# Accepted Manuscript


Disordered regions in transmembrane proteins

Gábor E. Tusnády, László Dobson, Péter Tompa

PII: S0005-2736(15)00243-6
DOI: doi: 10.1016/j.bbamem.2015.08.002
Reference: BBAMEM 81965

To appear in: *BBA - Biomembranes*

Received date: 21 May 2015
Revised date: 28 July 2015
Accepted date: 9 August 2015

Please cite this article as: Gábor E. Tusnády, László Dobson, Péter Tompa, Disordered regions in transmembrane proteins, *BBA - Biomembranes* (2015), doi: 10.1016/j.bbamem.2015.08.002

# Disordered regions in transmembrane proteins

Gábor E. Tusnády[1], László Dobson[1] and Péter Tompa[1,2]

[1]Institute of Enzymology, RCNS, HAS, 1117, Budapest, Magyar Tudósok körútja 2, HUNGARY

[2]VIB Structural Biology Research Center VUB Building E Pleinlaan 2 1050 BRUSSEL

Running title: Disordered regions in TMPs

**ABSTRACT**

The functions of transmembrane proteins in living cells are widespread; they range from various transport processes to energy production, from cell-cell adhesion to communication. Structurally, they are highly ordered in their membrane-spanning regions, but may contain disordered regions in the cytosolic and extra-cytosolic parts. In this study, we have investigated the disordered regions in transmembrane proteins by a stringent definition of disordered residues on the currently available largest experimental dataset, and show a significant correlation between the spatial distributions of positively charged residues and disordered regions. This finding suggests a new role of disordered regions in transmembrane proteins by providing structural flexibility for stabilizing interactions with negatively charged head groups of the lipid molecules. We also find a preference of structural disorder in the terminal – as opposed to loop – regions in transmembrane proteins, and survey the respective functions involved in recruiting other proteins or mediating allosteric signaling effects. Finally, we critically compare disorder prediction methods on our transmembrane protein set. While there are no major differences between these methods using the usual statistics, such as per residue accuracies, Matthew's correlation coefficients, etc.; substantial differences can be found regarding the spatial distribution of the predicted disordered regions. We conclude that a predictor optimized for transmembrane proteins would be of high value to the field of structural disorder.

**KEYWORDS:** Transmembrane protein, topology, intrinsically disordered residues, prediction, positive inside rule

## 1. INTRODUCTION

Transmembrane proteins (TMPs[1]) provide the gates to the interior of the cells. They play major roles in cellular processes, such as signaling, metabolism, transports, communication, sensing and energy production. Although their functions in the living cells are crucial, their structural characterization lags far behind that of globular proteins due to both their natural dual environment, which makes their purification and crystallization tedious and because of their commonly large size, which limits the success of NMR structure determination.

The membrane-spanning parts of these proteins are highly ordered due to the low dielectric constant imposed by the hydrophobic part of the double lipid layer, but their water-solvated regions may contain intrinsically disordered regions (IDRs). According to the early study of Iakoucheva et al, about 70% of TMPs involved in signaling contain long IDRs (longer than 30 consecutive residues) [1], with somewhat lower overall rate in all transmembrane proteins (41%) [2].

The importance of intrinsic disorder stems from the special functionality it endows on proteins. The inability to fold is imprinted in the biased amino acid composition of intrinsically disordered proteins (IDPs) or IDRs. They are depleted in hydrophobic residues that usually drive protein folding, and are enriched in charged and polar residues which prefer to stay in contact with water [3,4]. This makes IDPs/IDRs assume an unfolded and extended, highly flexible structural ensemble, which is compatible with special functional modes [5]. Often, disordered proteins function by molecular recognition, when their short motifs [6] or longer disordered domains [7] bind partner molecules in a process of induced folding [8], with the possible advantage of uncoupling specificity from binding strength and increased speed of interaction. Regions connecting binding motifs, and sometimes entire proteins, can also act as disordered – also termed entropic – chains, when they enable almost unrestricted conformational search for the flanking binding elements. Due to their extended conformation and exposure to other proteins, disordered regions are also the preferred sites of post-translational modifications (phosphorylation, ubiquitination, etc…), and hence they frequently mediate regulatory input [9]. As a result of these special functional modes, structural disorder correlates with signaling and regulatory functions, and is depleted in proteins playing biosynthetic, and metabolic roles [10]. In all, due to the special functional advantages endowed by structural disorder, it is more

[1] Abbreviation used: MCC, Matthew's correlation coefficient; IDP, intrinsically disordered protein; IDR, intrinsically disordered region; PDB, Protein Data Bank; TMP, transmembrane protein;

abundant in eukaryotes than prokaryotes, although it varies a lot in both domains of life in reflection of lifestyle [11,12].

IDRs often combine with folded domains in the same protein, thus structural disorder is often observed in the Protein Data Bank (PDB) [13]. In multidomain proteins, it gives rise to complex regulatory phenomena [14] by promoting interactions with external partners, and autoregulatory interactions within the protein, both of which are subject to complex input by other partners and modifications. This is clearly the case in TMPs, as demonstrated in several concrete instances, which provide some insight into the role of IDRs in TMPs. Structural disorder can occur both in their terminal [15,16] and loop regions [17,18], and it can mediate interactions with external partners [15], or send regulatory input within the protein [16,19]. Further to these individual examples that rely on detailed experimental investigations, however, general bioinformatics studies on the extent, localization and possible function of IDRs in TMPs, drawn from applying a single disorder predictor, should be treated with extreme care.

The skewed amino acid composition of IDPs/IDRs provides the basis of several disorder prediction methods [20–22]. Most of these methods average the amino acid composition within a large sequence window, however, the presence of large hydrophobic segments such as the transmembrane α-helices in TMPs may seriously compromise the use of prediction algorithms in case of transmembrane proteins. The accuracy of various prediction methods of structural disorder was investigated on TMPs with known 3D structures by Pryor and Wiener [23]. They showed a clear division between programs that accurately predict structural disorder in membrane proteins and programs that fail, which allowed the authors to integrate these methods into their membrane protein structural genomics pipeline.

The PDB is usually the primary source for the analysis of "observed" IDRs and a starting point to preparing training set for prediction methods both in globular [13,20,24] and transmembrane [2,23,25–30] proteins. However, this is a data source of limited value for disordered regions. The investigation of the human proteome revealed that approximately 40% of the residues fall within, and 60% outside, SCOP domains [4], and the fraction of disordered residues is only 1-2% inside SCOP domains. Therefore, the statistical analysis of protein domains in the PDB utilizes only very limited information about disordered residues. Since transmembrane protein structures are underrepresented in the PDB [31], this situation is even more severe in the case of TMPs. Thus, the results of any study of disordered residues in TMPs relying on structural information in the PDB should be strongly biased and may be representative of only a small fraction of proteins.

In this article, by applying a stringent definition for disorder regions, we analyzed their various statistics in TMPs on the currently available largest structural set. We also critically compare the results of the various prediction methods on this set to raise guidelines for their use. Finally, we describe the various roles of IDRs in TMPs currently described in the literature.

## 2. METHODS

### *2.1. Databases, determination of disordered regions and topology*

TOPDB database (version 2.0) [32,33] was downloaded from http://topdb.enzim.hu and each entry containing a cross reference to the PDB has been selected. These entries were filtered to 40% sequence identity by CD-HIT algorithm [34–36]. Then, homologous proteins in the PDB for each entry were collected by BLAST [37] (with e-value $10^{-10}$), and pairwise alignments were made by ClustalW [38] if the resolution of the homologous PDB entry, determined by X-ray crystallography, was less than 3.5 Å. The homolgous PDB structures were used, if the pairwise sequence identities were above 60%. Disordered regions were inferred indirectly, considering a residue disordered if atomic co-ordinates of its side-chain were missing from all homologous PDB entries. This information was mapped back onto the sequence of the original TOPDB entry through the sequence alignment. If a disordered region overlaps with a transmembrane region, the entire disordered region was disregarded. Disordered regions were also disregarded if they were closer than five residues to the C- or N-terminus of the corresponding PDB entry (for example, a water-soluble structured region of TMP) and were not longer than 5 residues. Those regions, which were not covered by any homologous PDB structure, were masked out. For evaluating the predictions, only unmasked regions were taken into account (Figure 1/A).

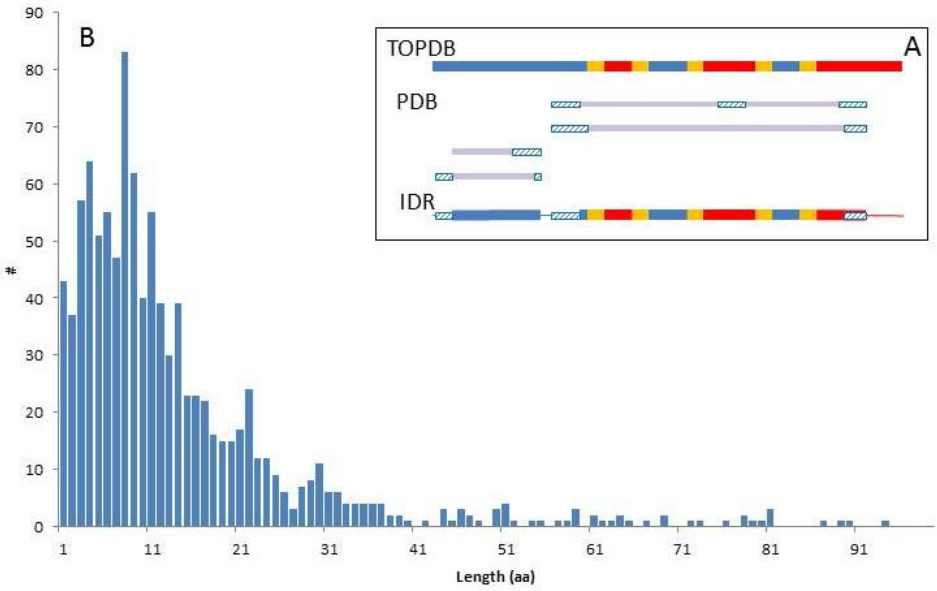



**Figure 1. Definition of IDRs and their length distribution. A:** The definition of IDRs. TOPDB line: topology defined in TOPDB database; PDB lines: 3D structures of identical or homologous proteins; IDR line: final definition of IDRs. Blue: outside, Orange: transmembrane region, Red: inside, striped box: residues in 3D structure with no atom coordinates, thin line:

regions wich are not covered in any PDB files, thick magenta: regions with determined 3D structure. **B:** The distribution of disordered region lengths.

The resulting sequences were divided further in several ways. First, we separated α-helical and β-barrel proteins, according to the annotations of TOPDB entries and called these two sets TMA and TMB, respectively. TMA set were further split into two subsets according to the number of transmembrane segments: the first contains bitopic proteins (TMAB set, 1TM proteins), the second one contains polytopic proteins (TMAP set, MultiTM proteins).

### *2.2. Statistics of disordered residues through the z-axis (z-coordinate dependent distribution of disordered residues)*

All entries with observed IDRs were transformed by bringing the membrane normal parallel to the z-axis using the transformation matrix provided by corresponding PDBTM entries. The proteins were then cut to 5Å slices parallel to the xy plane, and in each slice the positively charged residues, the starting and end point of disordered regions (actually, the coordinates of the last residue before, and the first after, the disordered region with a resolved structural position, see below) were counted. These counts were summed up for all entries in the TMA set. Finally, these counts were normalized to one.

### *2.3. Prediction methods*

For disorder prediction, we used the same in silico programs as Pryor and Wiener [23], plus Dynamine [39,40] and ANCHOR [41,42]. In case of IUPred [21,43] and ANCHOR, in addition to the regular usage, we implemented the following modifications on the input sequences: i, the transmembrane segments (io); ii, the transmembrane segments plus 15 amino acids in both directions (io15); iii, and the transmembrane segments plus 30 amino acids in both directions (io30) were cut out, and the remaining "inside" and "outside" sequence parts of each protein were all linked together to produce an arbitrary "in" and "out" protein, respectively. These "in" and "out" sequences were submitted to IUPred and ANCHOR programs, and the resulting predictions were mapped back onto the original sequence. The prediction results were modified, if a predicted disordered region was found to overlap with a transmembrane segment, then it was disregarded as a disordered region. For the other methods, we followed the procedure described in Pryor and Wiener [23]. In total, 12 prediction methods were tested, altogether with 28 flavors, as given in Table 1.

| Name | Flavor(s) | Description, Reference(s), URL |
|---|---|---|
| Anchor | -, IO, IO15, IO30 | Prediction of protein binding regions in disordered proteins. The original and an in house modified version were used (see Methods) [42,44]. http://anchor.enzim.hu |

| | | |
|---|---|---|
| Disembl | Coils, Hotloops, Rem365 | Disembl uses several alternative definitions, and a new one based on the concept of ``hot loops" was also introduced [45]. http://dis.embl.de/ |
| Disopred | Diso, Pbdat | The DISOPRED server uses a knowledge-based method to predict dynamically disordered regions from the amino acid sequence [46]. http://bioinf.cs.ucl.ac.uk/disopred/ |
| Disprot | VSL2B | The VSL2 predictors are applicable to disordered regions of any length and can accurately identify the short disordered regions that are often misclassified by other disorder predictors [20]. http://www.dabi.temple.edu/disprot/predictorVSL2.php |
| DynaMine | | DynaMine is a fast predictor of protein backbone dynamics [39,40]. http://dynamine.ibsquare.be |
| ESpritz | Disprot, NMR, X-ray | ESpritz methods based on bidirectional recursive neural networks and trained on three different flavors of disorder, including an NMR flexibility predictor [47]. http://protein.bio.unipd.it/espritz/ |
| FoldIndex | | FoldIndex uses the algorithm of Uversky and co-workers, which is based on the average residue hydrophobicity and net charge of the sequence [48]. http://bioportal.weizmann.ac.il/fldbin/findex |
| GlobPlot | | GlobPlot is a web service to plot the tendency within the query protein for order/globularity and disorder [22]. http://globplot.embl.de |
| IUPred | -, long, long IO, long IO15, long IO30, short, short IO, short IO15, short IO30 | IUPred bases on the estimated pairwise energy content of proteins. The original and an in house modified version were used (see Methods) [21,43]. http://iupred.enzim.hu |
| Predisorder | | PreDisorder is based on MULTICOM-CMFR ab initio prediction method, which was ranked among the top disorder predictors in CASP8 [49]. http://sysbio.rnet.missouri.edu/predisorder.html |
| Ronn | | The regional order neural network (RONN) software is based on bio-basis function neural network pattern recognition algorithm |

| | | |
|---|---|---|
| | | [50].<br>https://www.strubi.ox.ac.uk/RONN |
| SPINE-D | | SPINE-D is a three-state single neural-network-based method [51].<br>http://sparks-lab.org/SPINE-D/ |

**Table 1.** Disorder prediction methods evaluated on transmembrane protein sequences. The '-' flavors refers to the original methods.

### *2.4. Evaluation of the methods*

For evaluating the methods, we utilized the metrics previously introduced in CASP experiments [52–55]. Some of them were also used by Pryor and Wiener [23], and Walsh et al [56]. They are listed and defined in Supplementary Document S1.

## 3. RESULTS AND DISCUSSION

### *3.1. Investigation of observed disordered regions*

#### *3.1.1. Stringent definition of disordered regions*

Here we used a stringent definition of disorder, by requiring that coordinates of the given residue are missing in all the homologous protein structures found in the PDB. Although this definition decreases the number of determined disordered residues, it does eliminate the ad hoc errors of structure determinations.

We selected 1162 transmembrane proteins from the recently updated TOPDB database [32,33] below 40% sequence identity. 1056 of these proteins have at least one homologous structure in the PDB. Our stringent definition results in 631 TMPs (60%) with one or more IDRs, 44 (4%) of them are β-barrel proteins (TMB set), 332 (31%) are bitopic (TMAB set) and 255 (14%) are polytopic (TMAP set) α-helical proteins. In the whole dataset there are 20050 disordered residues (5.19%) in 1189 IDRs. Most of these residues are in short IDRs (shorter than 30 residues) (15964 residues in 1049 IDRs), and only 4086 residues were found in 140 IDRs longer than 30 residues. The distribution of the length of disordered regions is shown in Figure 1/B.

As expected, the number of short IDRs is lower than in the work of Pryor and Wiener [23], due to our stringent IDR definition, which eliminates the sporadic errors generated during structure determination.

Although the TMA and TMB sets contain proteins from different taxonomic sources, (there are 631 and 410 eukaryotic and prokaryotic proteins altogether in the two sets, respectively), the distribution of the number of transmembrane α-helices in the TMA is similar to that of human transmembrane proteome [57]. In both sets, 1TM proteins are the most abundant, and the second most prevalent group is the seven TM proteins with extra-cytosolic N-termini (Supplementary

Figure S1). In accord, the results and conclusions drawn in the present work may apply for the human proteome as well. In the following analyses we use data only from the TMA sets.

### *3.1.2. Disordered regions are abundant in the N- and C-terminal regions*

Although the short disordered regions near the N- or C-termini of PDB entries were removed during our IDR definition, the distribution of IDRs in sequences is still biased (Figure 2). We divided each sequence into ten equal parts, and the proportion of disordered residues was counted in each part. Residues before the first and after the last transmembrane segment (termed terminal regions) were counted separately from residues located between two transmembrane segments (loop regions). As seen in Figure 2, the first and last tenth of the sequences contain more than two times more disordered residues than their middle regions. Further, the IDRs tend to be separate from transmembrane regions, i.e. the proportion of disordered residues in terminal regions is 5 times more frequent than the proportion of disordered residues in loops. As discussed above, data sets using PDB files and inferring disordered residues from missing coordinates may have been biased. From these data, we cannot ascertain whether this non-uniform distribution of IDRs along the sequence results from a bias in structure determination techniques or it is an intrinsic property of proteins.

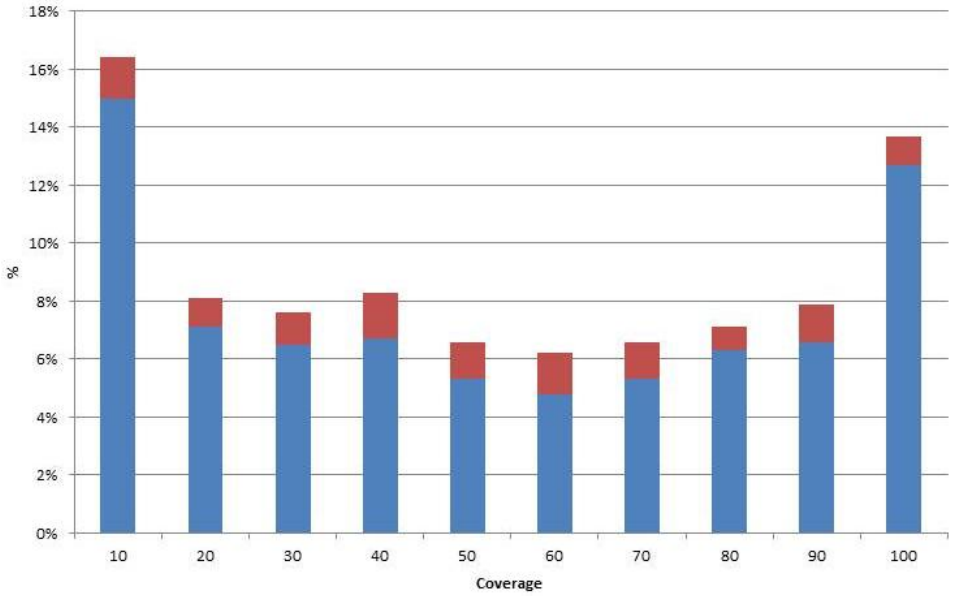



**Figure 2. The distribution of IDRs in the sequence.** X-axis shows the coverage of the sequences. Blue: proportion of disordered residues before the first transmembrane segment or after the last transmembrane segment (terminal regions). Red: proportion of disordered residues between two transmembrane segments (loop regions).

### *3.1.3. Outside parts of terminals are more abundant in IDRs than inside parts*

We have also examined the distribution of IDRs and the distribution of residues in IDRs in the terminal regions (before the first or after the last transmembrane segment) and between two transmembrane segments (loop regions). Figure 3 shows the distribution of IDR containing (Panel A) and all (panel B) terminals and loop regions, and the ratios of these two (Panel C). The highest proportion of IDR can be found in the outside N-terminal part of proteins in TMAB set, while the second most abundant class regarding the absolute values are the loop regions in proteins in TMAP set. Taking into consideration the relative values, the most prevalent classes are the proteins containing IDRs outside of the N- and C- terminal parts in 1TM proteins (TMAB set). In the inside regions, the relative frequency of IDR containing regions is larger only in the loop regions in multiTM proteins (TMAP set).

We have also calculated the relative frequencies of residues within IDRs in the regions mentioned above (terminal and loop regions) (Figure 3, Panel D). In this calculation, only the C-terminal regions of bitopic transmembrane proteins contain more disordered residues in the outside than in the inside part, whereas in all the other cases the inside parts tend to contain relatively more disordered residues than the outside parts. Summarizing these findings, while the outside parts of transmembrane proteins contain more IDRs in general, with respect to the relative frequencies of residues to be within IDRs, the inside regions contain more disordered residues.

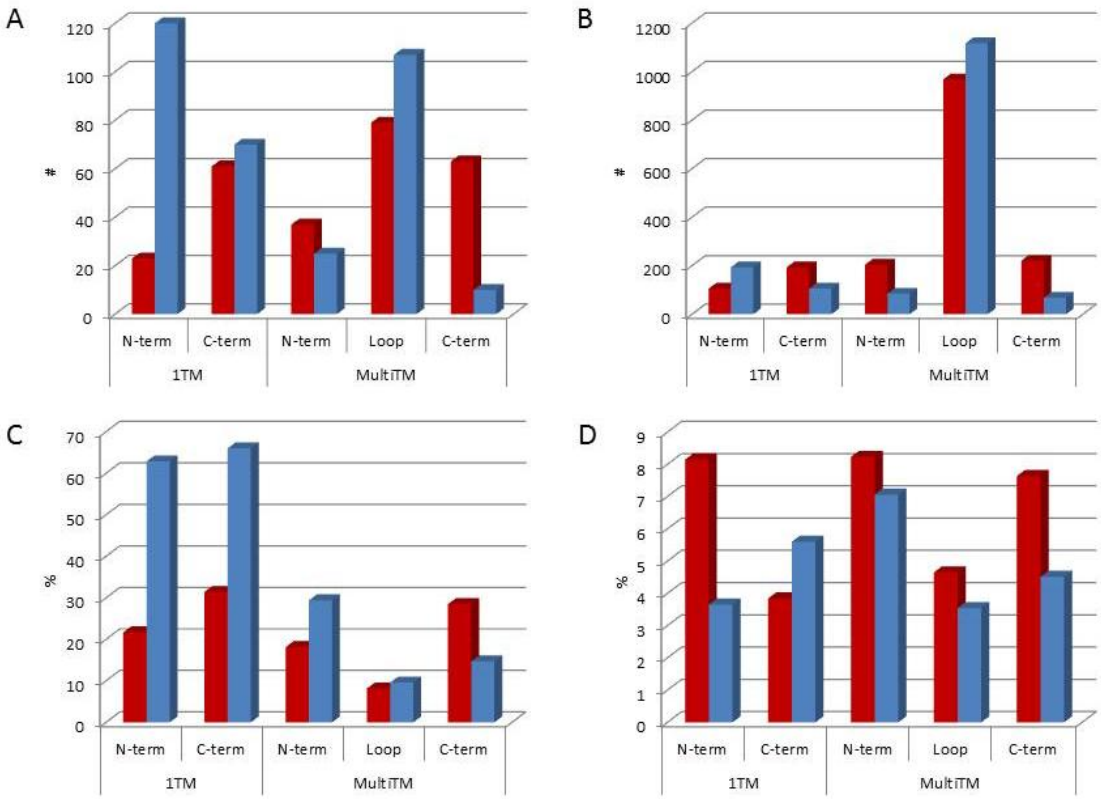



**Figure 3. Distribution of IDRs in terminals and loop regions**. A: Number of proteins containing IDRs in terminal and loop regions; B: Number of regions in the indicated parts of the transmembrane proteins; C: Relative frequencies of regions containing IDRs in the specified regions (i.e. the number of IDR containing regions divided by the number of regions in the certain part of the protein). D: Relative frequencies of residues in IDRs (i.e.the number of residues in IDRs divided by the number of all residues in the specified regions). Red: inside, Blue: outside.



### *3.1.4. Distribution of IDRs in terminals differs in 1TM and multiTM proteins*

We have also investigated the distributions of IDRs in the terminal regions for 1TM (TMAB set) and multiTM (TMAP set) proteins (Table 2 and Table 3, respectively). In 1TM proteins, most frequently, if only one terminal contains IDR or IDRs, it is C terminal part in Nin-Cout proteins, whereas in Nout-Cin proteins it is the N terminal, which is in line with the observed relative frequencies of IDRs described in the previous paragraph. In some cases both terminals contain IDRs (4.64% and 1.89% for Nout-Cin and Nin-Cout proteins, respectively), but this situation is the less frequent. The second less frequent case is when none of the terminals contains IDRs, nevertheless the relative frequencies of these are above 10%.

| | | Nin | | | Nout | |
|---|---|---|---|---|---|---|
| Cin | | N+ | N- | | N+ | N- |
| | C+ | - | - | C+ | 4.64 | 26.8 |
| | C- | - | - | C- | 58.25 | 10.31 |
| Cout | | N+ | N- | | N+ | N- |
| | C+ | 1.89 | 64.15 | C+ | - | - |
| | C- | 19.81 | 14.15 | C- | - | - |

**Table 2. Joint distributions of IDRs in TMAB set (1TM proteins).** + indicates that the specified region contains IDR, - indicates the whole region is ordered.

| | | Nin | | | Nout | |
|---|---|---|---|---|---|---|
| Cin | | N+ | N- | | N+ | N- |
| | C+ | 2.4 | 28.74 | C+ | 5.17 | 13.79 |
| | C- | 14.97 | 53.89 | C- | 32.76 | 48.28 |
| Cout | | N+ | N- | | N+ | N- |
| | C+ | 0 | 16.22 | C+ | 0 | 14.89 |
| | C- | 21.62 | 62.16 | C- | 11.11 | 74.07 |

**Table 3. Joint distributions of IDRs in TMAP set (MultiTM proteins).** + indicates that the specified region contains IDR, - indicates the whole region is ordered.

For multiTM proteins, most frequently none of the terminals contains IDRs, but similarly to 1TM proteins the less frequent cases are when both terminals contain IDRs. Intriguingly, this can be observed only in transmembrane proteins with their C terminus inside.

#### *3.1.5. Positively charged residues and disordered residues are highly correlated in the 3D structure*

For each α-helical transmembrane protein, we have determined the relative frequencies of positively charged amino acids along the z-axis. To this end, all proteins were transformed according to the rotational matrix given in the corresponding PDBTM entry, so that the membrane planes became parallel to the XY plane, and the positive z-axis pointed toward the inside (cytosolic) part of the membrane, while the negative toward the outside (extra-cytosolic) part. Then, we cut the proteins into 5Å wide slabs parallel to the XY plane and counted positively charged amino acids in each slab for all proteins in TMA set. Finally, we calculated the relative frequencies of positively charged amino acids along the z-axis by dividing these

counts by the number of all positively charged amino acids in our TMA set. Since the 3D coordinates of disordered residues are not known, we used the coordinates of the last known Ca before the disordered region and the coordinates of the first known Ca after the disordered region. We called these positions as "stems" of the IDRs. We calculated the relative frequencies of "stems" of IDRs along the z-axis for IDRs, which are in terminal or loop region, separately (Figure 4). As expected, positively charged amino acids have a peak at +15Å where the negatively charged lipid head groups interact with positively charged amino acids, and according to the positive inside rule [58], the peak inside is higher than outside. Interestingly, the distribution of IDRs in loop regions shows a similar distribution, with a correlation coefficient of 0.87. However, IDRs in terminal regions show a different distribution, with a smaller peak at about 25Å inside with a correlation coefficient of 0.42 only. This difference between the distributions of IDRs to be in loop or terminal regions may reflect the different roles of residues in these regions. The IDRs in loop regions may have the role of promoting the positively charged amino acids to form favorable interaction with lipid head groups, which may be not so favorable if the positively charged amino acids are in a rigid structure. However, the IDRs in terminal regions may have different roles, as discussed later.

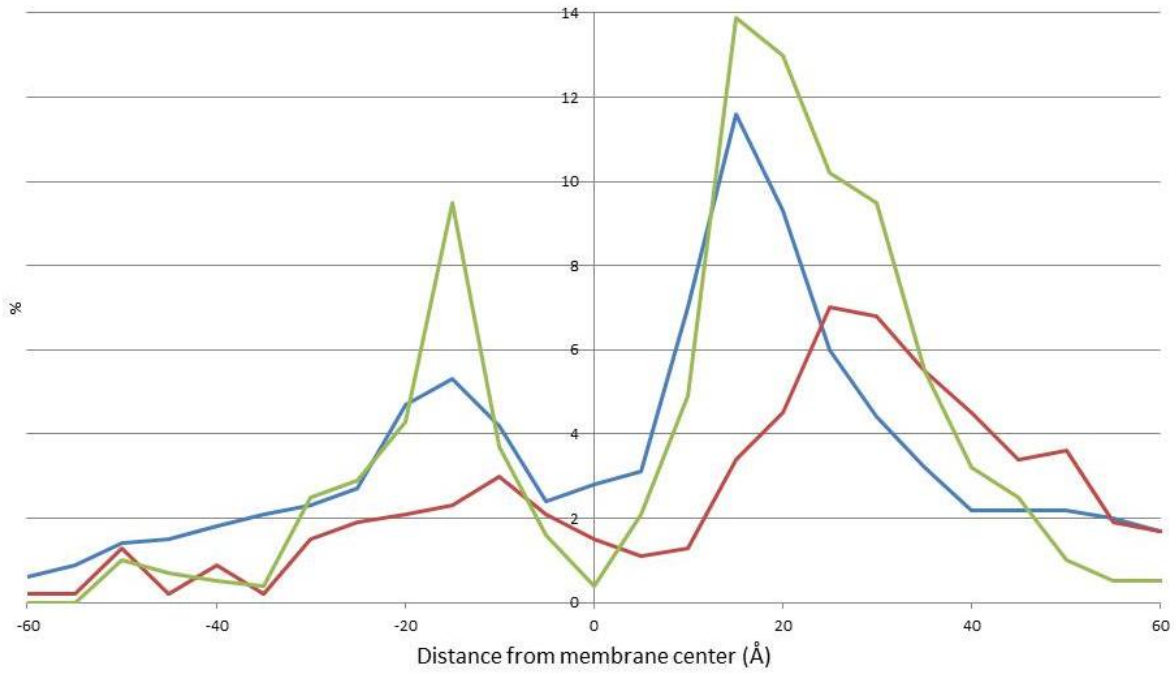



**Figure 4. Z-coordinate dependent distributions.** Distribution of positive charged amino acids (blue line) and the stem position of IDR to be in terminal (red line) or loop (green line) region along the z-axis, normalized by the number of all amino acids having the given properties. Negative z-coordinate means outside, positive one means inside regions.

We have to note that a similar observation was described in the case of archaerhodopsin-2 (aR2), where disorder was observed for lipids filling the inter-trimer space and the side chains of arginine residues (Arg34 and Arg230) that can interact with negatively charged lipid head groups [59]. Moreover, it was shown that arginine and lysine are highly "disorder-promoting" amino acids [60], therefore the high frequencies of these amino acids close to the lipid head groups of the inside leaflet may promote disordered structure in loop regions.

### *3.2. Prediction methods*

#### *3.2.1. Evaluating of disorder prediction methods*

Our large collection of stringent IDR cases in TMPs also allows us to compare different disorder prediction algorithms to formulate guidelines for their use. Even on soluble proteins, different predictors tend to disagree on functionally important disordered regions [61]. Since most methods take amino acid frequencies of longer regions into consideration, and TMPs contain long hydrophobic segments close to their IDRs, it can be assumed that disorder prediction accuracies may be lower on transmembrane proteins than on globular proteins. In fact, the accuracy and Matthew's Correlation Coefficients are very low for most prediction methods (between 0.77-0.53 and 0.36-0.05 for Acc and MCC, respectively). Moreover, the variations of the values in Table 4 are small, i.e. the methods used have similar performance. Regarding the modified IUPred predictions, prediction accuracies of all flavors are better than the original version on IDRs in loop regions, supporting our original hypothesis that the presence of hydrophobic transmembrane segments interferes with prediction methods, which use the average amino acid composition of longer segments. In terminal regions, the effect of hydrophobic segments is smaller, therefore this trend changes in the case of the IUPred short predictions.

| Method | ACC | ACC2 | SENS | SPEC | MCC | TM% | RD | RX2 | SOL | FPREG | AUC | RMSE | PCC | Z-corr | Order (avg) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predisorder | 0.76 (2) | 0.82 (19) | 0.69 (4) | 0.83 (22) | 0.33 (2) | 1.21 (18) | 73.08 (2) | 66.73 (13) | 0.56 (2) | 6488 (25) | 0.6 (14) | 0.43 (19) | 0.4 (2) | 0.95 (1) | 4.57 |
| Espritz_xray | 0.71 (4) | 0.88 (17) | 0.45 (9) | 0.92 (17) | 0.32 (3) | 3.9 (24) | 50.08 (9) | 12.11 (3) | 0.42 (7) | 2226 (7) | 0.63 (9) | 0.34 (17) | 0.38 (4) | 0.88 (4) | 7.14 |
| Disopred_diso | 0.7 (8) | 0.91 (5) | 0.47 (8) | 0.95 (8) | 0.36 (1) | 4.33 (25) | 47.87 (11) | 99.6 (17) | 0.41 (9) | 3090 (12) | 0.66 (2) | 0.3 (6) | 0.41 (1) | 0.47 (19) | 7.57 |
| Disprot | 0.73 (3) | 0.81 (23) | 0.62 (6) | 0.81 (23) | 0.28 (8) | 0.33 (14) | 66.26 (6) | 11.72 (2) | 0.51 (5) | 4801 (21) | 0.59 (16) | 0.44 (23) | 0.36 (5) | 0.92 (2) | 7.57 |
| Sipne-D | 0.77 (1) | 0.79 (24) | 0.76 (1) | 0.79 (24) | 0.32 (4) | 8.9 (26) | 72.23 (3) | 30.16 (5) | 0.53 (3) | 4312 (17) | 0.59 (15) | 0.46 (24) | 0.4 (3) | 0.53 (17) | 7.57 |
| IUPred_short_io | 0.67 (9) | 0.9 (15) | 0.42 (11) | 0.93 (15) | 0.3 (7) | 0 (1) | 50.59 (8) | 109.3 (19) | 0.41 (8) | 5073 (22) | 0.63 (6) | 0.32 (15) | 0.35 (8) | 0.73 (12) | 10.00 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RONN | 0.7 (5) | 0.82 (21) | 0.59 (7) | 0.83 (21) | 0.26 (14) | 1.93 (21) | 52.12 (7) | 54.25 (11) | 0.39 (10) | 3452 (13) | 0.58 (17) | 0.43 (21) | 0.34 (9) | 0.82 (7) | 10.00 |
| Dynamine | 0.7 (5) | 0.7 (26) | 0.71 (3) | 0.7 (26) | 0.22 (19) | 0.73 (16) | 73.93 (1) | 13.49 (4) | 0.59 (1) | 9895 (27) | 0.56 (22) | 0.55 (26) | 0.32 (12) | 0.86 (5) | 10.14 |
| IUPred_short_io15 | 0.67 (10) | 0.9 (11) | 0.38 (12) | 0.94 (13) | 0.31 (5) | 0 (1) | 48.38 (10) | 109.4 (20) | 0.39 (11) | 4566 (19) | 0.64 (5) | 0.31 (10) | 0.36 (6) | 0.53 (17) | 10.14 |
| Espritz_nmr | 0.7 (5) | 0.74 (25) | 0.66 (5) | 0.75 (25) | 0.23 (18) | 1.51 (19) | 69.33 (5) | 9.55 (1) | 0.52 (4) | 7113 (26) | 0.56 (20) | 0.51 (25) | 0.33 (11) | 0.91 (3) | 10.29 |
| IUPred_short | 0.65 (11) | 0.91 (4) | 0.34 (16) | 0.95 (7) | 0.31 (6) | 0.02 (10) | 43.1 (13) | 150.6 (23) | 0.35 (13) | 3785 (14) | 0.66 (3) | 0.3 (4) | 0.36 (7) | 0.41 (20) | 11.00 |
| IUPred_long_io | 0.65 (13) | 0.89 (16) | 0.36 (14) | 0.93 (16) | 0.26 (12) | 0 (1) | 39.86 (15) | 210.1 (25) | 0.29 (17) | 5332 (23) | 0.62 (12) | 0.33 (16) | 0.32 (14) | 0.86 (5) | 12.71 |
| IUPred_long_io15 | 0.65 (12) | 0.9 (14) | 0.32 (20) | 0.94 (14) | 0.28 (10) | 0 (1) | 37.81 (17) | 205.5 (24) | 0.28 (18) | 4789 (20) | 0.63 (10) | 0.32 (14) | 0.33 (10) | 0.68 (14) | 13.86 |
| Disembl_hotloops | 0.61 (20) | 0.93 (1) | 0.27 (21) | 0.98 (1) | 0.28 (9) | 0.11 (12) | 27.08 (21) | 83.94 (14) | 0.26 (19) | 1135 (2) | 0.69 (1) | 0.27 (1) | 0.32 (15) | 0.74 (11) | 14.00 |
| IUPred_short_io30 | 0.64 (14) | 0.9 (11) | 0.37 (13) | 0.94 (11) | 0.27 (11) | 0 (1) | 39.86 (15) | 113.1 (21) | 0.32 (14) | 4285 (16) | 0.63 (7) | 0.31 (12) | 0.32 (12) | 0.27 (22) | 14.00 |
| Espritz_disprot | 0.63 (16) | 0.9 (9) | 0.33 (17) | 0.95 (10) | 0.26 (13) | 1.86 (20) | 26.91 (22) | 98.27 (15) | 0.2 (22) | 626 (1) | 0.63 (8) | 0.31 (9) | 0.31 (16) | 0.82 (7) | 14.29 |
| Disembl_rem465 | 0.64 (15) | 0.83 (18) | 0.42 (10) | 0.85 (19) | 0.19 (20) | 0.51 (15) | 41.9 (14) | 107.7 (18) | 0.37 (12) | 3069 (11) | 0.56 (19) | 0.42 (18) | 0.27 (20) | 0.77 (9) | 15.14 |
| Disembl_coils | 0.61 (19) | 0.49 (27) | 0.73 (2) | 0.48 (27) | 0.11 (24) | 2.48 (22) | 71.89 (4) | 98.87 (16) | 0.49 (6) | 6400 (24) | 0.53 (27) | 0.71 (27) | 0.26 (21) | 0.54 (16) | 16.14 |
| Disopred_pbdat | 0.6 (22) | 0.92 (2) | 0.22 (23) | 0.97 (4) | 0.24 (15) | 10.54 (27) | 36.79 (18) | 121.5 (22) | 0.29 (16) | 2923 (9) | 0.65 (4) | 0.29 (2) | 0.29 (18) | 0.38 (21) | 17.00 |
| IUPred_long_io30 | 0.62 (17) | 0.9 (13) | 0.32 (19) | 0.94 (12) | 0.24 (16) | 0 (1) | 33.04 (19) | 215 (26) | 0.23 (20) | 4470 (18) | 0.62 (13) | 0.32 (13) | 0.29 (17) | 0.12 (24) | 17.86 |
| GlobPipe | 0.61 (20) | 0.82 (20) | 0.32 (18) | 0.86 (18) | 0.15 (21) | 2.79 (23) | 44.46 (12) | 39.26 (6) | 0.31 (15) | 2976 (10) | 0.55 (24) | 0.43 (20) | 0.24 (22) | 0.61 (15) | 18.29 |
| IUPred_long | 0.61 (18) | 0.91 (8) | 0.25 (22) | 0.95 (8) | 0.24 (17) | 0.07 (11) | 31.17 (20) | 219.9 (27) | 0.21 (21) | 4176 (15) | 0.62 (11) | 0.31 (8) | 0.29 (18) | -0.03 (26) | 18.43 |
| Anchor_io15 | 0.55 (24) | 0.91 (7) | 0.1 (25) | 0.97 (3) | 0.11 (23) | 0 (1) | 16.01 (24) | 41.46 (8) | 0.09 (24) | 1898 (5) | 0.57 (18) | 0.3 (7) | 0.16 (24) | 0.7 (13) | 20.00 |
| Anchor_io | 0.53 (25) | 0.9 (9) | 0.12 (24) | 0.96 (6) | 0.08 (26) | 0 (1) | 15.84 (25) | 39.66 (7) | 0.08 (25) | 2113 (6) | 0.55 (23) | 0.31 (10) | 0.13 (26) | 0.75 (10) | 20.71 |
| Foldindex | 0.6 (23) | 0.81 (22) | 0.36 (15) | 0.84 (20) | 0.13 (22) | 0.18 (13) | 26.74 (23) | 61.85 (12) | 0.17 (23) | 2429 (8) | 0.54 (25) | 0.43 (22) | 0.22 (23) | 0.1 (25) | 21.86 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Anchor_io30 | 0.53 (25) | 0.91 (5) | 0.1 (26) | 0.97 (5) | 0.09 (25) | 0 (1) | 14.31 (26) | 45.43 (9) | 0.07 (26) | 1723 (4) | 0.56 (21) | 0.3 (5) | 0.14 (25) | 0.16 (23) | 22.14 |
| Anchor | 0.52 (27) | 0.91 (3) | 0.05 (27) | 0.98 (2) | 0.05 (27) | 0.79 (17) | 9.54 (27) | 53.83 (10) | 0.03 (27) | 1355 (3) | 0.54 (26) | 0.29 (3) | 0.09 (27) | -0.32 (27) | 23.57 |

**Table 4. Evaluation of disorder prediction methods on TMA set.** Abbreviation of columns are Balanced Accuracy (ACC), Accuracy (ACC2), Sensitivity (SENS), Specificity (SPEC), Matthew's Correlation Coefficient (MCC), False Transmembrane percent (TM%), Per protein disorder prediction accuracy (RD), $R\chi^2$ metric (RX2), Segment Overlap (SOL), Number of false positive regions (FPREG), Area Under the Curve (AUC), Root Mean Square deviation (RMSE), Pearson Correlation Coeffitient (PCC), z-coordinate dependent distribution (Z-corr). For the definitions of the columns refer to Supplementary Document S1.

#### *3.2.2. In silico prediction methods have a similar bias in over predicting the terminal regions.*

We have investigated the distribution of predicted disordered residues along the sequences (Supplementary Figure S2 and Supplementary Figure S3). All prediction methods show the same tendency as observed in the 3D structures, i.e., that the first and last decades contain the most predicted disordered residues, whereas the middle part contains less. However, the difference between these parts differs for the various prediction methods, for example, Disembl_coils predicts almost the same amount of disordered residues in the middle and in the terminal parts, whereas Disopred_pbdat predicts more the 15 times more disordered residues for the N-terminal region than in the middle part. In Supplementary Figure S2, it can be seen that the relative frequency of predicted disorder differs in loop and terminal regions, i.e., some methods predict disordered residues in loop regions very rarely (e.g. Disprot, IUPred_long), others predict about the same frequency for loop and terminal regions (e.g. Disembl_coils, Espritz_nmr). However, these latter predictors tend to overpredict disorder in general (Supplementary Figure S3).

#### *3.2.3. Prediction methods differ strongly in z-coordinate dependent prediction accuracy*

The distribution of stem positions of predicted disordered regions along the z-axis has been calculated for each prediction method (Supplementary Figure S4), along with the correlation coefficients of these distributions with the observed distributions of residues in loop and terminal regions (Figure 5). These distributions differ strongly from each other; some have strong correlation with the distribution of observed IDRs in loop regions (e.g. Predisorder, Espritz_nmr, Disprot), some follow the tendencies of observed IDRs in terminal regions (e.g. DisEmbl_hotloops), but some of them have a maximum near the middle of the membrane (e.g. Disopred_diso, Spine-D). Although these latter methods are listed in the first part of Table 4, these should not be used for transmembrane proteins, whereas prediction methods in the former

two classes can be used to predict IDRs in loop regions of multiTM proteins and IDRs in terminal regions, respectively.

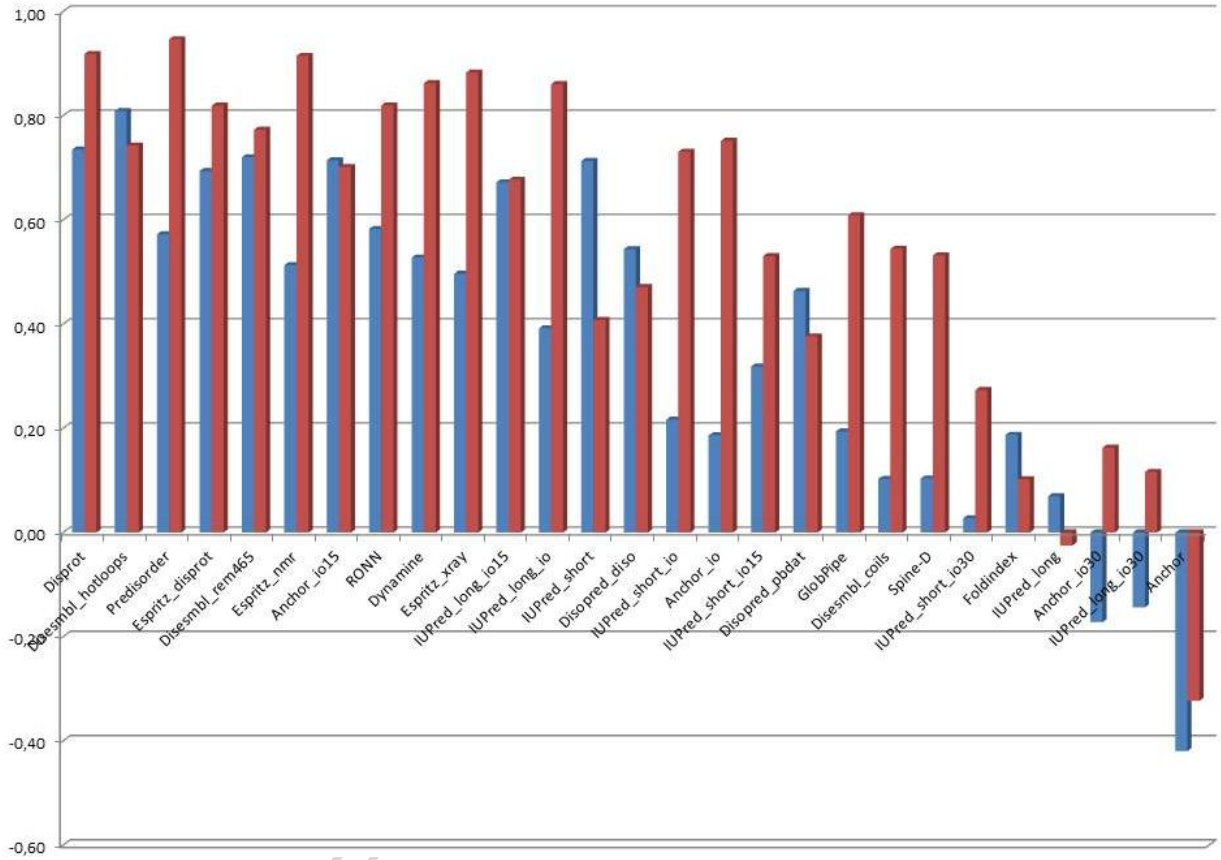



**Figure 5. Correlation coefficients of z-axis dependent distributions of predicted IDRs and IDRs observed in loop (blue) and in terminal (red) regions.**

### *3.3. The roles of disordered regions in transmembrane proteins*

The predictions provide important insight into significant structural disorder in TMPs, which is of potential functional relevance. Evidence for this has to come from detailed structure-function studies of individual proteins, for which there are many documented cases. IDRs in these locate most often at chain termini or chain ends, bind to other partners, or bind another region of the very same protein in a regulatory fashion. That is, the interaction may mediate signaling protein-protein interactions inside (such as T-cell receptor, E-cadherin) or outside (e.g. fibronectin binding protein) the cell, or it may have regulatory consequences (e.g. CFTR and Shaker channel), as outlined in the next sections.

One of the best characterized example is cystic fibrosis transmembrane conductance regulator (CFTR), a plasmamembrane chloride conductance channel. This channel, which belongs to the MRP subfamily [62] of the ATP-binding cassette (ABC) superfamily of proteins [63], is mutated in cystic fibrosis, that causes a loss of activity due to degradation in the endoplasmic reticulum

secretion pathway. The protein has two membrane-spanning domains (MSD1 and MSD2), two nucleotide-binding domains (NBD1 and NBD2), and a region between the transmembrane segments that harbors an intrinsically disordered regulatory (R) domain of about 200 amino acids in length [17,64]. This R domain has nine consensus protein kinase A (PKA) phosphorylation sites, within short regions that sample transient helical conformations. PKA phosphorylation leads to the activation of chloride conductance, via synergistic action of several short sites, by reducing their affinity to the NBD domains [19].

Fibronectin binding protein A (FnBPA) is a TMP with large extracellular regions, anchored to the membrane by a single TM helix, which mediates adherence to the host tissues by the specific recognition of host extracellular matrix (ECM). *S. aureus* FnBPA has a long, extracellular segment with a 130-amino acid repeat region, D1–D4, which is highly disordered by NMR [15]. This IDR contains transiently structured elements, which are involved in fibronectin binding, which results from induced folding to a unique extended β-strand structure (a tandem β-zipper) to two tandem Fn1 modules [65]. This binding ensures high affinity and specificity in binding, which are critical for survival and successful host invasion of bacteria.

Solution NMR, as a method for protein and nucleic acid structure determination, has had considerable impact, especially for smaller proteins or molecules with partially disordered regions that have inhibited crystallization attempts. However, NMR of membrane proteins has been difficult because of the usually large size of TMPs. However, in some cases, when structure of a membrane protein is determined both NMR spectroscopy and X-ray crystallography, they may provide complementary structural information. DsbB and DsbA are dynamic enzymes that can form intramolecular and intermolecular disulphide bonds resulting in various intermediate states. The NMR structure of their complex shows an N-terminal amphipathic helix that lies parallel to the membrane forming a so-called interfacial helix (IFH), which is disordered in the crystal structure. IFH have various structural roles, for example, they are responsible for the regulation of channel gating in both the KirBac 1.1 inward rectifying potassium channel [66] and the MscS mechanosensitive channel [67], while in photosystem I, IFHs appear to shield cofactors from the aqueous phase [68].

The Shaker-channel, a voltage-dependent potassium channel, is located in the plasma membrane of *D. melanogaster* neurons. The primary role of the channel is to conduct depolarizing potassium currents when membrane potential becomes more positive [16], by rapid transitions between inactive and active states. The channel has a special kinetic gating mechanism, in which its disordered cytoplasmic N-terminal tail moves around due to its entropic freedom and occludes the mouth of the channel. The region has transient helical propensity and assumes stable helical conformation when bound to the body of the channel, in a typical “ball-and-chain”

entropic clock mechanism. In this, the kinetics of binding and unbinding are regulated by post-translational modifications.

The Shaker channel also presents another functional trick by induced folding of a disordered tail, at the other end of the molecule. The channel has an IDR within its cytoplasmic C-terminal tail, which harbors a short recognition motif for PDZ domains. The C-terminal region of the protein is in a random coil state, and its length is critical in fine-tuning interactions of the channel, because it is specifically bound to intracellular scaffold proteins, such as the postsynaptic density 95 (PSD-95) [69]. The function of specific binding of this IDR to scaffold proteins is to promote channel clustering at unique membrane sites, which is important in proper synapse assembly and function.

The regulation of receptor activity is also the major theme in the action of a random coil C fragment of dihydropyridine receptor (DHPR), which interacts with, and activates, skeletal muscle ryanodine receptor (RyR) [18]. The two receptors are functionally connected in excitation-contraction coupling of muscle, when depolarization of the plasma membrane (where DHPR resides) causes a massive calcium influx into muscle cells, which activates RyR, furthering calcium release from the endo-(sarco) plasmic reticulum. It has been shown that activation also has a direct physical component, when the random coil C fragment within loop II–III of DHPR interacts with RyR, and activates (but also inhibits) it in a stochastic physical process.

An intracellular IDR is also critical for the function of calcium-dependent cell adhesion glycoproteins, cadherins, which mediate cell-to-cell and cell-to-ECM communication [70]. E-cadherin is a single-pass transmembrane protein with a fully disordered cytoplasmic tail of about 70 amino acids in length [71]. This IDR can bind β-catenin, which is a signaling hub protein that can bind several other partners, such as APC, Tcf and axin. These interactions are all mediated by IDRs, and take part in the complex regulation of two important developmental processes, cell–cell adhesion and the regulation of gene expression. The extended binding mode of β-catenin partners is supportive of their disorder in the unbound form, which is one of the examples of a homologous intrinsically disordered domain appearing in distinct proteins (termed catenin-binding domain, CBD) [7].

The role in homotypic interactions by IDRs has been suggested in T-cell signaling, in which homo-oligomerization mediated by cytoplasmic domains of T-cell receptor ζ-chains is a key signaling event [72]. The regions are fully disordered and have been initially claimed to bind each other while retaining structural disorder [73] (which would be a case of fuzziness [74]), although recently this mechanism has been criticized [75]. Whereas the exact binding mechanism still remains to be seen, the weak and transient interactions of receptors, mediated by IDRs, are indispensable for T-cell activation.

## 4. CONCLUSION

Here we have investigated the IDRs in TMPs using a stringent definition for extracting IDRs from structures determined by X-ray crystallography on the currently available largest experimentally dataset containing 631 TMPs with 1189 IDRs. According to our study, IDRs tend to be in the N- or C-terminal of TMPs while loop regions contain significantly less IDRs. The z-axis dependent distributions of residues in IDRs revealed a strong correlation between disordered residues and the positively charged amino acids suggesting a new function of disordered residues in TMPs near to the inner bounder of the double lipid layerThe observed correlation infers that structural disorder is related to an excess of positive residues in cytoplasmic loops, and with other disorder-promoting features in other cases (extracellular and terminal regions). Therefore, the observed structural disorder in TM proteins is not simply a manifestation of, or reason behind, the positive inside rule, rather it represents a functional modality on its own that has co-evolved with an excess of positive charges in the loop regions of multiTM proteins, but with a deficit of positive charges in extracellular regions and intracellular terminal sections of TM proteins. This infers that positive charges plus disorder inside might have one function (membrane binding and stabilization) in certain cases, but structural disorder and the lack of positive residues another function (mediating protein-protein interactions) in other cases. In other words, although positive (charged, in general) residues tend to promote structural disorder (they are disorder-promoting residues), these two features (disorder and positive charge) are uncoupled in TM proteins and have been under different selection pressures.

Generating a sizeable and stringent database of structural disorder in membrane proteins has also allowed us to critically compare the performance of disorder predictors on TMPs. We observe that most in silico methods overpredict IDRs in TMP, especially in the N- and C-terminal regions, and show big difference in Z-coordinate dependent distributions in opposite to the usual metrics used for evaluating these predictors. Therefore, we give recommendations as to the preferences of using predictors for terminal and loop regions. By collecting individual examples from the literature, we suggest that IDRs in loop region may have the role to stabilize the protein via positioning positively charged amino acids to lipid head groups, and also quite often to mediate allosteric regulatory interactions with the very same TMP or adjacent proteins. IDRs in terminal regions tend to have different roles. Most often, they mediate interactions with other proteins, imparting localization effect that can act on the whole cell (e.g. anchoring pathogens) or on the protein (clustering receptors, or recruiting signaling partners). In all, the examples presented for IDRs in TMPs are involved in signaling, protein-protein interactions, and regulation, which, given the central position of TMPs in signal transduction, suggests that they represent the first control point on signaling by the cell. Given the great number of TMPs, their central role in signaling and the abundance of structural disorder in them, targeted studies on the

most important examples are definitely a worthy investment to gain deeper insight into cell regulation.

## 5. ACKNOWLEDGEMENT

This work was supported by the Hungarian Scientific Research Fund (K104586, http://www.otka.hu). GET was supported by "Momentum" Program of the Hungarian Academy of Sciences, PT received the Odysseus grant G.0029.12 from Research Foundation Flanders (FWO).

## 6. REFERENCES

[1] L.M. Iakoucheva, C.J. Brown, J.D. Lawson, Z. Obradović, A.K. Dunker, Intrinsic disorder in cell-signaling and cancer-associated proteins., J. Mol. Biol. 323 (2002) 573–84.

[2] Y. Minezaki, K. Homma, K. Nishikawa, Intrinsically disordered regions of human plasma membrane proteins preferentially occur in the cytoplasmic segment, J. Mol. Biol. 368 (2007) 902–913. doi:10.1016/j.jmb.2007.02.033.

[3] I. Kotta-Loizou, G.N. Tsaousis, S.J. Hamodrakas, Analysis of Molecular Recognition Features (MoRFs) in membrane proteins., Biochim. Biophys. Acta. 1834 (2013) 798–807. doi:10.1016/j.bbapap.2013.01.006.

[4] R. van der Lee, M. Buljan, B. Lang, R.J. Weatheritt, G.W. Daughdrill, A.K. Dunker, et al., Classification of Intrinsically Disordered Regions and Proteins., Chem. Rev. (2014). doi:10.1021/cr400525m.

[5] M. Varadi, S. Kosol, P. Lebrun, E. Valentini, M. Blackledge, A.K. Dunker, et al., pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins., Nucleic Acids Res. 42 (2014) D326–35. doi:10.1093/nar/gkt960.

[6] M. Fuxreiter, P. Tompa, I. Simon, Local structural disorder imparts plasticity on linear motifs., Bioinformatics. 23 (2007) 950–6. doi:10.1093/bioinformatics/btm035.

[7] P. Tompa, M. Fuxreiter, C.J. Oldfield, I. Simon, A.K. Dunker, V.N. Uversky, Close encounters of the third kind: disordered domains and the interactions of proteins., Bioessays. 31 (2009) 328–35. doi:10.1002/bies.200800151.

[8] P.E. Wright, H.J. Dyson, Linking folding and binding., Curr. Opin. Struct. Biol. 19 (2009) 31–8. doi:10.1016/j.sbi.2008.12.003.

[9] H. Xie, S. Vucetic, L.M. Iakoucheva, C.J. Oldfield, A.K. Dunker, Z. Obradovic, et al., Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins., J. Proteome Res. 6 (2007) 1917–32. doi:10.1021/pr060394e.

[10] H. Xie, S. Vucetic, L.M. Iakoucheva, C.J. Oldfield, A.K. Dunker, V.N. Uversky, et al., Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions., J. Proteome Res. 6 (2007) 1882–98. doi:10.1021/pr060392u.

[11] P. V Burra, L. Kalmar, P. Tompa, Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes., PLoS One. 5 (2010) e12069. doi:10.1371/journal.pone.0012069.

[12] R. Pancsa, P. Tompa, Structural disorder in eukaryotes., PLoS One. 7 (2012) e34687. doi:10.1371/journal.pone.0034687.

[13] T. Le Gall, P.R. Romero, M.S. Cortese, V.N. Uversky, A.K. Dunker, Intrinsic disorder in the Protein Data Bank., J. Biomol. Struct. Dyn. 24 (2007) 325–42. doi:10.1080/07391102.2007.10507123.

[14] P. Tompa, Multisteric regulation by structural disorder in modular signaling proteins: an extension of the concept of allostery., Chem. Rev. 114 (2014) 6715–32. doi:10.1021/cr4005082.

[15] C.J. Penkett, C. Redfield, I. Dodd, J. Hubbard, D.L. McBay, D.E. Mossakowska, et al., NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein., J. Mol. Biol. 274 (1997) 152–9. doi:10.1006/jmbi.1997.1369.

[16] T. Hoshi, W.N. Zagotta, R.W. Aldrich, Biophysical and molecular mechanisms of Shaker potassium channel inactivation., Science. 250 (1990) 533–8.

[17] T. Hegedus, A.W.R. Serohijos, N. V Dokholyan, L. He, J.R. Riordan, Computational studies reveal phosphorylation-dependent changes in the unstructured R domain of CFTR., J. Mol. Biol. 378 (2008) 1052–63. doi:10.1016/j.jmb.2008.03.033.

[18] C.S. Haarmann, D. Green, M.G. Casarotto, D.R. Laver, A.F. Dulhunty, The random-coil "C" fragment of the dihydropyridine receptor II-III loop can activate or inhibit native

skeletal ryanodine receptors., Biochem. J. 372 (2003) 305–16. doi:10.1042/BJ20021763.

[19] J.M.R. Baker, R.P. Hudson, V. Kanelis, W.-Y. Choy, P.H. Thibodeau, P.J. Thomas, et al., CFTR regulatory region interacts with NBD1 predominantly via multiple transient helices., Nat. Struct. Mol. Biol. 14 (2007) 738–45. doi:10.1038/nsmb1278.

[20] K. Peng, P. Radivojac, S. Vucetic, A.K. Dunker, Z. Obradovic, Length-dependent prediction of protein intrinsic disorder., BMC Bioinformatics. 7 (2006) 208. doi:10.1186/1471-2105-7-208.

[21] Z. Dosztányi, V. Csizmók, P. Tompa, I. Simon, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins, J. Mol. Biol. 347 (2005) 827–839. doi:10.1016/j.jmb.2005.01.071.

[22] R. Linding, R.B. Russell, V. Neduva, T.J. Gibson, GlobPlot: Exploring protein sequences for globularity and disorder., Nucleic Acids Res. 31 (2003) 3701–8.

[23] E.E. Pryor, M.C. Wiener, A critical evaluation of in silico methods for detection of membrane protein intrinsic disorder., Biophys. J. 106 (2014) 1638–49. doi:10.1016/j.bpj.2014.02.025.

[24] E. Potenza, T.D. Domenico, I. Walsh, S.C.E. Tosatto, MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins, Nucleic Acids Res. (2014) gku982–. doi:10.1093/nar/gku982.

[25] J.Y. Yang, M.Q. Yang, A.K. Dunker, Y. Deng, X. Huang, Investigation of transmembrane proteins using a computational approach, BMC Genomics. 9 Suppl 1 (2008) S7. doi:10.1186/1471-2164-9-S1-S7.

[26] A. De Biasio, C. Guarnaccia, M. Popovic, V.N. Uversky, A. Pintar, S. Pongor, Prevalence of intrinsic disorder in the intracellular region of human single-pass type I proteins: the case of the notch ligand Delta-4., J. Proteome Res. 7 (2008) 2496–506. doi:10.1021/pr800063u.

[27] B. Xue, L. Li, S.O. Meroueh, V.N. Uversky, A.K. Dunker, Analysis of structured and intrinsically disordered regions of transmembrane proteins, Mol. Biosyst. 5 (2009) 1688–1702. doi:10.1039/B905913J.

[28] A. De Biasio, C. Guarnaccia, M. Popovic, V.N. Uversky, A. Pintar, S. Pongor, Prevalence of intrinsic disorder in the intracellular region of human single-pass type I proteins: the case of the notch ligand Delta-4, J. Proteome Res. 7 (2008) 2496–2506. doi:10.1021/pr800063u.

[29] A. Venkatakrishnan, T. Flock, D.E. Prado, M.E. Oates, J. Gough, M. Madan Babu, Structured and disordered facets of the GPCR fold., Curr. Opin. Struct. Biol. 27C (2014) 129–137. doi:10.1016/j.sbi.2014.08.002.

[30] I. Stavropoulos, N. Khaldi, N.E. Davey, K. O'Brien, F. Martin, D.C. Shields, Protein disorder and short conserved motifs in disordered regions are enriched near the cytoplasmic side of single-pass transmembrane proteins, PLoS One. 7 (2012) e44389. doi:10.1371/journal.pone.0044389.

[31] D. Kozma, I. Simon, G.E. Tusnády, PDBTM: Protein Data Bank of transmembrane proteins after 8 years, Nucleic Acids Res. 41 (2013) D524–529. doi:10.1093/nar/gks1169.

[32] G. Tusnady, L. Kalmar, I. Simon, TOPDB: topology data bank of transmembrane proteins., Nucleic Acids Res. 36 (2008) D234–9. doi:10.1093/nar/gkm751.

[33] L. Dobson, T. Lango, I. Remenyi, G.E. Tusnady, Expediting topology data gathering for the TOPDB database, Nucleic Acids Res. 43 (2014) D283–D289. doi:10.1093/nar/gku1119.

[34] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences., Bioinformatics. 22 (2006) 1658–9. doi:10.1093/bioinformatics/btl158.

[35] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics. 26 (2010) 680–2. doi:10.1093/bioinformatics/btq003.

[36] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data., Bioinformatics. 28 (2012) 3150–2. doi:10.1093/bioinformatics/bts565.

[37] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410. doi:10.1016/S0022-2836(05)80360-2.

[38] J.D. Thompson, T.J. Gibson, D.G. Higgins, Multiple sequence alignment using ClustalW and ClustalX., Curr. Protoc. Bioinformatics. Chapter 2 (2002) Unit 2.3. doi:10.1002/0471250953.bi0203s00.

[39] E. Cilia, R. Pancsa, P. Tompa, T. Lenaerts, W.F. Vranken, From protein sequence to dynamics and disorder with DynaMine., Nat. Commun. 4 (2013) 2741. doi:10.1038/ncomms3741.

[40] E. Cilia, R. Pancsa, P. Tompa, T. Lenaerts, W.F. Vranken, The DynaMine webserver: predicting protein dynamics from sequence., Nucleic Acids Res. 42 (2014) W264–70. doi:10.1093/nar/gku270.

[41] B. Mészáros, P. Tompa, I. Simon, Z. Dosztányi, Molecular principles of the interactions of disordered proteins, J. Mol. Biol. 372 (2007) 549–561. doi:10.1016/j.jmb.2007.07.004.

[42] Z. Dosztányi, B. Mészáros, I. Simon, ANCHOR: web server for predicting protein binding regions in disordered proteins, Bioinformatics. 25 (2009) 2745–2746. doi:10.1093/bioinformatics/btp518.

[43] Z. Dosztányi, V. Csizmok, P. Tompa, I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, Bioinformatics. 21 (2005) 3433–3434. doi:10.1093/bioinformatics/bti541.

[44] B. Mészáros, I. Simon, Z. Dosztányi, Prediction of protein binding regions in disordered proteins, PLoS Comput. Biol. 5 (2009) e1000376. doi:10.1371/journal.pcbi.1000376.

[45] R. Linding, L.J. Jensen, F. Diella, P. Bork, T.J. Gibson, R.B. Russell, Protein disorder prediction: implications for structural proteomics., Structure. 11 (2003) 1453–9.

[46] J.J. Ward, L.J. McGuffin, K. Bryson, B.F. Buxton, D.T. Jones, The DISOPRED server for the prediction of protein disorder., Bioinformatics. 20 (2004) 2138–9. doi:10.1093/bioinformatics/bth195.

[47] I. Walsh, A.J.M. Martin, T. Di Domenico, S.C.E. Tosatto, ESpritz: accurate and fast prediction of protein disorder., Bioinformatics. 28 (2012) 503–9. doi:10.1093/bioinformatics/btr682.

[48] J. Prilusky, C.E. Felder, T. Zeev-Ben-Mordehai, E.H. Rydberg, O. Man, J.S. Beckmann, et al., FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded., Bioinformatics. 21 (2005) 3435–8. doi:10.1093/bioinformatics/bti537.

[49] X. Deng, J. Eickholt, J. Cheng, PreDisorder: ab initio sequence-based prediction of protein disordered regions., BMC Bioinformatics. 10 (2009) 436. doi:10.1186/1471-2105-10-436.

[50] Z.R. Yang, R. Thomson, P. McNeil, R.M. Esnouf, RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins., Bioinformatics. 21 (2005) 3369–76. doi:10.1093/bioinformatics/bti534.

[51] T. Zhang, E. Faraggi, B. Xue, A.K. Dunker, V.N. Uversky, Y. Zhou, SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method., J. Biomol. Struct. Dyn. 29 (2012) 799–813. doi:10.1080/073911012010525022.

[52] J.M.G. Izarzugaza, O. Graña, M.L. Tress, A. Valencia, N.D. Clarke, Assessment of intramolecular contact predictions for CASP7, Proteins. 69 Suppl 8 (2007) 152–158. doi:10.1002/prot.21637.

[53] O. Noivirt-Brik, J. Prilusky, J.L. Sussman, Assessment of disorder predictions in CASP8., Proteins. 77 Suppl 9 (2009) 210–6. doi:10.1002/prot.22586.

[54] B. Monastyrskyy, K. Fidelis, J. Moult, A. Tramontano, A. Kryshtafovych, Evaluation of disorder predictions in CASP9., Proteins. 79 Suppl 1 (2011) 107–18. doi:10.1002/prot.23161.

[55] B. Monastyrskyy, A. Kryshtafovych, J. Moult, A. Tramontano, K. Fidelis, Assessment of protein disorder region predictions in CASP10., Proteins. 82 Suppl 2 (2014) 127–37. doi:10.1002/prot.24391.

[56] I. Walsh, M. Giollo, T. Di Domenico, C. Ferrari, O. Zimmermann, S.C.E. Tosatto, Comprehensive large-scale assessment of intrinsic protein disorder., Bioinformatics. (2014). doi:10.1093/bioinformatics/btu625.

[57] L. Dobson, I. Reményi, G.E. Tusnády, The human transmembrane proteome., Biol. Direct. 10 (2015) 31. doi:10.1186/s13062-015-0061-x.

[58] G. von Heijne, The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology, EMBO J. 5 (1986) 3021–3027.

[59] T. Kouyama, R. Fujii, S. Kanada, T. Nakanishi, S.K. Chan, M. Murakami, Structure of archaerhodopsin-2 at 1.8 Å resolution., Acta Crystallogr. D. Biol. Crystallogr. 70 (2014) 2692–701. doi:10.1107/S1399004714017313.

[60] A. Campen, R.M. Williams, C.J. Brown, J. Meng, V.N. Uversky, A.K. Dunker, TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder., Protein Pept. Lett. 15 (2008) 956–63.

[61] Z. Dosztányi, B. Mészáros, I. Simon, Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins, Brief. Bioinform. 11 (2010) 225–243. doi:10.1093/bib/bbp061.

[62] G. Tusnády, E. Bakos, A. Váradi, B. Sarkadi, Membrane topology distinguishes a subfamily of the ATP-binding cassette (ABC) transporters, FEBS Lett. 402 (1997) 1–3. doi:9013845.

[63] J.R. Riordan, J.M. Rommens, B. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, et al., Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA., Science. 245 (1989) 1066–73.

[64] L.S. Ostedgaard, O. Baldursson, D.W. Vermeer, M.J. Welsh, A.D. Robertson, A functional R domain from cystic fibrosis transmembrane conductance regulator is

predominantly unstructured in solution., Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 5657–62. doi:10.1073/pnas.100588797.

[65] U. Schwarz-Linek, J.M. Werner, A.R. Pickford, S. Gurusiddappa, J.H. Kim, E.S. Pilka, et al., Pathogenic bacteria attach to human fibronectin through a tandem beta-zipper., Nature. 423 (2003) 177–81. doi:10.1038/nature01589.

[66] D.A. Doyle, Structural themes in ion channels., Eur. Biophys. J. 33 (2004) 175–9. doi:10.1007/s00249-003-0382-z.

[67] R.B. Bass, K.P. Locher, E. Borths, Y. Poon, P. Strop, A. Lee, et al., The structures of BtuCD and MscS and their implications for transporter and channel function., FEBS Lett. 555 (2003) 111–5.

[68] P. Jordan, P. Fromme, H.T. Witt, O. Klukas, W. Saenger, N. Krauss, Three-dimensional structure of cyanobacterial photosystem I at 2.5 A resolution., Nature. 411 (2001) 909–17. doi:10.1038/35082000.

[69] E. Magidovich, I. Orr, D. Fass, U. Abdu, O. Yifrach, Intrinsic disorder in the C-terminal domain of the Shaker voltage-activated K+ channel modulates its interaction with scaffold proteins., Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 13022–7. doi:10.1073/pnas.0704059104.

[70] J.M. Gooding, K.L. Yap, M. Ikura, The cadherin-catenin complex as a focal point of cell adhesion and signalling: new insights from three-dimensional structures., Bioessays. 26 (2004) 497–511. doi:10.1002/bies.20033.

[71] A.H. Huber, D.B. Stewart, D. V Laurents, W.J. Nelson, W.I. Weis, The cadherin cytoplasmic domain is unstructured in the absence of beta-catenin. A possible mechanism for regulating cadherin turnover., J. Biol. Chem. 276 (2001) 12301–9. doi:10.1074/jbc.M010377200.

[72] A. Sigalov, D. Aivazian, L. Stern, Homooligomerization of the cytoplasmic domain of the T cell receptor zeta chain and of other proteins containing the immunoreceptor tyrosine-based activation motif., Biochemistry. 43 (2004) 2049–61. doi:10.1021/bi035900h.

[73] A.B. Sigalov, Unusual biophysics of immune signaling-related intrinsically disordered proteins., Self. Nonself. 1 (2010) 271–281. doi:10.4161/self.1.4.13641.

[74] P. Tompa, M. Fuxreiter, Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions., Trends Biochem. Sci. 33 (2008) 2–8. doi:10.1016/j.tibs.2007.10.003.

[75] A. Nourse, T. Mittag, The cytoplasmic domain of the T-cell receptor zeta subunit does not form disordered dimers., J. Mol. Biol. 426 (2014) 62–70. doi:10.1016/j.jmb.2013.09.036.

## TABLES

**Table 1.** Disorder prediction methods evaluated on transmembrane protein sequences.

**Table 2.** Joint distributions of IDRs in TMAB set (1TM proteins). + indicates that the specified region contains IDR, - indicates the whole region is ordered.

**Table 3.** Joint distributions of IDRs in TMAP set (MultiTM proteins). + indicates that the specified region contains IDR, - indicates the whole region is ordered.

**Table 4.** Abbreviation of columns are Balanced Accuracy (ACC), Accuracy (ACC2), Sensitivity (SENS), Specificity (SPEC), Matthew's Correlation Coefficient (MCC), False Transmembrane percent (TM%), Per protein disorder prediction accuracy (RD), $R\chi^2$ metric (RX2), Segment Overlap (SOL), Number of false positive regions (FPREG), Area Under the Curve (AUC), Root Mean Square deviation (RMSE), Pearson Correlation Coeffitient (PCC), z-coordinate dependent distribution (Z-corr). Evaluation of disorder prediction methods on TMA set.

## FIGURE LEGENDS

**Figure 1. Definition of IDRs and their length distribution. A:** The definition of IDRs. TOPDB line: topology defined in TOPDB database; PDB lines: 3D structures of identical or homologous proteins; IDR line: final definition of IDRs. Blue: outside, Orange: transmembrane region, Red: inside, striped box: residues in 3D structure with no atom coordinates, thin line: regions wich are not covered in any PDB files, thick magenta: regions with determined 3D structure. **B:** The distribution of disordered region lengths.

**Figure 2. The distribution of IDRs in the sequence.** X-axis shows the coverage of the sequences. Blue: proportion of disordered residues before the first transmembrane segment or

after the last transmembrane segment (terminal regions). Red: proportion of disordered residues between two transmembrane segments (loop regions).

**Figure 3.** Distribution of IDRs in terminals and loop regions. A: Number of proteins containing IDRs in terminal and loop regions; B: Number of regions in the indicated parts of the transmembrane proteins; C: Relative frequencies of regions containing IDRs in the specified regions (i.e. the number of IDR containing regions divided by the number of regions in the certain part of the protein). D: Relative frequencies of residues in IDRs (i.e.the number of residues in IDRs divided by the number of all residues in the specified regions). Red: inside, Blue: outside.

**Figure 4.** Distribution of positive charged amino acids (blue line) and the stem position of IDR to be in terminal (red line) or loop (green line) region along the z-axis, normalized by the number of all amino acids having the given properties. Negative z-coordinate means outside, positive one means inside regions.

**Figure 5.** Correlation coefficients of z-axis dependent distributions of predicted IDRs and IDRs observed in loop (blue) and in terminal (red) regions.

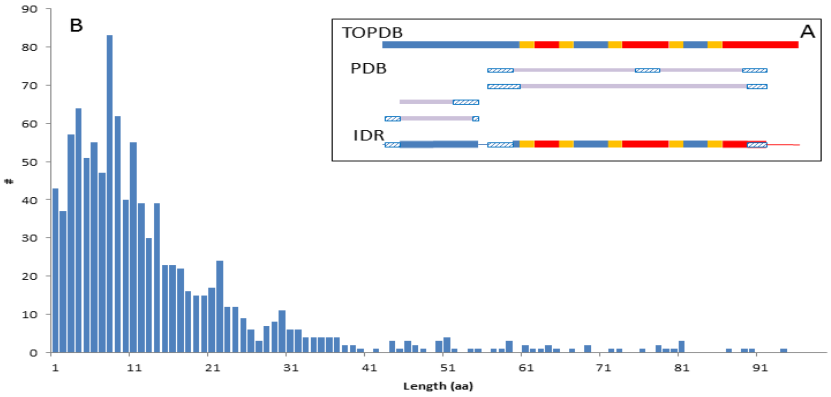



Figure 1

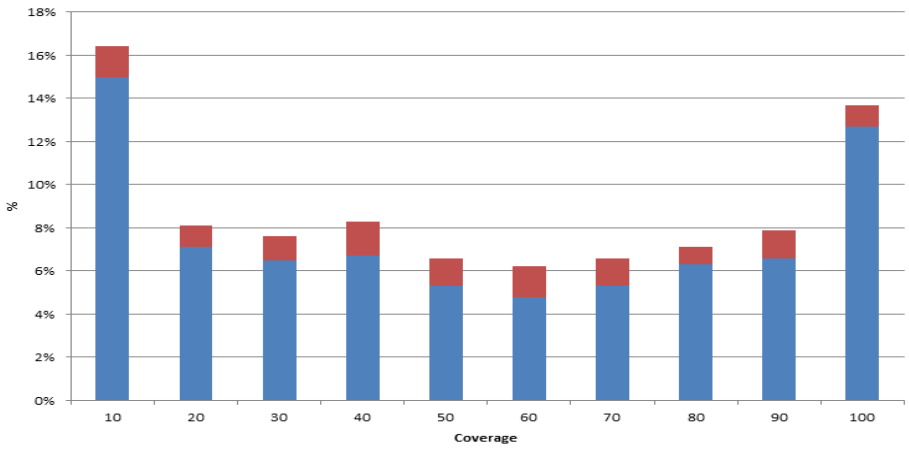



Figure 2

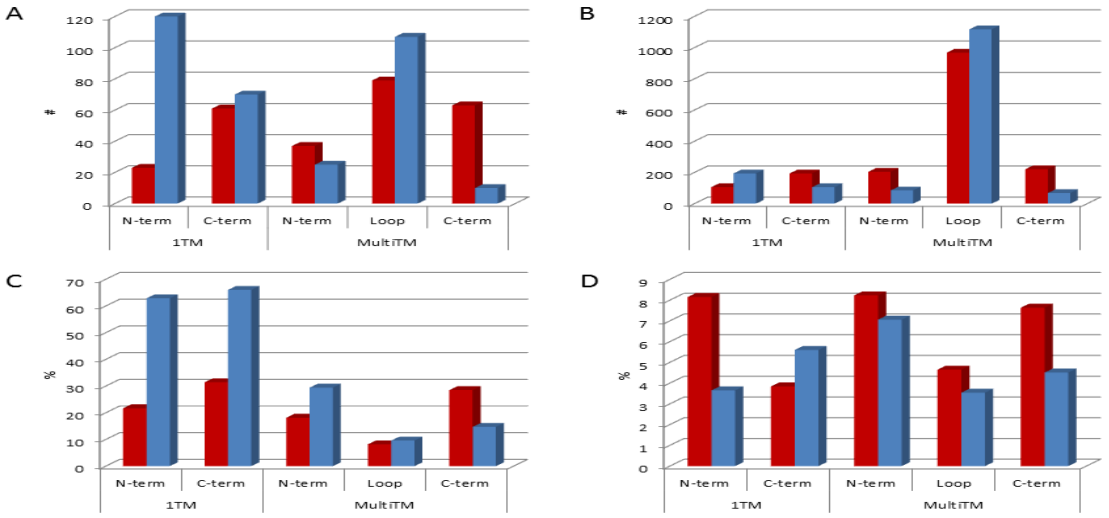



Figure 3

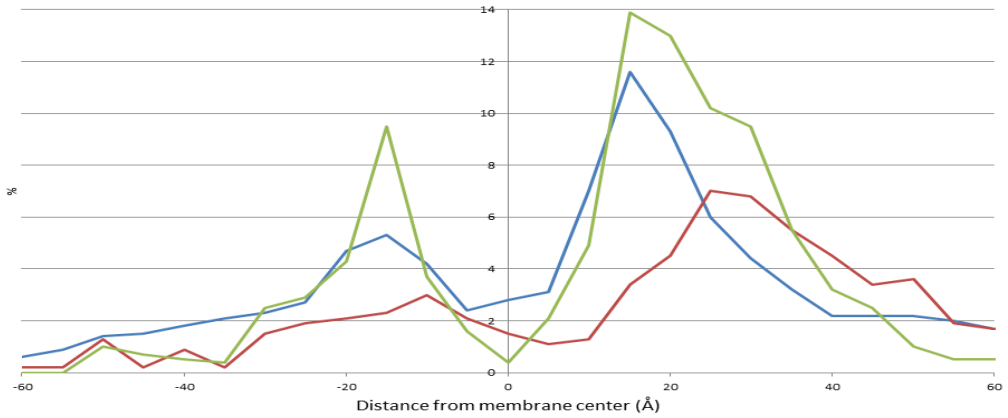



Figure 4

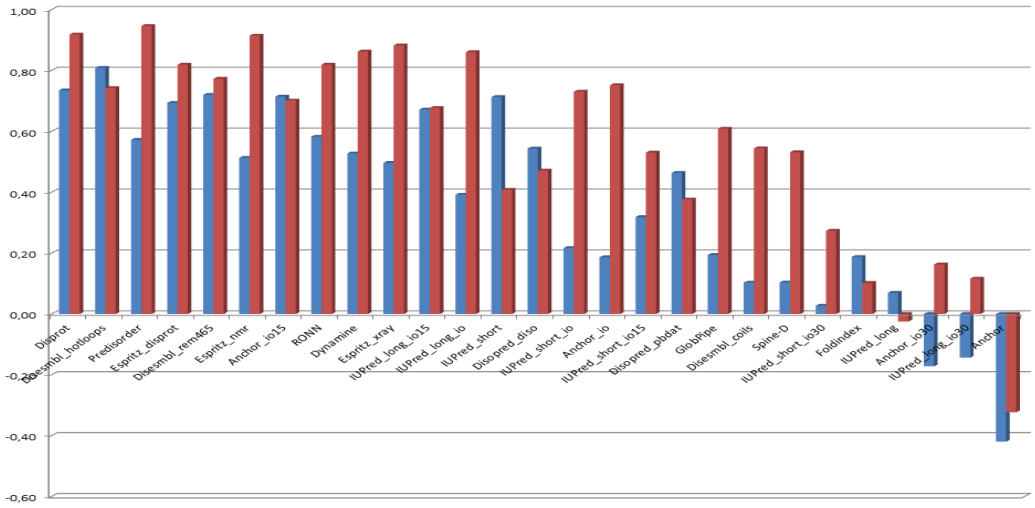
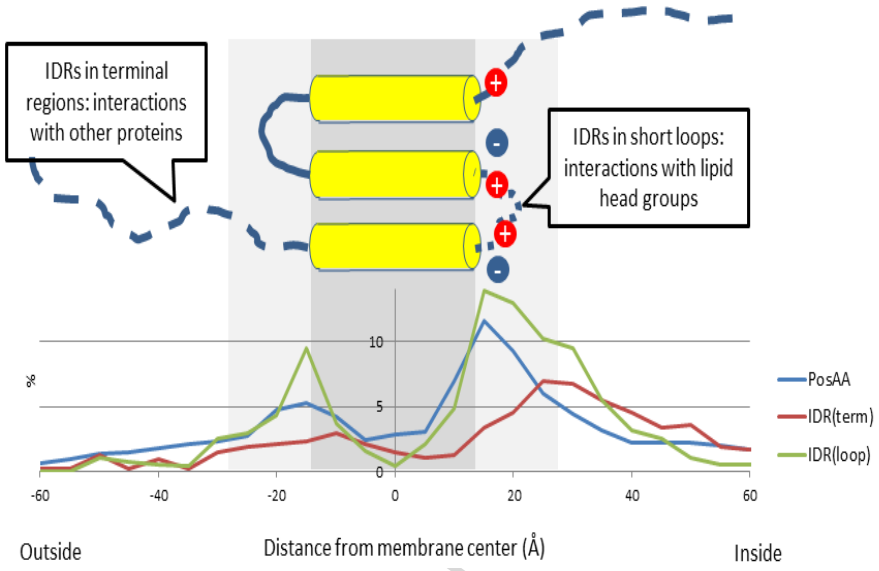



Figure 5




## Graphical abstract

**HIGHLIGHTS:**

- IDRs tend to be in the N- or C-terminal regions of transmembrane proteins
- There is a strong correlation between disordered residues in loop regions and the positively charged amino acids
- IDRs in loop regions may stabilize the protein via positioning positive amino acids to lipid head groups
- In silico methods overpredict IDRs in TMPs, especially in the N- and C-terminal regions
- IDRs in terminal regions may have different roles by mediating interactions with other proteins