devised a computational approach for the prediction of alpha-helical transmembrane protein higher-order structures through data mining, sequence analysis, motif search, extraction, identification and characterization of the amino acid residues at the interface of the complexes. **Results**: Our template motif-based approach using experimental recognition sites predicts thousands of binary complexes across species between membrane proteins.

Availability and Implementation: The TransINT online database of the annotated predicted interactions can be accessed on https://transint.shinyapps.io/transint/.

Contact: charbel.maroun@inserm.fr

Supplementary information: available at Bioinformatics online.

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

INTRODUCTION

Proteins are the core of the cell machinery in organisms and encompass a large variety of functions, including but not limited to activation, transport, degradation, stabilization, apoptosis, and participation in the production of other proteins. Since their complexes are the functional units of cells, it is fundamental to understand the interactions between their components. More specifically, knowledge of the 3D structure of the proteins involved and of the interfaces needed for complex formation is a fundamental phenomenon governing all processes of life (Kastritis and Bonvin, 2013).

The membrane protein (MP) interactome is the complete set of direct interactions that take place among integral-to-the-membrane or transmembrane (TM) proteins. These TM span the cell membrane in its entirety and represent approximately one third of the proteomes of organisms (Stevens and Arkin, 2000). In eukaryotes they come mostly in the form of alpha-helical domains that cross different types of cell membranes and are approximately $3 * 10^5$ in number, excluding polymorphisms or rare mutations. The number of possible binary and multiple interactions between them is thus vastly larger (for a review on the interactome, see (Lage, 2014)).

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 many different physiological processes. 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 MPs may lead to emergent properties, as a relationship exists between oligomerization and function (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 600,000 (Bork *et al.*, 2004; Stumpf *et al.*, 2008; Venkatesan *et al.*, 2009), to 3,000,000 (Luck *et al.*, 2020), several

orders of magnitude larger than the *D melanogaster* interactome. High-throughput experimental and theoretical approaches are being used to build PPI networks. The data covering PPIs is increasing exponentially. Indeed, in year 2012, there were more than 235,000 binary interactions reported (Licata *et al.*, 2012). Most protein interaction databases (DB) – 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), MENTHA (Calderone *et al.*, 2013), Clinically relevant GPCR PPI mapping (Sokolina *et al.*, 2017), HPRD (Peri *et al.*, 2003; Keshava Prasad *et al.*, 2009), HINT (Das and Yu, 2012), CORUM (Ruepp *et al.*, 2010; Giurgiu *et al.*, 2019), APID (Alonso-López *et al.*, 2019, 2016), HI-II-14 (Hwang *et al.*, 2016), HI-III-20 (Luck *et al.*, 2020), BioPlex (Huttlin *et al.*, 2015), QUBIC (Huang, 2015), IID (Kotlyar *et al.*, 2019) - 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 DBs are mostly concerned with water-soluble globular proteins and the juxtamembrane interactions of MPs. However, unlike globular proteins, MPs are water-insoluble, and for technical reasons their large-scale investigation has lagged behind (Carpenter *et al.*, 2008).

Proteome-wide maps of the human interactome network have been generated in the past (Miller *et al.*, 2005; Mosca *et al.*, 2013; Hwang *et al.*, 2016). Traditional experimental techniques like yeast two-hybrid (Y2H) assays (Iyer *et al.*, 2005) are not well suited for identifying MP interactions. Other assays, just like the Y2H assay, are depleted of MPs or unreliable (Rual *et al.*, 2005). A new biochemical technique using a split-ubiquitin membrane two-hybrid (MYTH) system has been developed; however, only a very limited number of membrane complexes have so far been determined using it (Miller *et al.*, 2005; Mosca *et al.*, 2013). This procedure has been further extended with a mammalian-membrane two-hybrid assay (MaMTH) technique (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 MP interactome. On the other hand, a variety of methods for the prediction of PPIs have been developed based on: i) machine learning and classifiers based on sequence alone: PIPE2 (Pitre *et al.*, 2008), SigProd (Martin *et al.*, 2005), PCA-EELM (You *et al.*, 2013), MCDPPI (You *et al.*, 2014), amino acid composition (Roy *et al.*, 2009), deep learning (Hashemifar *et al.*, 2018); ii) a statistical integration of biological data together with experts' feedback, HMRI, limited to the human membrane receptor interactome (Qi *et al.*, 2009a); and iii) template structures (Lu *et al.*, 2002; Szilagyi and Zhang, 2014). Again, most of the approaches address soluble globular proteins only (Keskin *et al.*, 2016). *Ab-initio* prediction of MP interfaces is rendered difficult, as these have amino acid compositions that are not very different from the rest of the protein surface, decorated by hydrophobic residues in the membrane-exposed surface.

To circumvent this problem, we developed a knowledge-based predictive approach based on the detection of alpha-helical TM contact residues issued from experimental 3D structures of MP multimers reported in the PDB (Burley *et al.*, 2017), and validated by the OPM DB (Lomize *et al.*, 2006). Querying thereafter the PDBsum DB (de Beer *et al.*, 2014; Laskowski *et al.*, 1997), we obtained the atomic details of the protein-protein interfaces, i.e. the contacts at the interface of the complexes. We then gathered those amino acids at the recognition interface to generate regular expressions or patterns that represent in a linear form the interaction motifs in space, including wildcard regions in between the interface residues. With this information in hand, we proceeded to find the obtained motifs in other MPs and a variation of them by allowing certain degrees of mutation in the membrane-embedded interface residues. This allowed us to obtain local, spatial, non-homologous interface contact sequences. Our assumption is that homologs of the template interface motifs are expected to interact in an analogous manner (Aloy *et al.*, 2003), as opposed to a global approach that would limit the search to functionally related partners without paying attention to the particular sequence at the

interface. In all cases, it is reasonable to assume that the number of interface motifs is limited in nature (Keskin and Nussinov, 2005). Thus, our approach is not focused in the overall sequence homology, such as in other template-based predictions (Zhang *et al.*, 2012; Szilagyi and Zhang, 2014). The linear 1D motifs we obtain represent thus 3D epitopes. Given that membrane proteomes may vary between species, tissues, cell types, metabolic pathways and developmental stages, we focus here on the plasma membrane of eukaryotes, ensuring thus that we are probing proteins with the same subcellular localization.

SYSTEMS AND METHODS

Algorithm

Sources of information retrieval and data filtering

Figure 1 summarizes the steps followed to collect, filter and process the input data, and generate the output. We started by obtaining a list of all eukaryote "reviewed" proteins from UniProtKB (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 identified the subset of proteins that have an experimental 3D structure in the PDB (Burley *et al.*, 2017) 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 PDB structure to be considered as valid, it had to have a resolution of 3.5 Å or less. We took in all different conformational states, regardless of 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. PDB MPs presenting engineered mutations, insertions and deletions in the TM segment with respect to the wild-type or natural variant sequence in the UniProtKB

DB were eliminated, just like chimeras of which the xenophobic part is TM. We also ignored redundant MPs, those whose 3D structure shows no significant TM segments, and those pair interactions that are redundant due to symmetry. Manual curation excluded non-parallel or perpendicular configurations, head-to-head or head-to-tail orientations of the protomers, out-of-membrane interactions only, and TM segments that do not interact. As a corollary, our approach includes implicitly the intra-molecular interactions that take place within each protomer composing an oligomer.

Finally, to ensure that the oligomer structures we took into account are quaternary structures biologically speaking, we used EPPIC, a protein-protein interface classifier (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.

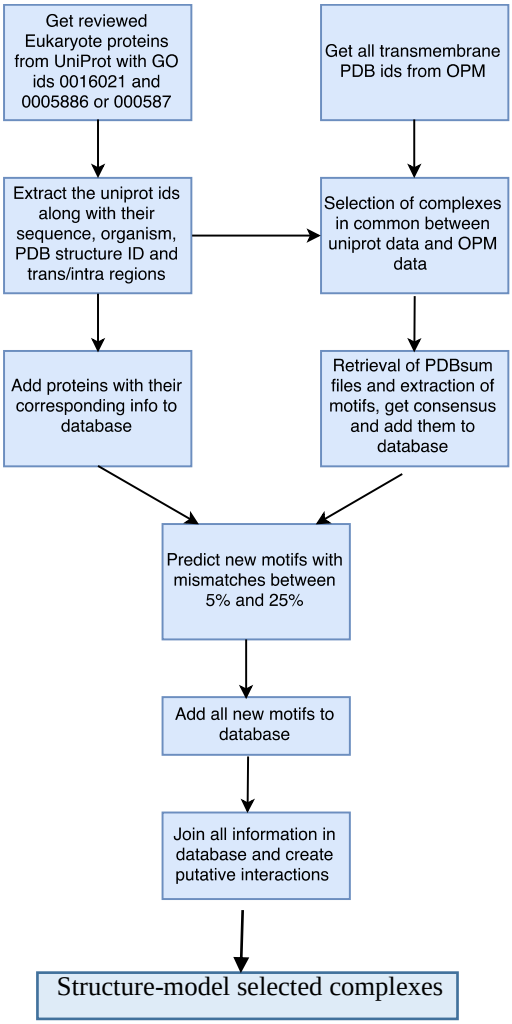


Figure 1: Workflow illustrating the TransINT algorithm, from information retrieval, to detection of recognition motifs, to generation of putative interactions, to 3D modeling of complexes.

Motif extraction

For choosing the PDB structures to work on, we referred to the OPM DB (Lomize and Pogozheva, 2013) 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 map to the MPs we extracted from UniProtKB and we extracted all the available PDBsum files of these structures. PDBsum

(Laskowski *et al.*, 1997) contains the non-bonded contacts between amino acid residues at the interface of molecules in a multimer complex. We double-checked the chosen PDB structures in the MPStruc DB of MPs of known 3D structure (https://blanco.biomol.uci.edu/mpstruc/). We then used the information in PDBsum to extract the contact residues between the couple of interacting proteins at the binding sites and the residues in between. We then defined the binding motifs by obtaining the corresponding linear sequence. From the PDBsum file listing the contacts we formulated 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 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 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, such as in I.{3}V.{3}L.{2}VL, equivalent to $IX_3VX_3LX_2VL$.

Searching for identified motifs in other protein sequences

For eliminating redundancy in our data, we grouped similar motifs together and built a consensus motif for each cluster using multiple sequence alignments. 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 UniProtKB dataset. We defined a Cost parameter as the percent of mutations allowed per motif, depending on the number of contact residues it contains and the mutation rate for the run. Cost was given as the score for substitution, while both insertions and deletions were given a score of 0 to ensure no contact residue is lost. For instance, when generating new sites from a valid motif with eight contact residues, Cost is two (1.6 rounded up) for a mutation rate of 20%. The values of Cost_A and Cost_B vary from 0-8.

To keep track of which A motif is interacting with which motif B, 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_A, this time based on an amino acid substitution matrix for MPs (Hwang *et al.*, 2016), updated by Betts and Rusell (Betts and Russell, 2003) and corrected by our means. The higher the score is, the more the specific sequence of the concerned MP fastens to the consensus binding motif. The values oscillate between 3 and 140. Afterwards, we associated the predicted motifs from new interactions based on the PDBsum-validated interactions. This way, we were sure that pattern A from new protein A "binds" to its complementary pattern B of new protein B. Since motifs can be found anywhere in the sequence, we considered only those ones in 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 that 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 (version 5.0.11) to keep all the information collected in an orderly manner. To access the DB and search for our data, a web interface was built using PHP (version 5.6.14) and HTML5, which allows the user to query the needed information. Users 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 a part thereof, using the regex format. 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 MPs of the same species. Our DB is updated following each release of UniProtKB and OPM datasets. All statistics are then regenerated.

Molecular docking

To illustrate and partially validate our approach, we generated 3D complexes derived from selected predicted pairs of MPs. 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 large amounts of experimental 3D MP structure files using several docking programs. Since we have the docking interfaces, we did not perform *ab initio* molecular calculations, nor were we concerned if 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 (Tovchigrechko and Vakser, 2006, 2005) for creating the new protein-protein 3D complexes since it 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 has also options in which the residues required to be in contact are listed. To verify the performance of GRAMM-X for MPs, we benchmarked it against several MP complexes in the PDB (unpublished data). 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-protein complexes, we chose examples in which the 3D PDB structures of proteins were already available or represented very highly homologous templates. After the 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. The obtained 3D structures of the MP-MP complexes were visualized via PyMol 2.3 (www.pymol.org).

RESULTS

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

UniProtKB provided us with 13,352 MPs that have the GO annotations mentioned in the S&M section. Overall, these proteins mapped to 954 distinct oligomer MP PDB structures. As we were only interested in structures that satisfy the requirements signified in the S&M section, we validated 50 PDB files of membrane protein-protein complexes. After checking which corresponding PDBsum files to consider, we ended up with 53 distinctly different template interactions, associated with 48 unique reviewed UniProtKB entries across species, of which 50 are structural homomers and 3 structural heteromers (Table S1). The set includes X-ray, solution NMR and electron microscopy structures. The MPs include classes such as 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 eukaryote branch of the tree of life is represented, such as *A. thaliana, B. taurus, G. gallus, M. musculus, O. sativa, R. norvegicus,* and *S. cerevisiae*. On another hand, there are 21 structures of complexes between bitopic proteins, 32 between polytopic proteins, and one mixed bitopic-polytopic protein complex. The oligomerization order is well represented, with 23 homo 2-mers, five homo 3-mers, 12 homo 4-mers, one hetero 4-mer, four homo 5-mers, three hetero 6-mers, one hetero 10-mer, and one homo 12-mer. The following EC numbers are represented in this set: 2 (10 transferases), 3 (one hydrolase) and 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 S2). The NMR-determined complexes were not submitted to the test. This is thus, the number of experimentally solved structures of MP protein complexes that we used as the template set.

The total number of motifs found after removing redundancies due to different chains of the same structures interacting more than once was 98 (Table S3). Moreover, we observed that some amino acid residues were more favored than others in the TM recognition sites. For instance, the hydrophobic side chains Leu, Ile, Val and Phe were the most common, with Leu being found more than 300 times, making about more than a third of all contact residues (Jha *et al.*, 2010) (Figure 2). The physicochemical properties of MPPI 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.*, 2019). Figure 3 shows, as a symmetric "heat map" 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 His-His for a homotypic pair, and Trp-Cys for a heterotypic pair. This outcome led 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 MP Bundles DB for alpha-helical MPs (Mayol *et al.*, 2019). The statistical trends in the contacts imply correlated mutations.

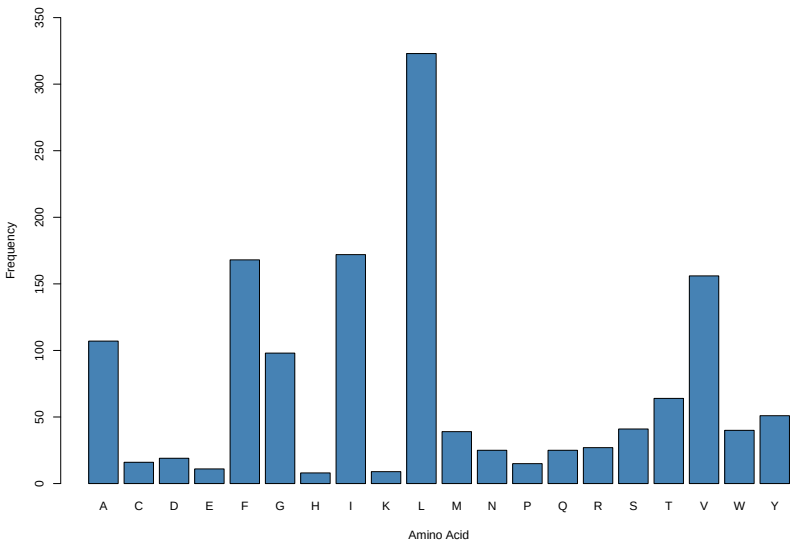


Figure 2 Frequency of amino acids in the extracted motifs from the PDBsum contact maps at the interfaces of MP complexes.

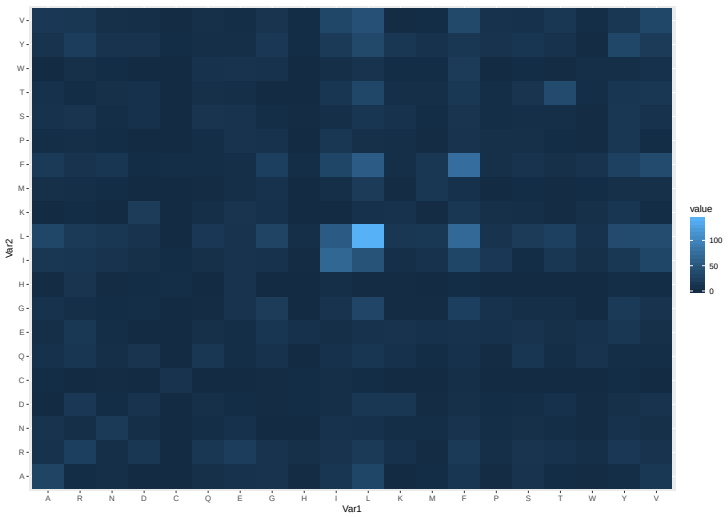
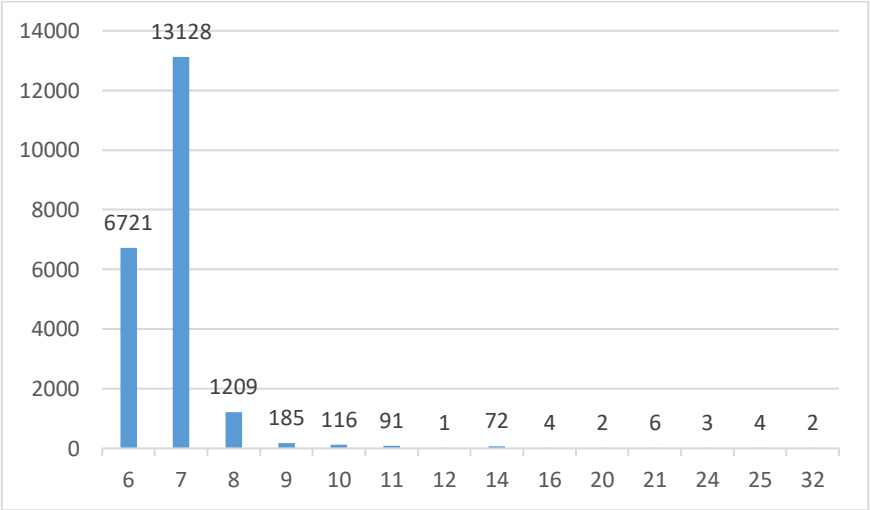


Figure 3 Symmetrical heat map of pairs of contact residues at the MP-MP interface of the template complexes. Amino acid residue names are represented by the one-letter code.

We then wanted to look at the number of motifs resulting for each number n of contact residues for the species recorded in the PDB and UniProtKB. As seen in Figure 4a and Figure 4b, the count occurrence of contact residues is largest for n = 7 for both protein A and protein B of the interacting pair. The quasi-periodicity of several interface recognition residues, reflecting a spatial arrangement corresponding to an alpha-helical pattern, along with n = 7, corresponding to one TM alpha-helix, suggests a limited set of binding motifs in nature. The corresponding statistics for *H. sapiens* show the same trends (peak at n = 7, frequency ~4827).

a)



b)

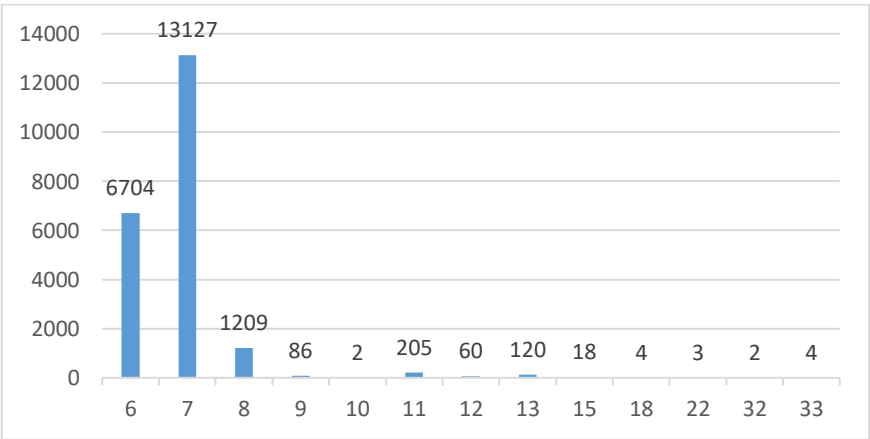


Figure 4 Count occurrence of contact residues for six or more contact residues. a) Protein A motifs; b) Protein B motifs. Mutation rate 0-20%. Organisms extracted from the PDB and UniProtKB.

The protein-protein docking benchmark 5.0 PPDB (Vreven *et al.*, 2015) 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 MP. From our 53 non-redundant template structures of protein-protein complexes, we could extract a subset of them along with the unbound structures of their components in order to define a true benchmark. This benchmark is made up of 10 sets of structures (Table 1).

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

Complex PDB code	Unbound PDB code protein A	Unbound PDB code protein B	Type Homo- (Hm) or Hetero- (Ht) meric
2LOH	2LLM	2LLM	Hm
2QTS	3IJ4	3IJ4	Hm
	4FZ1	4FZ1	Hm
	4NTW	4NTW	Hm
	4NTX	4NTX	Hm
	4NTY	4NTY	Hm

	4NYK	4NYK	Hm
2ZW3	5ER7	5ER7	Hm
	5ERA	5ERA	Hm
3SYQ	3SYA	3SYA	Hm
	3SYC	3SYC	Hm
	3SYO	3SYO	Hm
	3SYP	3SYP	Hm
	4KFM	4KFM	Hm
4JKV	4N4W	4N4W	Hm
	4O9R	4O9R	Hm
	4QIM	4QIM	Hm
	4QIN	4QIN	Hm
4NEF	4OJ2	4OJ2	Hm
4WO1	2L34	2L34	Hm
	4WOL	4WOL	Hm
4X5T	2M6B	2M6B	Hm
5A63	2N7Q	2N7Q	Ht
		(for Q92542)	
	2N7R	2N7R	Ht
		(for Q92542)	
5O9H	6C1Q	6C1Q	Hm
	6C1R	6C1R	Hm

Note: this table is derived from Table S2, column C.

When comparing our benchmark to the "Dimeric complexes" table of the Membrane Protein Complex Docking Benchmark, MemCplxDB (Koukos *et al.*, 2018), we only recover the PDB 5A63 complex, given that MemCplxDB shows many interactions between MP and non-MP, 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 MPs of beta-barrel structure. Our benchmark represents thus a gold-standard set of positives for integral-to-membrane proteins interacting through their alpha-helical TM segments.

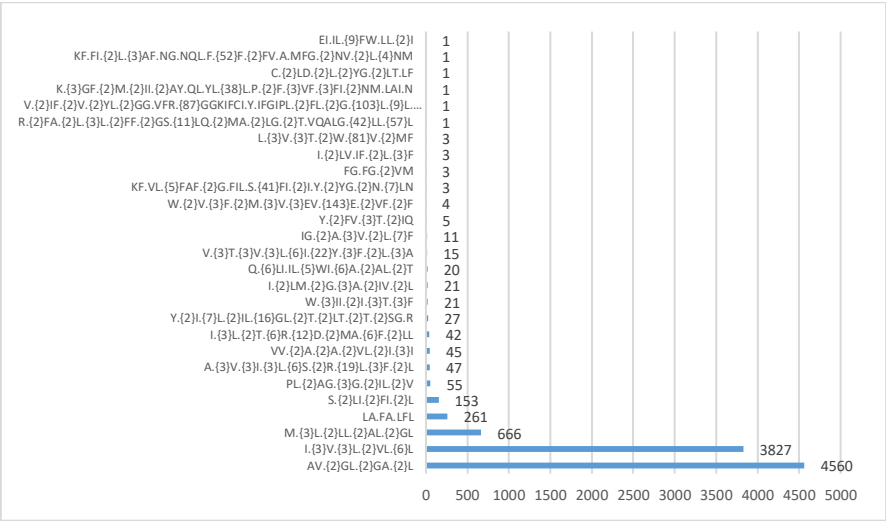
Predicted interactions

We identified many potential recognition sites and new binary complexes at different mutation rates (Figure 5). We found small motifs with six to nine contact residues extremely abundant (Figure 6). The range of contact residues in the motifs was of 6-32. As the number of contact residues increased, the number of hits decreased drastically, with only three predictions for both motifs at 32 contact residues, the longest motif. As a function of the mutation rate, TransINT predicts 346 interactions for *H. sapiens* for a mutation rate of 0% (motif A). For the mutations rates 5-20%, the number of interactions is in Figure 7.

As mentioned in the S&M section, through multiple sequence alignments, we built a consensus sequence motif for all the matched sequences of a given binding site in the 0-20 % mutation range, considering the contact residues only. This led to detection of conserved amino acid residues among the contact residues. The most prevalent consensus motif A found was AV.{2}GL.{2}GA.{2}L, illustrated by the specific sequence ^{423}AVFSGLICVAMYL^{436} of the sodium-dependent proline transporter (UniProtKB Q99884) at 15% mutation allowance (Figure 5a). For consensus motif B, two motifs are highly populated - I.{2}LV.IF.{2}L.{3}F and LL.{3}VL.SI.{2}AF.{2}G (Figure 5b). The least common motifs are characterized by

important sequence lengths and long segments in between the contact residues. An example is motif B LL.AS.{90}LV.EA.FAI.NI.{2}S.{2}L.{3}F.{19}I.{100}I, found in the short transient receptor potential channel 4 (UniProtKB Q9UBN4). Mutation rates of 0 % for protein A and protein B result in those proteins whose binding motifs conform exactly to the consensus motifs. This produces 246 entries of complexes across species and includes homo- and hetero-oligomers. Our motifs are part of conserved sequences of TM regions, especially in the case of MPs of the same family. Being contact residues, the amino acids composing our binding motifs do not represent necessarily hot-spots, i.e. those that contribute to the binding free energy (Thanos *et al.*, 2006). But, because of their degeneracy in hydrophobic amino acid type, it may be that the TM interface residues contribute equally to the free energy of binding.

a)



b)

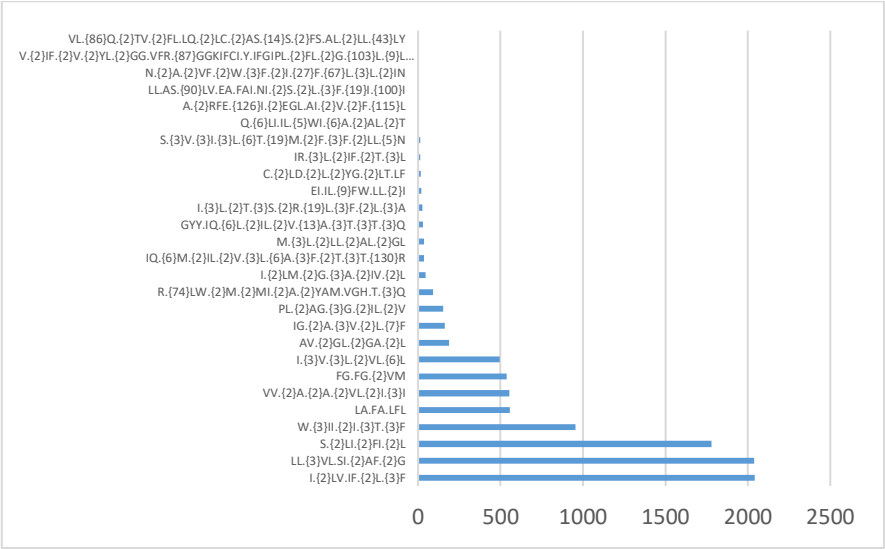


Figure 5 Number of hits per motif. a) motif A, b) motif B. Mutation rate 0-20%

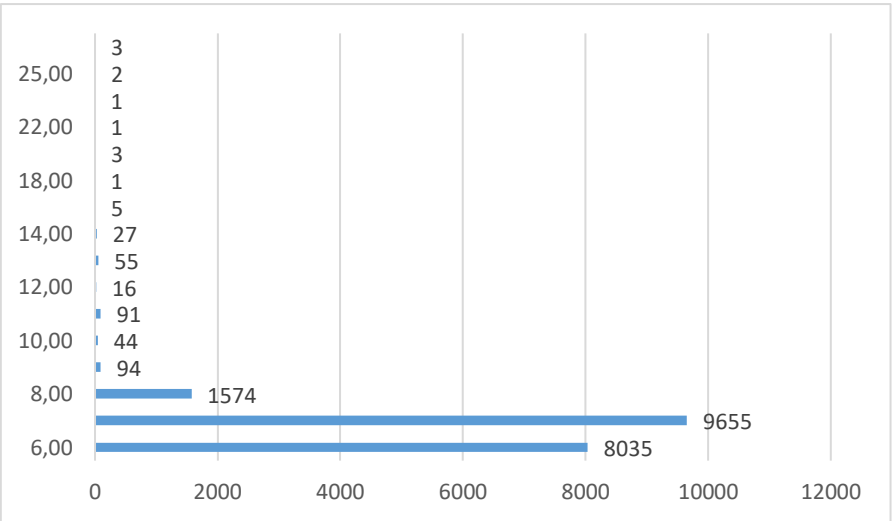


Figure 6 Frequency of predicted *H. sapiens* proteins per number of contact residues for 0-20% mutation rates. All A and B motifs combined.

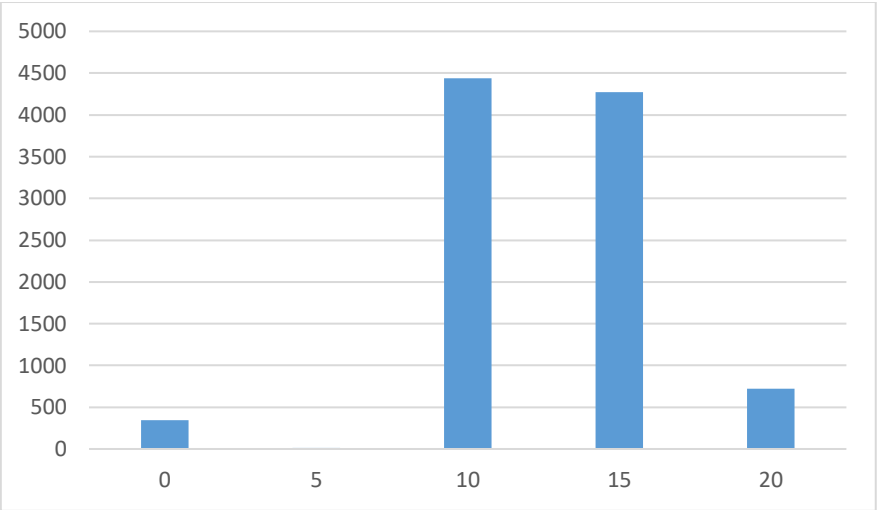


Figure 7 Number of interactions as a function of mutation rate for *H sapiens*.

As mentioned, a multiple alignment of several different protein sequences results in a consensus sequence. Thus, for example, the VV.{2}A.{2}A.{2}VL.{2}I.{3}I motif of length 19 and up to 20% mutation rate is shown in Figure 8. The consensus sequence leads to a fingerprint of that binding motif - $VVX_2AX_2AX_2VL$.

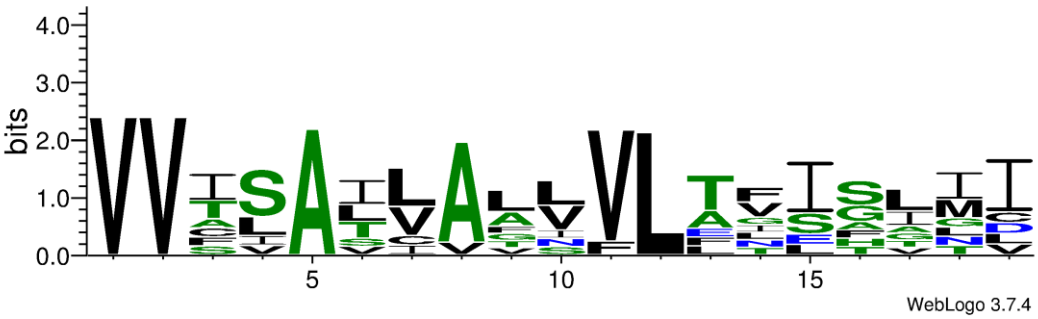


Figure 8 Multiple sequence alignment WebLogo representation (https://weblogo.berkeley.edu/) showing consensus for a motif of length 19. Black denotes hydrophobic residues; green, polar residues.

Some of the proteins, like the ligand-gated chloride channel human glycine receptor subunit alpha-3 (O75311), show several distinctly different interface motifs, suggesting a promiscuous binding behavior (Levy, 2010):

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, found by TransINT as involved in direct interactions, providing an explanation of the multiplicity of binding modes of proteins and their multifunctionality.

The PPI template-based prediction algorithm and server PRISM2.0 (Baspinar *et al.*, 2014) uses also 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 in order to propose an interaction. Thus, when having an interface template corresponding to a MP, it may propose not only TM protein complexes, but also globular protein complexes. Many of our MP template interfaces are absent in the PRISM dataset. On another hand, when comparing our dataset of interactions with that of HMRI (Qi *et al.*, 2009b), which seems to list only heteromers and for which not all the interactions are between MPs, we find a correspondence for the heterotypic pair TYROBP- KLRC2 (p value = 0,034341), for example. The prediction of putative interactions by BIPS (Garcia-Garcia *et al.*, 2012) 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 zeta chain (P20963) and high immunity immunoglobulin epsilon receptor subunit gamma (P30273). BIPS predicts a similar pair between T-cell surface glycoprotein CD3 zeta chain (P20963) and low affinity immunoglobulin gamma Fc region receptor III-A (P08637). When relating our TransINT predictions at a 0-20% mutation rate for human proteins to the FPClass dataset of genome-wide predicted human PPIs (Miller *et al.*, 2005), a data mining-based method, we find several consistencies, listed in Table 2.

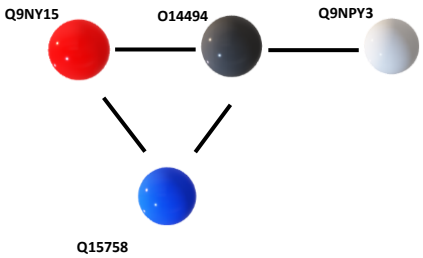
Table 2 Predicted interactions common to TransINT (0-20% mutation rate, *H. sapiens*) and FPClass.

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

*UniProtKB AC: UniProtKB accession code

Interaction networks

Networks link overlapping pairs of proteins, from which one can propose multimer complexes if the binding sites are different and non-overlapping, and if the interactions are simultaneous. From the predictions, unexpected connections can be found, and new possible cellular roles can be proposed for some of the complexes, including linking the network to a disease pathway. The architecture of a network reflects biological processes and molecular function. We illustrate below a subnetwork of 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 are present in two tissues of *H. sapiens'* -adipose tissue (major) and breast (minor) (Protein Atlas). In addition, a *CD93* - *STAB1* interaction in human is reported in the String DB of PPI networks, and *CD93* and *PLPP1* are co-expressed in *G. gallus* (Szklarczyk *et al.*, 2019).

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 (Wodak *et al.*, 2013; Rao *et al.*, 2014)). Thus, the experimental work of Rolland and colleagues on the human interactome network (Hwang *et al.*, 2016) found only 41 interactions were between MPs. Twenty-eight of these proteins were found interacting in the IntAct DB. Nevertheless, none of the interactions we extracted from the structural PDBsum DB were found among the 41 interactions. Perhaps some of these interactions are 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 (Luck *et al.*, 2020), none of the high-quality binary PPIs are MPs ($\log_2$ odds ratio < 0, Extended Data Figure 6a of Luck et al.). As for the BioPlex 3.0 DB, using the TAP-MS technology for detecting PPIs, the authors found more than 2,000 MP interactions. On another hand, TransINT predicts interactions between human Gamma-aminobutyric acid

receptor subunit beta-3 (P28472) and subunit beta-2 (P47870); and between subunit beta-3 (P28472) and subunit alpha-6 (Q16445). The latter interaction is confirmed in the BioPlex (p(interaction)>0.99) and ComplexPortal DBs (Complex AC: CPX-2164 and CPX-2951). The two interactions are listed in the MENTHA experimentally-determined direct protein interactions DB (Calderone *et al.*, 2013). In addition, when assessing our data for *H. sapiens* with the IID DB, we find, for example, that the predicted Vascular endothelial growth factor receptor 2 (P35968) – receptor 3 (P35916) interaction is validated experimentally.

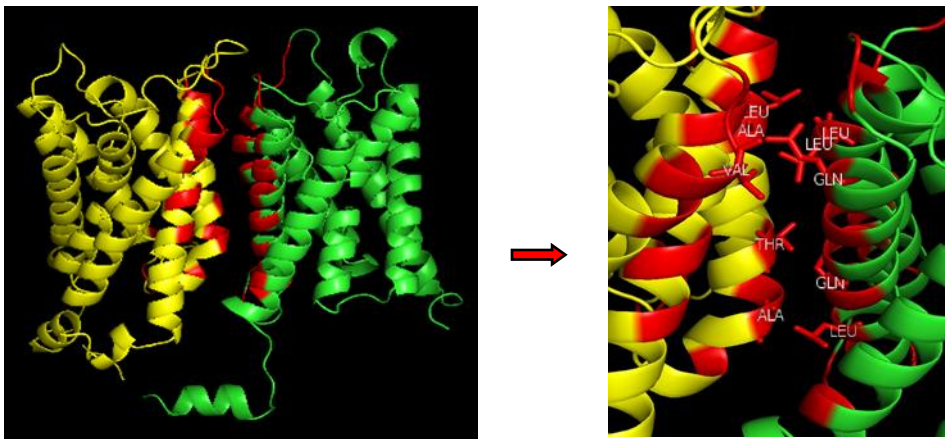
The importance of recording negative results of PPI assays in interatomic DBs, i.e. those indicating that the tested proteins do not interact, has been raised (Alvarez-Ponce, 2016). But identification of them is less straightforward. This stage should lead us to define a set of true negative interactions 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 DB for our TransINT proteins, we find two negative interactions of two TransINT proteins, but with two non-MPs! Analogously, the Negatome (Smialowski *et al.*, 2010; Blohm *et al.*, 2014) and Stelzl (Stelzl *et al.*, 2005) datasets compile sets of protein pairs that are unlikely to engage in physical direct interactions. We observed that spanning the Negatome dataset with our predicted positive interactions for *H. sapiens* results in no negative interactions, but several MPs in the Negatome set are absent from TransINT.

Molecular docking simulations

We then selected several MP pairs for which TransINT predicted an interaction based on the interface epitopes. To begin with, we took the human AQP5 (P55064) and AQP2 (P41181) protomers and proceeded to a molecular docking as described in the S&M section. In this case, the crystal structures existed for both MPs -PDBs 3D9S and 4NEF, respectively.

Figure 9a shows that a heterodimer is feasible involving the corresponding membrane-exposed interface regions. We then selected the predicted rat AQP2-ErbB-3 pair, interacting through the AV.{2}GL.{2}GA.{2}L pattern present on both protein surfaces. Since the experimental structures are not available for either MP, we homology-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, just like those of ErbB in the TM region; the resulting individual 3D models are thus highly reliable. Figure 9b shows a modeled rat AQP2-ErbB-3 heterodimer respecting the query interface, suggesting the complex is viable.

a)



b)

Figure 9: Low-resolution cartoon structural model of predicted MP-MP complexes obtained by steered 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. The figure to the right shows a zoom of the interface with the contact residues; b) 3D model of the *Rattus norvegicus'* protomers of aquaporin 2 (P34080) and the receptor tyrosine-protein kinase ErbB-3 (Q62799) in complex, with TransINT interface residues in purple and in red, respectively. In both complexes the contact residues are indeed at the interface of the complex and allow its formation.

CONCLUSION

In this work, we developed TransINT, an interface residue-based protocol destined to predict at large scale MP complexes through binary interactions among their alpha-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 different types of necessary information about them. It contains MP recognition sites based on the assumption that close homolog structural interacting motifs interact always in similar orientations and with equivalent interfaces. Only interactions in the eukaryote plasma membrane interactome in which the binding sites involve specifically TM regions are reported in this paper. TransINT predicts a MP interactome with thousands of *de novo* interactions, including multiple recognition sites, i.e. MPs showing more than one interface, leading to multimer formation. The obtained sequence motifs identify thus homodimeric interaction interface amino acid residues, which could be the first step to generate higher order macromolecular edifices.

The resulting membrane interactome represents "first draft" predictions and contains 42,746 entries for all species dealt with (19,101 for *H. sapiens*). The large number of protein partners we predict suggest that even distantly related proteins often use regions of their surface

with similar arrangements to bind to other proteins. The predicted interaction partners allowed us to generate low-resolution 3D structures for a selected number of predicted complexes, showing that complex formation is feasible as the interacting surfaces of the individual proteins manage to face each other in the docked complex.

Complementary to the sequence-based co-evolution PPI prediction methods (Liu *et al.*, 2013; Hamp and Rost, 2015; Sun *et al.*, 2017; Lage, 2014), our 3D-to-1D approach encodes 3D into 1D and adds the spatial dimension to a given membrane interactome. It may lead to new biological hypotheses concerning PPIs at the membrane level, to genotype – phenotype relationships, to investigate the effect of pathological mutations on the interaction between MPs and to propose molecular mechanisms of action.

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 protein-protein interface, such as membrane insertable, metabolically stable, non-toxic small-molecule active MPPI inhibitors or stabilizers influencing *in vivo* the assembly of intact MPs involved in various diseases. Examples of this modulators are exogenous peptides and peptidomimetics (Yin *et al.*, 2007; Caputo *et al.*, 2008; Yin and Flynn, 2016; Stone and Deber, 2017; Scott *et al.*, 2016; Corbi-Verge and Kim, 2016). 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 (Murakami *et al.*, 2011; Watanabe *et al.*, 2006). In addition, our protein interaction data implies direct interactions and can lead to the construction of protein interaction networks. Thus, our results may help understand biochemical pathways, the crosstalk between them, transactivation, may suggest potential drug targets, 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 protein complexes integral to these membranes become available. Our methodology can also be extended by using homologous protein networks in different organisms. In addition, its predictions can be refined, i.e. the number of false positives and the noise reduced by insuring the MPs belong to the same developmental stage, tissue, cell type, site of expression, reaction and metabolic pathways, display functional similarity and whose protomers in a complex do not show a gene distance of more than 20 (Hopf *et al.*, 2014), and by applying machine learning methods.

Several notes are to be found in the Supplemental material.

Funding

This work was supportted by:

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

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

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

- French Embassy in Armenia. Bourse du Gouvernement Français.

Acknowledgments

REFERENCES

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

Alonso-López,D. *et al.* (2019) APID database: redefining protein-protein interaction experimental evidences and binary interactomes. *Database (Oxford)*, **2019**.

Alonso-López,D. *et al.* (2016) APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res.*, **44**, W529-535.

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

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.

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.

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.

Burley,S.K. *et al.* (2017) Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol. Biol.*, **1607**, 627–641.

Calderone,A. *et al.* (2013) mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods*, **10**, 690–691.

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.

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

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

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

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. (2015) More challenges for machine learning protein interactions. *Bioinformatics*, **2**, 1–5.
Hashemifar,S. *et al.* (2018) Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, **34**, i802–i810.
Hopf,T.A. *et al.* (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*, **3**.
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.
Huttlin,E.L. *et al.* (2015) The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, **162**, 425–440.
Hwang,H. *et al.* (2016) A hybrid method for protein-protein interface prediction. *Protein Science*, **25**, 159–165.
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.
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.
Keshava Prasad,T.S. *et al.* (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Res.*, **37**, D767-772.
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.
Koukos,P.I. *et al.* (2018) A Membrane Protein Complex Docking Benchmark. *J. Mol. Biol.*, **430**, 5246–5256.

Lage,K. (2014) Protein–protein interactions and genetic diseases: The interactome. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, **1842**, 1971–1980.
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,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.
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*, **49**, 350–364.
Luck,K. *et al.* (2020) A reference map of the human binary protein interactome. *Nature*, **580**, 402–408.
Martin,S. *et al.* (2005) Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**, 218–226.
Mayol,E. *et al.* (2019) Inter-residue interactions in alpha-helical transmembrane proteins. *Bioinformatics*, **35**, 2578–2584.
Miller,J.P. *et al.* (2005) Large-scale identification of yeast integral membrane protein interactions. *Proceedings of the National Academy of Sciences*, **102**, 12123–12128.
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.
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*.
Peri,S. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
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.* (2009a) Systematic prediction of human membrane receptor interactions. *Proteomics*, **9**, 5243–5255.
Qi,Y. *et al.* (2009b) 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.
Roy,S. *et al.* (2009) Exploiting Amino Acid Composition for Predicting Protein-Protein Interactions. *PLOS ONE*, **4**, e7813.
Rual,J.-F. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
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. *Biochim Biophys Acta Biomembr*, **1859**, 577–585.
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. *Curr. Opin. Struct. Biol.*, **24**, 10–23.
Szklarczyk,D. *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.
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.

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**.

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. *Annu Rev Biomed Eng*, **18**, 51–76.

You,Z.-H. *et al.* (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.

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.