Original paper

PolarProtPred: Predicting apical and basolateral localization of transmembrane proteins using putative short linear motifs and deep learning

Laszlo Dobson[1,2], András Zeke[1], Gábor E. Tusnády[1,*]

[1]Institute of Enzymology, Research Centre for Natural Sciences, Magyar Tudósok Körútja 2, 1117 Budapest, Hungary; [2] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Cell polarity refers to the asymmetric organization of cellular components in various cells. Epithelial cells are the best-known examples of polarized cells, featuring apical and basolateral membrane domains. Mounting evidence suggests that short linear motifs play a major role in protein trafficking to these domains, although the exact rules governing them are still elusive.

Results: In this study we prepared neural networks that capture recurrent patterns to classify transmembrane proteins localizing into apical and basolateral membranes. Asymmetric expression of drug transporters results in vectorial drug transport, governing the pharmacokinetics of numerous substances, yet the data on how proteins are sorted in epithelial cells is very scattered. The provided method may offer help to experimentalists to identify or better characterize molecular networks regulating the distribution of transporters or surface receptors (including viral entry receptors like that of COVID-19).

Availability: The prediction server PolarProtPred is available at http://polarprotpred.ttk.hu.

Contact: tusnady.gabor@ttk.hu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Polarity is an essential feature of many cells, especially in differentiated, multicellular organisms. In these cells, macromolecular complexes (e.g., plasma membrane proteins, cytoskeletal structures) are often organized asymmetrically. Many mammalian cell types exhibit a certain level of polarity, such as neurons, migratory cells, epithelial cells, and more. Epithelial cells possess a highly organized architecture establishing an apical-basolateral axis separated by tight junctions to maintain physiological barriers (Bryant and Mostov, 2008), for example they maintain ion homeostasis in the eccrine glands and ducts (Hanukoglu et al., 2017) or play a role in nutrient up-take (Inukai et al., 2004). Many viruses exploit epithelial cells to invade their host: Influenza A Virus targets M2 apical protein for virion entry (Wohlgemuth et al. 2018). Coronaviruses also aim for apical entry, while virus release may occur on both side: apical release promotes horizontal infec-tion to nearby cells, upon basolateral exit the virus eventually reach the bloodstream and gets circulated in the body (Cong and Ren, 2014). Although we have an increasingly detailed knowledge of the main determinants of apical and basolateral polarity networks, the exact composition of these membranes is still elusive for most tissues (Riga et al., 2020). Elements (proteins) required for the proper transport greatly differ on the apical and basolateral part of the membrane. In turn, polarity also relies on the correct sorting of these molecules to particular locations. In many cases, trafficking of these proteins from the Trans-Golgi Network to the plasma membrane does not occur in a single step, but rather via an indirect route through endosomal pathways (Laird and Spiess, 2000). The journey to the cell surface proteins is tightly regulated via post-translational modifications and transient interactions with other molecules (Stoops and Caplan, 2014).

Many of these sorting processes are mediated via Short linear motifs (SLiMs), flexible protein segments composed of a restricted number of residues (typically between 3-10), that usually bind to ordered protein domains via coupled folding and binding. These properties enable them to bind to a diverse range of partners with low micromolar affinity and establish transient interactions (Van Roey et al., 2014). Besides mediating protein interactions, they also provide sites for post-translational notifications or proteolytic cleavage sites (Davey et al., 2012). Recent decades pro-vided a handful of evidence of motifs playing a crucial role in the trafficking of proteins to polarized membranes.

Trafficking to the basolateral and to the apical membranes includes multiple pathways (Farr et al., 2009; Weisz and Rodriguez-Boulan, 2009) and often includes cargo sorting (Di Martino et al., 2019). The basolateral targeting of transmembrane proteins (TMPs) may rely on cytosolic tyrosine-based (Le Bivic et al., 1991), mono- and dileucine motifs (Hunziker and Fumey, 1994; Martín et al., 2019). Localization may also be proteolytic processing and glycosylation dependent (Evdo-kimov et al., 2016). In contrast, the apical targeting can occur in the absence of basolateral signal and can also involve rafts (Simons and Ikonen, 1997). Both N- and O-glycans can play a role in apical sorting (Urquhart et al., 2005; Yeaman et al., 1997), as well as interaction between trans-membrane (TM) regions and their surroundings (Dunbar et al., 2000). Apical trafficking is sometimes functionally redundant or combinatorial, with each protein possessing a set of motifs, individually capable of proper targeting (Stoops and Caplan, 2014). Piggybacking on other partner proteins is also widespread, thus no single targeting system exists. The divided nature of apical membranes adds further complication to trafficking (Garcia-Gonzalo and Reiter, 2012).

In theory, having so many linear motifs involved in sorting would imply that predicting the localization of a given protein would be a straightforward task. However, this is not the case: A major problem with the above-mentioned SLiMs is that they are often members of large, multifunctional motif families. Dileucine motifs, PDZ ligands, or tyrosine-based motifs are generic ligands for many proteins carrying the corresponding binding domains, with diverse roles, while polarized sorting is only driven by a subset of these motifs. Therefore, detecting a very generic motif on a protein does not assure that it automatically localizes to the corresponding membrane domain. One often has to re-define motifs or establish new SLiMs in order to capture those actually governing sorting.

Although now we have a decent understanding of which regions/residues/modifications play a critical role in individual proteins to reach their destination, their complexity makes it hard to apply simple, "hard" rules to them. Hence, we propose a novel approach to classify alpha-helical TMPs in polarized cells, based on their topology and putative SLiMs driving their localization. Previously we collected hundreds of proteins with reliable experimental evidence of their destination in PolarProtDb (Zeke et al., 2020). In this work we use computational biology approaches to predict the localization of TMPs using protein sequence alone.

2 Methods

2.1 Dataset

As an initial step, we downloaded the reference proteome of Pan troglodytes, *Gorilla gorilla, Pon-go abelii, Macaca mulatta, Felis catus, Canis familiaris, Equus caballus, Ovis aries, Bos taurus, Oryctolagus cuniculus, Callithrix jacchus, Mus musculus, Rattus norvegicus, Sus scrofa* and gen-erated orthologous groups based on the OMA database (Supplementary Table 1). We used the PolarProtDb (Zeke et al., 2020) to define apical and basolateral membrane proteins. Furthermore, we collected plasma membrane proteins from the RBCC database (Hegedűs et al., 2015) and our previous experimental pipelines (Langó et al., 2017, 2020; Müller et al., 2019). We collected proteins localizing to other membranes (mitochondrial membrane, endoplasmic reticulum, lysosomal membrane etc.) from SwissProt (reviewed, evidence level: protein) (UniProt Consortium, 2019).

We aligned the sequences with ClustalOmega (Sievers et al., 2011). Those alignments were dis-carded, where we found discrepancies in the aligned TM topologies predicted by CCTOP (Dobson et al., 2015a). Next, we divided the collected proteins into training and testing subsets. I) The "Training dataset" contains 1011, 759, 2037, 5464 apical, basolateral, generic outer plasma membrane and other membrane proteins, respectively (Supplementary Table 2). II) The "Independent test set" contains proteins from 40 orthologous group (10 from each localization group), including 70, 67, 60, 56 proteins (Supplementary Table 3). The "training&validation" and the "in-dependent test" datasets do not contain any sequence that share higher than 40% sequence identity.

For training and validation, the "Training&validation dataset" was used in a manner, that each binary Neural Network received roughly equal amount of positive and negative examples. The least populated localization group (Basolateral) was used as a standard: the number of proteins selected for each class is equal to 75% of the basolateral protein class size. As a next step, 80% of the examples were selected for training, 20% for validation (Supplementary Table 2). For the final predictors, roughly even number of sequences were selected for each label, using the same ratio (80-20) for training and validation (Supplementary Table 2). Furthermore, no sequence from the same cluster (Supplementary Table 1) was selected for training and validation of individual Neural Networks.

For testing an independent test set was used, that was not used in any manner during the training (Supplementary Table 3). To avoid information leakage, we removed all BLAST hits occurred dur-ing training (Supplementary Table 4) from the background database (SwissProt) when testing our method.

A third dataset (the "Human AB dataset") was also created, that contains all apical and basolateral proteins from Homo sapiens (Supplementary Table 5).

2.2 Putative linear motifs

Teiresias (Rigoutsos and Floratos, 1998) was used to detect patterns in the sequence. We only accepted those occurrences, where the detected pattern fell into disordered non-membrane regions. We also random shuffled sequences ten times, and accepted motifs when the average plus three times the standard deviation was lower compared to real hits.

2.3 Clustering of pre-aligned linear motifs

We collected motifs from the literature or using computational methods (Supplementary Table 6,7). Then we aligned SLiMs belonging to the same class and built a distance matrix based using following equations. Linear motifs defined using regular expressions, where brackets define (multiple) amino acid(s) permitted at a given position.

First, we calculated the pairwise distance of any two amino acid, using all linear motifs from ELM database (Kumar et al., 2019). We considered the number of each residue in total, or as a given pair within brackets (allowing for multiplicities but ignoring wildcards) across all valid ELM motif definitions. For those very few pairs that do not occur across the entire ELM, we set a pseudocount to 1. For any i,j residue pair, we defined Ei,j as:

$$E_{i,j} = \log(\frac{\sum_{k=1}^{n}(S_{k,i} + C_{k,i}) * (S_{k,j} + C_{k,j})}{\sum_{k=1}^{n} O_{k,i,j}})$$

where *n* is the number of all all brackets across all motifs, and

$$S_{k,i} = \begin{cases} 1, & if\ bracket\ k\ contains\ one\ amino\ acid, and\ that\ is\ amino\ acid_i \\ 0, & else \end{cases}$$
$$C_{k,i} = \begin{cases} 1, & if\ bracket\ k\ has\ more\ element\ including\ amino\ acid_i \\ 0, & else \end{cases}$$
$$O_{k,i,j} = \begin{cases} 1, & if\ bracket\ k\ includes\ both\ amino\ acid_i\ and\ \ amino\ acid_j \\ 0, & else \end{cases}$$

Next, we calculated the Information content for each position (I_p) in the alignment:

$$I_p = \sum_{i=1}^{20} \frac{aa_i}{M} * \log(\frac{aa_i}{M})$$

where *M* is the number of all linear motifs, aa_i is the occurrence of a given amino acid at a given position across all instances of the alignment.

We define $T_{m,g}$ as the sum of amino acids of group *g*, in linear motif instance *m*, normalized with the length of the alignment. We defined amino acid groups as polar and positive: RKH; polar and negative: DE, polar and neutral: STQN; aliphatic: AILVM; aromatic: FYW, turn-forming: PG, covalent: C.

$$T_{m,g} = \frac{\sum_{p=1}^{L} aa_p}{L}$$

Finally, the distance of motifs *x* and *y* is defined as:

$$D(x, y) = (\sum_{p=1}^{L} E_{aa_{p,x}aa_{p,y}} * I_p + \sum_{g=1}^{G}(T_{g,x} - T_{g,y}))/2$$

where *L* is the alignment length and *G* is number the amino acid categories.
These distance matrices were then fed into K-means or hierarchical clustering algorithms, typically yielding comparable results.

2.3 Prediction

To train the predictors we randomly selected proteins from each localization group. The least populated localization group (Basolateral) was used as a standard: the number of proteins selected for each class is equal to 75% of the basolateral protein class size. Each predictor had five variants, to cover most of the dataset, each with different training sets (Supplementary Figure 1).

We used CCTOP (Dobson et al., 2015a, 2015b) to predict TM regions, IUPred (Mészáros et al., 2018) to detect disordered regions and SEG through PlatoLoCo (Jarnot et al., 2020) to detect low complexity regions. We used BLAST on SwissProt to define conservation: only the first hit of each species was accepted if the sequence length did not differ more than 25% of the query sequence. A position was defined conserved, if at least 60% of the aligned positions matched. We built two 20x20 matrices, where each cell contained the frequency of a given residue pair (distance is maximum two residues, adjacent residues were multiplied with two) defined by the row/column. The order of amino acids reflected the physico-chemical properties, similarly as for the clustering steps. The bottom triangle of the matrix represented extracellular regions, while the top triangle represented cytoplasmic regions. Furthermore, we built two different matrices, and used disordered regions and low complexity regions as filters.

The CNNs had 20x20x2 dimension input, with 'Adam' optimizer, and used 2 Conv2D layers (32 layer; size 3, activation relu and 64 layers; size 3; activation relu), followed by a MaxPooling layer (size 2), Dropout layer (0.25), Dense layer (128 neurons; activation: relu) and a final Dropout layer (0.5) before the binary output. The CNNs were trained for 20 epochs, with early stopping if validation loss decresased for 3 constitutive epochs. The classical fully connected NNs had 7 input features, 28 hidden layers, sigmoid activations and 'Adam' optimizer and was trained for 100 epochs. The final NNs uses the input of all these predictors: the first layer contains 10 hidden neurons with sigmoid activation, the final layer has two or three output neurons (for the binary and the categorical versions, respectively) with softmax activation. These Neural Networks were trained for 100 epochs. The output is the mean of these individual predictors (Supplementary Table 8-9).

3 Results

3.1 Datasets

The core of our dataset is derived from the PolarProtDb database. The number of unique genes was relatively low, therefore we applied a similar approach to what we used in PolarProtDb previously: We noticed that relatively close vertebrate orthologs usually localize to the same membrane domain, therefore the collected set of proteins was extended by its orthologs (for more details see Methods). We created 3 datasets containing TMPs with different localization (see Methods). I) The "Training dataset" contains 1011, 759, 2037, 5464 apical, basolateral, generic outer plasma membrane and other membrane proteins, respectively. II) For the "Independent test set" 70, 67, 60, 56 proteins were selected from the same localization. III) The "Human AB dataset" is derived from the "Training dataset" as well as the "Independent test set", and it contains apical and basolateral proteins in *Homo sapiens*.

3.2 Information content of generic SLiM classes offer a natural classification scheme correlating with apical and basolateral localization

During the preparation of PolarProtDb, we noticed certain short linear motifs appear frequently and they can often be related to the sorting. We focused our attention at two particular cases, dileucine-like motifs and PDZ ligands. Dileucine motifs (most typically having an architecture of Glu-x-x-x-φ-φ, with φ being hydrophobic, most commonly Leu) are short linear motifs found on the cytoplasmic tails of many TM proteins, and widely known for their role in endocytosis, basolateral localization, or lysosomal targeting. These linear motifs bind to the sigma subunits of the four major, conserved adaptin complexes (AP-1 to AP-4). The name "dileucine" is a misnomer because motifs belonging to this class can carry different hydrophobic amino acids, as long as the glutamate is preserved. In certain instances, only one hydrophobic position is detected (sometimes labelled "monoleucine" motifs), or the glutamate can also be missing. We used previously published articles to establish a collection of dileucine-type motifs with a known role in either the basolateral polarization (that is thought to be governed by AP-1 complexes), or endocytosis (AP-2 associated) as well as lysosomal trafficking (that is mostly AP-3 driven) (Park et al, 2014).

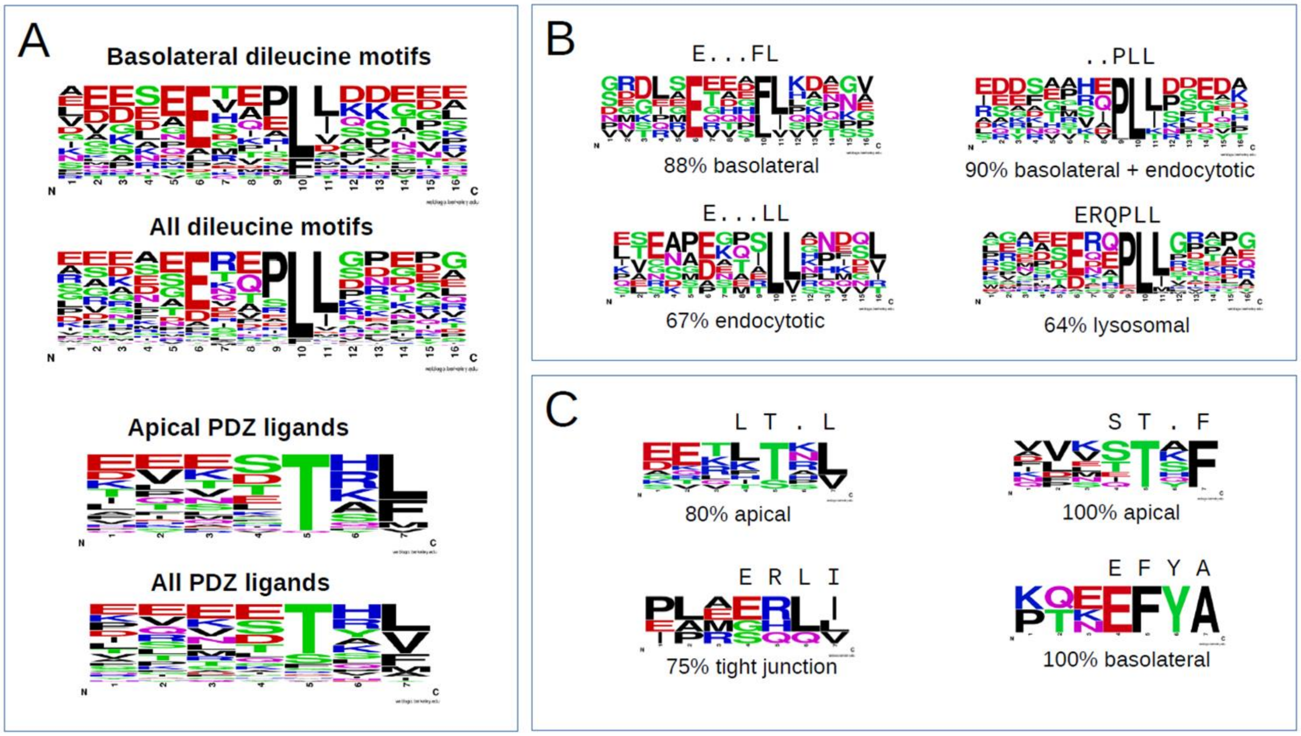


Figure 1: A). Sequence logos (frequency plots) of the apical-specific PDZ-binding motifs and the basolaterally localizing dileucine motifs, compared to more generic motif sets. Panels B) and C): Selected clusters (n=8 total for both motif families, using agglomerative clustering) with high enrichment of specific subcellular localizations. The most striking amino acid preferences are highlighted above each cluster logo.

We assumed that motifs associating with the same adaptin complexes would display similar amino acid preferences in each position, while those binding to different adaptins are likely to be more different. Notably, basolaterally localizing dileucine-type motifs are prone to be suboptimal, rarely use double leucine and frequently omit one hydrophobic position.

The other motifs we recovered from the literature belong to the sphere of PDZ ligand motifs. These usually strictly C-terminal linear motifs associate with PDZ domain containing proteins, that regulate assembly of cell-cell contacts, tight junctions and apical or basal membrane compartments. We observed that while the basolateral proteins carry highly diverse PDZ ligand motifs, those located in the apical compartment tend to look far more uniform, with the generic consensus T[RKH][LFM]$ covering the majority of instances. Literature searches for matching PDZ domain proteins suggest that NHERFs (important components of the subapical cytoskeleton) might be among the PDZ partners that could explain this observation. Phage display data on NHERF3 (earlier called PDZK1) also suggests that this presumed partnership might be correct (Gfeller et al, 2012). Therefore, we assumed that apically polarized proteins are possible to predict based on this specific PDZ ligand motif (whenever present), similarly to the basolaterally-localizing dileucine motifs.

To better delineate these two motif families, (neither of which are exclusive to sorting) we devised an information-based clustering method. First, we collected the apical PDZ-binding and basolateral dileucine motifs from the literature (Supplementary Table 6,7), while also adding many counter-examples to each set. The localization-specific motifs show relatively limited difference from the complete set (i.e., those including motifs from all other locations as well), as shown on Figure 1A. However, we managed to split them into meaningful subclasses using the information content of their sequence alone (see Methods), recovering much of these specialized, polarized sorting-associated subsets, with features unique to each subgroup (Figure 1B and C).

3.3 Pattern discovery combined with structural information provides novel putative motifs indicating localization

The above detailed clustering approach is a great indication of localization, but they are too specific for a proteome wide analysis as many proteins use different sorting routes to reach their final destination. To generalize this method, we scanned proteins of the "Human dataset" for repeating patterns using combinatorial approaches (see Methods). To reduce the possible false positive hits, only those putative motifs were accepted, where the conservation was visible across orthologs and the region was intrinsically disordered. We also randomly shuffled sequences ten times, and accepted motifs when their average discovery rate plus three times the standard deviation was lower compared to real hits to ensure high specificity.

Unfortunately, the list of patterns at this step was still very noisy, and many patterns belonged to apical and basolateral proteins as well. To reveal meaningful hits, we incorporated topology information and clustered the hits using the same approach as for PDZ-binding apical and dileucine-like basolateral motifs. Figure 2 shows the most specific putative motifs (if at least 70% of the protein hits belong to the same localization) based on I) apical, basolateral localization and II) which side of the membrane they appear, i.e., if they fall into a cytoplasmic/extracellular disordered region. We counted the occurrence of these motifs in apical and basolateral TM proteins and found this approach to be robust enough to highlight patterns more frequently appearing in one side of the polarized cell, even

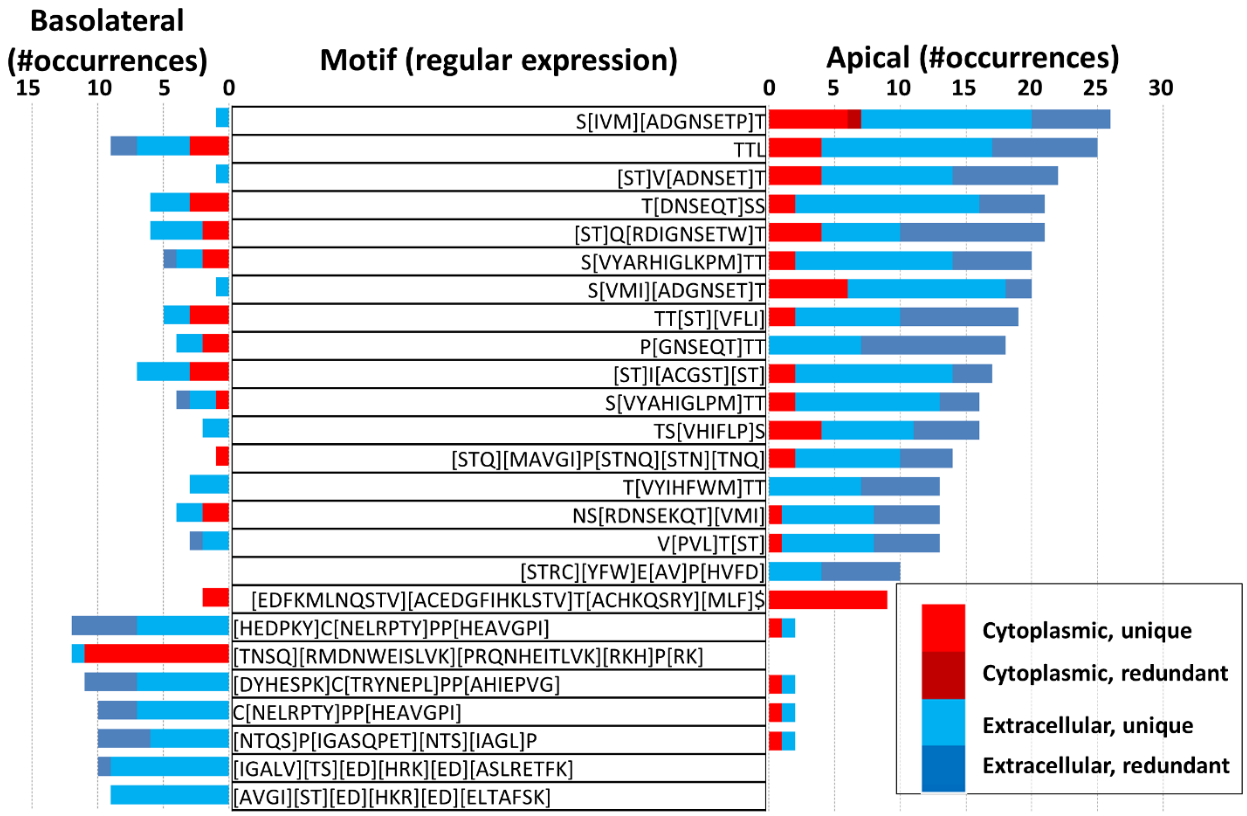


Figure 2: Putative short linear motifs dominating in proteins of apical or basolateral membranes. Motifs appearing in multiple apical or basolateral proteins were collected and were then sorted based on the number of their occurrences. Cytoplasmic motifs: red; Extracellular motifs: blue. Unique and redundant motifs have lighter and darker shade, respectively.

if we lack biochemical context for most of these linear motifs. Notably, some of the putative motifs are somewhat redundant; However, these small variations seem to code important information, as prediction accuracy (see later) drops when we remove them.

To achieve maximum performance, the motif identification was done on human proteins only. Although the full dataset would have yielded much more hits, evolutionary relatives would inherently bias the above-mentioned clustering method: In many cases, orthologs show only few residue changes and the clustering would only group species instead of true subclasses. Furthermore, we removed collagens and mucin-like proteins, as they are highly repetitive, and their incorporation would highlight non-specific motifs.

3.4 Machine learning approach to classify protein localization in polarized cells from sequence information

Although highlighting putative linear motifs can be promising, the presence of such a motif is far from certainty regarding protein localization, considering they are not experimentally verified, moreover they do not cover every protein. To overcome these limitations, we built several Neural Networks (NNs) to classify protein localization in polarized cells. Since apical and basolateral membranes can be considered as plasma membranes, we prepared four datasets containing apical, basolateral, plasma and other (endoplasmic reticulum, mitochondria, etc.) TMPs. As the dataset contains a different number of proteins for each localization, we used bootstrap aggregating to split data into smaller, but roughly even sets.

We prepared binary fully connected NNs to classify proteins based on their localization (predicting apical, basolateral, or other membrane classes). Input features include the detected motifs, if they are included in a disordered region and they are conserved (threshold above 0.6, see Methods). The input array also distinguishes extracellular and cytoplasmic localization. Although we achieved moderate success with this method (Supplementary Table 10, columns F-H: 54-71% accuracy on different sets), we concluded that there is still room for improvements.

Current deep learning techniques often utilize Position Specific Scoring Matrices of different segments of the protein. In this case, however, the sorting signal is more likely included in more compact linear motifs, therefore we prepared an architecture that takes close residue pairs into consideration. The few residue distances between amino acids in SLiMs suggest that adjacent or near residue pairs may provide an abstract level of information that Convolutional Neural Networks (CNNs) can capture. Each protein sequence was converted into two 20x20 matrices, representing the 20 standard amino acids. Values in this matrix were calculated based on the distance of different residue pairs, where adjacent amino acid pairs increase the value of a point with a higher value compared to distant ones (see Methods). The top and the bottom triangle specifies cytoplasmic and extracellular localization. Furthermore, each protein has two matrices, one for disordered regions and another one for low complexity regions (Supplementary Figure 1). The CNN achieved 62-85% accuracy on predicting

Table 1: Performance metrics of the binary (apical vs basolateral) and the ternary (apical vs basolateral vs other) predictors. Since the ternary predictor classifies proteins into three classes, results are show for each category (Apical vs. not apical, basolateral vs not basolateral, other vs apical, basolateral). BAC: Balanced Accuracy. Sens: Sensitivity. Spec: Specificity. MCC: Matthews Correlation Coefficient. AUC: Area Under Curve.

Predictor	True Class	BAC	Sens	Spec	MCC	AUC
Binary	Apical	0.90	0.91	0.88	0.80	0.96
Ternary	Apical	0.86	0.89	0.84	0.67	0.94
	Basolateral	0.71	0.54	0.88	0.43	0.77
	Not apical/basolateral	0.69	0.61	0.77	0.39	0.80

different classes (Supplementary Table 10, column B-E). Although predicting apical localization is quite accurate, the basolateral classifiers suffer from some level of overfitting.

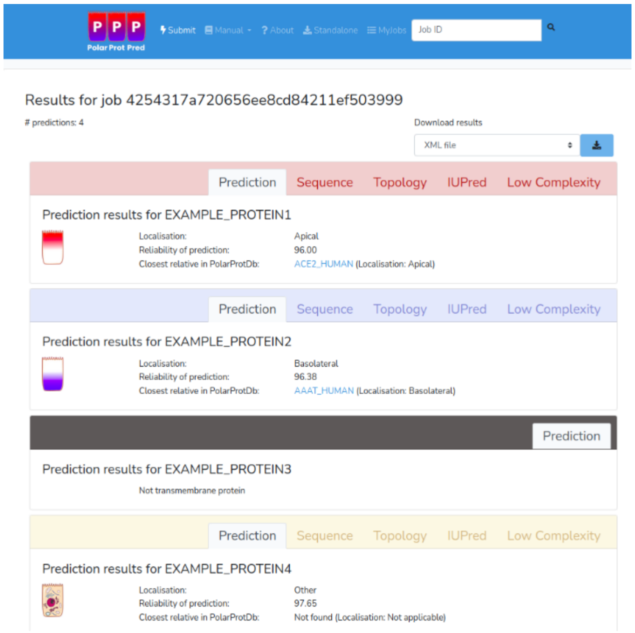
Last, but not least, we combined the output of these predictors to classify proteins (see the architecture on Supplementary Figure 1). We prepared two predictors: the 'binary' mode predicts apical vs. basolateral localization in polarized cells, the 'ternary' mode distinguishes apical, basolateral, and other membrane proteins. According to the independent test sets, the binary predictor has 89% balanced accuracy, while the categorical predictor reached 70-85% accuracy, with outstanding AUC in case of predicting apical localization, regardless of 'binary'/'categorical' mode (Table1, Supplementary Figure 2-3). Notably, while classes may be represented with different number of proteins in some cases, most of the metrics (Balanced accuracy, Sensitivity, Specificity and Matthews Correlation Coefficient) can be used even if the classes have different sizes.

3.5 Availability

Protein sequences can be submitted at http://polarprotpred.ttk.hu. The user may ask for an email alert containing a link referring to the results. Users can select from apical/basolateral ('binary') prediction mode (if they are confident that their protein is expressed in polarized cells) or apical, basolateral, other ('ternary') prediction. When the submitted job is finished, a five-panel window is produced by the PolarProtPred web server (Figure 3). Results are stored on our server and can be accessed later.

The first panel contains the predictions results. If the sequence was found in the PolarProtDb, it is also linked on the main page. The second panel displays the submitted sequence. Panel 3-5 shows the predicted TM, disordered, low-complexity regions, respectively.

The main purpose of the web server is to provide an easy access user interface for the PolarProtPred method. Although some of the utilized methods have high computation requirements and setting up these programs locally is rather time consuming, its source code is available on https://github.com/brgenzim/PolarProtPred. PolarProtPred can also be accessed via a programmable direct interface.



Results for job 4254317a720656ee8cd84211ef503999

predictions: 4

Prediction results for EXAMPLE_PROTEIN1

Localisation: Apical
Reliability of prediction: 96.00
Closest relative in PolarProtDb: ACE2_HUMAN (Localisation: Apical)

Prediction results for EXAMPLE_PROTEIN2

Localisation: Basolateral
Reliability of prediction: 96.38
Closest relative in PolarProtDb: AAAT_HUMAN (Localisation: Basolateral)

Prediction results for EXAMPLE_PROTEIN3

Not transmembrane protein

Prediction results for EXAMPLE_PROTEIN4

Localisation: Other
Reliability of prediction: 97.65
Closest relative in PolarProtDb: Not found (Localisation: Not applicable)

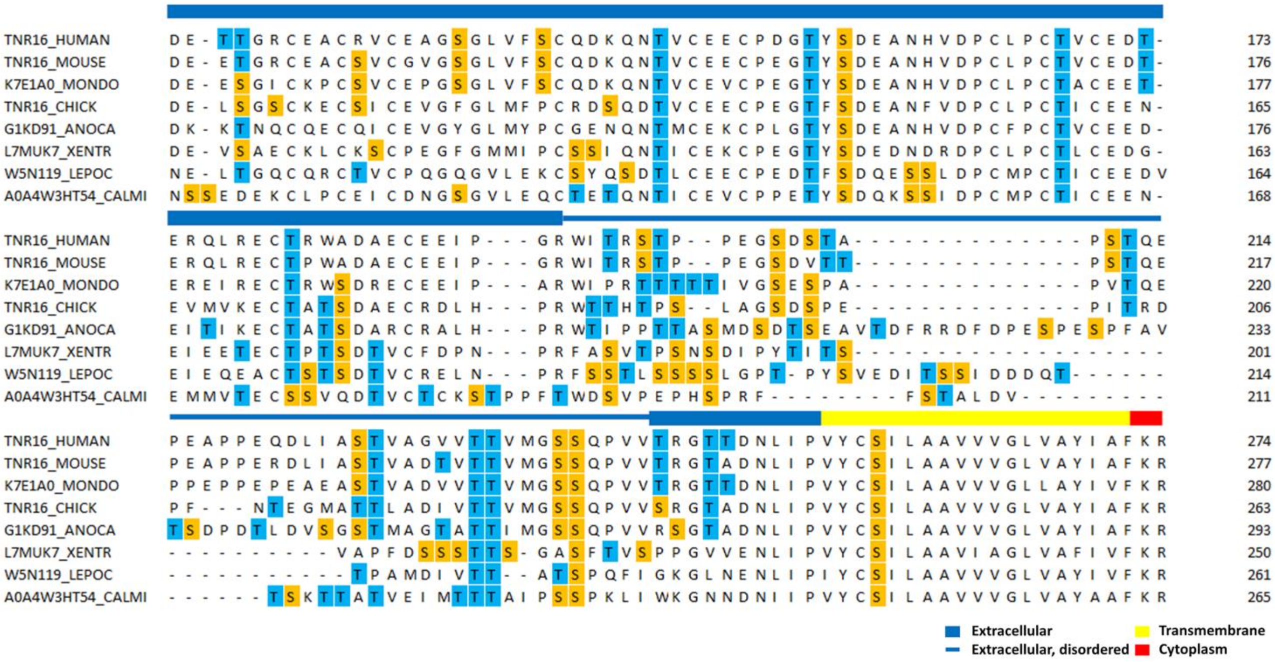
Figure 3: Layout of the PolarProtPred web server. The first panel summarizes the prediction results and shows cross-reference to PolarProtDB (if applicable). The second panel shows the submitted sequence, while panel 3-5 shows predicted structural features, respectively.

4 Discussion

4.1 Limitations

Despite the simplicity of our approach and its merits, we are aware that the model also has many shortcomings. One of the most important problem relates to the fact that protein localization is not a binary variable in cells. What is more, apico-basal sorting of many proteins is not stationary but depends on developmental stage of the cell as well as the actual tissue type. To circumvent these problems, our learning set was mostly based on proteins expressed or experimentally validated in mature, polarized MDCK cells, or tissues known to obey highly similar sorting rules (e.g., small intestine enterocytes or the Caco-2 cell monolayers). However, there are other epithelial tissues whose sorting rules appear to be mildly (e.g., choroid plexus epithelium, with apically localized K+/Na+ pumps) or highly different (e.g., placental chorion epithel). Obviously, we need to learn much more of these specialized tissue polarities before similar machine learning approaches could become universally applicable.

Another caveat is that current analytical methods provide limited information about polarized cells: They usually characterize individual proteins with immunolocalization or apical/basal membrane-specific labeling of amino acids before proteomic analysis. All these techniques provide a relative measure only, whereas the other side of the polarized cell is often not monitored. The main bottleneck of immunolocalization is the limited availability of appropriate antibodies. Monoclonal antibodies can be highly specific but might also recognize multiple epitopes. Thus in the latter case we, cannot safely assume that they are specific enough to the protein of interest, or mark the correct isoform. Selective labeling uses primary amine-specific reagents, enabling them to identify only those proteins that have such available regions. Coverage of TM proteins is relatively low in proteomic analyses, therefore differences between the two sides are based on a limited number of peptides.



```
TNR16_HUMAN         DE-TTGRCEACRVCEAGSGLVFSCQDKQNTVCEECPDGTYSDEANHVDPCLPCTVCEDT-   173
TNR16_MOUSE         DE-ETGRCEACSVCGVGSGLVFSCQDKQNTVCEECPEGTYSDEANHVDPCLPCTVCEDT-   176
K7E1A0_MONDO        DE-ESGICKPCSVCEPGSGLVFSCQDKQNTVCEVCPEGTYSDEANHVDPCLPCTACEET-   177
TNR16_CHICK         DE-LSGSCKECSICEVGFGLMFPCRDSQDTVCEECPEGTFSDEANFVDPCLPCTICEEN-   165
G1KD91_ANOCA        DK-KTNQCQECQICEVGYGLMYPCGENQNTMCEKCPLGTYSDEANHVDPCFPCTVCEED-   176
L7MUK7_XENTR        DE-VSAECKLCKSCPEGFGMMIPCSSIQNTICEKCPEGTYSDEDNDRDPCLPCTLCEDG-   163
W5N119_LEPOC        NE-LTGQCQRCTVCPQGQGVLEKCSYQSDTLCEECPEDTFSDQESSLDPCMPCTICEEDV   164
A0A4W3HT54_CALMI    NSSEDEKCLPCEICDNGSGVLEQCTETQNTICEVCPPETYSDQKSSIDPCMPCTICEEN-   168

TNR16_HUMAN         ERQLRECTRWADAECEEIP---GRWITRSTP--PEGSDSTA--------------PSTQE   214
TNR16_MOUSE         ERQLRECTPWADAECEEIP---GRWITRSTP--PEGSDVTT--------------PSTQE   217
K7E1A0_MONDO        EREIRECTRWSDRECEEIP---ARWIPRTTTTTIVGSESPA--------------PVTQE   220
TNR16_CHICK         EVMVKECTATSDAECRDLH---PRWTTHTPS--LAGSDSPE--------------PITRD   206
G1KD91_ANOCA        EITIKECTATSDARCRALH---PRWTIPPTTASMDSDTSEAVTDFRRDFDPESPESPFAV   233
L7MUK7_XENTR        EIEETECTPTSDTVCFDPN---PRFASVTPSNSDIPYTITS-------------------   201
W5N119_LEPOC        EIEQEACTSTSDTVCRELN---PRFSSTLSSSSLGPT-PYSVEDITSSIDDDQT------   214
A0A4W3HT54_CALMI    EMMVTECSSVQDTVCTCKSTPPFTWDSVPEPHSPRF--------FSTALDV---------   211

TNR16_HUMAN         PEAPPEQDLIASTVAGVVTTVMGSSQPVVTRGTTDNLIPVYCSILAAVVVGLVAYIAFKR   274
TNR16_MOUSE         PEAPPERDLIASTVADTVTTVMGSSQPVVTRGTADNLIPVYCSILAAVVVGLVAYIAFKR   277
K7E1A0_MONDO        PPEPPEPEAEASTVADVVTTVMGSSQPVVTRGTTDNLIPVYCSILAAVVVGLLAYIVFKR   280
TNR16_CHICK         PF---NTEGMATTLADIVTTVMGSSQPVVSRGTADNLIPVYCSILAAVVVGLVAYIAFKR   263
G1KD91_ANOCA        TSDPDTLDVSGSTMAGTATTIMGSSQPVVRSGTADNLIPVYCSILAAVVVGLVAYIAFKR   293
L7MUK7_XENTR        ----------VAPFDSSSTTS-GASFTVSPPGVVENLIPVYCSILAAVIAGLVAFIVFKR   250
W5N119_LEPOC        -----------TPAMDIVTT--ATSPQFIGKGLNENLIPIYCSILAAVVVGLVAYIVFKR   261
A0A4W3HT54_CALMI    ------TSKTTATVEIMTTTAIPSSPKLIWKGNNDNIIPVYCSILAAVVVGLVAYAAFKR   265
```

Figure 4: Sequence alignment of the O-glycosylated linker region between the extracellular domains (blue) and the TM helix (yellow) of the low affinity Neurotrophin receptor (p75NTR or TNFR16). Despite excellent architectural homology between vertebrate receptors, only a few Ser-Thr rich repeats are consistently conserved. This figure was generated from an automated ClustalOmega aligment with slight corrections. HUMAN = Homo sapiens, MOUSE = Mus musculus, MONDO = Monodelphis domestica, CHICK = Gallus gallus, ANOCA = Anolis carolinensis, XENTR = Xenopus tropicalis, LEPOC = Lepisosteus oculatus, CALMI = Callorhincus milii.

Computational limitations also arise: Disordered regions are rather hard to predict in TMPs (Tusnády et al., 2015), thus any kind of prediction or analysis that relies on these regions has a disadvantage. Additional problem is the detection of short linear motifs: these segments are extremely hard to capture using bioinformatics tools alone, and only a handful of experimentally verified instances are available. Hopefully, we can successfully overcome these limitations by manipulating cutoffs for disordered regions programs and by defining new motif clusters.

4.2 Topology constraints and motif clustering increase the specificity of combinatorial pattern detection

The computational identification of SLiMs is a challenging task in general due to the low information content of motifs. Although scanning services (such as the ELM server) offer various filtering tools (conservation, accessibility, localization), they still result in a large number of false positive hits. One important bottleneck is that in most cases we do not have information about the conformation of the motif or the interacting partner, obviously limiting the description of the motif. We can define SLiMs using regular expressions, sequence logos or Position Specific Scoring Matrices, however all these representations have several limitations: all definitions are quite permissive and therefore allow a lot of different conformation, as they handle logical statements poorly (for example often a less specific definition can be splitted into two or more highly specific one - which would be much more easier knowing the interacting partner). So far these limitations can be only overcome by complicated experiments and laborious analysis, as in the case of PP2A motif (Hertz *et al.*, 2016). Here we showed the information in general SLiM classes combined with subcellular localization can be utilized to break them down into meaningful subclasses. Our clustering approach efficiently disentangles some of the above-mentioned limitations by defining subclasses of proteins using sequence information alone.

4.3 Examples with O-glycosylation regions governing sorting

Our motif discovery approach yields a rather large number of potential linear motifs, with a clearly greater success for apical determinants. The extracellularly located, serine/threonine rich apical motifs are relatively straightforward to interpret, although they are still not trivial from a biological perspective (Figure 2*)*. First, we observed that these motifs are built from simpler underlying principles: Most commonly a serine and a threonine amino acid being located in a +3 relationship (T..T, S..T, T..S or S..S). The other basic pattern involves two adjacent threonines or serines (most commonly TT, sometimes TS or SS).

Protein N-acetylgalactosamine (GalNac) transferases, the key enzymes of mucin-type O-glycosylation have a rather loose direct substrate site consensus (apart from being highly serine-threonine-proline rich, with a preference for threonine for the N-acetylgalactosamine attachment), but often display a striking processivity (Revoredo et al, 2016). Certain GalNacTs (GalNAcT4, GalNAcT7 and GalNAcT10) also preferentially glycosylate directly adjacent acceptor sites, explaining our Thr-Thr or Thr-Ser motifs. Other GalNacTs are also enhanced by pre-existing nearby O-glycans spaced further apart (such as GalNac-T12), yielding two target sites in a +3 relationship. Yet other spacings are also possible, up to a 5-spacing arrangement (GalNacT2, GalNacT3, GalNacT5), yet these enzymes are less stringent, and are not expected to yield clean linear motifs. The presence of lectin domains in GalNacTs also allows long-range processivity (6 to 15 amino acids apart), resulting in highly saturated O-glycosylation regions, whose exact glycan attachment sites are difficult if not impossible to predict (De las Rivas et al, 2019). At the same time, the O-glycoregion is relatively simple to detect with bioinformatic methods (Nishikawa et al 2010). The fact that many of the O-glycosylation motifs overlap each other, often in imperfect copies, also explains their high redundancy within the same protein.

A cautious alignment of these regions reveals that although they are architecturally conserved within most vertebrate proteins, exact sequence matches are rare, as expected by the numerous, imperfect, partially redundant O-glycosylation sites (Figure 4). Although extracellular disordered regions are depleted in TMPs (Tusnády et al., 2015), they are definitely present and one important function of them is to serve sites for glycosylation mediating sorting (Goutham et al., 2020).

4.4 Other resources, similar approaches

Recent decades provided several experimentally derived datasets that utilized high-throughput experimental methods to clarify localization of human membrane proteins (Caceres *et al.*, 2019). There is also a good number of prediction methods that predict localization information, either the presence of a signal peptide (Armenteros *et al.*, 2019), or the exact localization of proteins (Almagro Armenteros *et al.*, 2017). Some of these methods were trained on data automatically downloaded from computationally annotated databases, thus any bias in their sources might affect their prediction accuracy, however, without any visible sign as their performance was measured on noisy datasets. In contrast, we manually annotated each apical, basolateral membrane protein, providing a clean training set for our method. Although the prediction performance is lower compared to the high accuracy of most current bioinformatic tools, it is reasonable to assume that accuracy has limitations in this case, as many proteins show multiple localization in different experiments.

Funding

This work was supported by grants from Hungarian Research and Developments Fund [OTKA K119287 and 132522]. L.D. was funded by EMBO STF-8784. G.E.T. [LP2012-35] was supported by the Momentum Grant of the Hungarian Academy of Sciences.

Conflict of Interest: none declared.

References

Almagro Armenteros,J.J. *et al.* (2017) DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33, 3387–3395.

Armenteros,J.J.A. *et al.* (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.*, 37, 420–423.

Bryant,D.M. and Mostov,K.E. (2008) From cells to organs: building polarized tissue. *Nat. Rev. Mol. Cell Biol.*, 9, 887–901.

Caceres,P.S. *et al.* (2019) Quantitative proteomics of MDCK cells identify unrecognized roles of clathrin adaptor AP-1 in polarized distribution of surface proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 116, 11796–11805.

Cong,Y. and Ren,X. (2014) Coronavirus entry and release in polarized epithelial cells: a review. *Rev. Med. Virol.*, **24**, 308–315.

Davey,N.E. *et al.* (2012) Attributes of short linear motifs. *Mol. Biosyst.*, 8, 268–281.

Di Martino,R. *et al.* (2019) Regulation of cargo export and sorting at the trans-Golgi network. *FEBS Lett.*, 593, 2306–2318.

Dobson,L. *et al.* (2015a) CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.*, 43, W408–12.

Dobson,L. *et al.* (2015b) The human transmembrane proteome. *Biol. Direct*, 10, 31.

Dunbar,L.A. *et al.* (2000) A transmembrane segment determines the steady-state localization of an ion-transporting adenosine triphosphatase. *J. Cell Biol.*, 148, 769–778.

Evdokimov,K. *et al.* (2016) Leda-1/Pianp is targeted to the basolateral plasma membrane by a distinct intracellular juxtamembrane region and modulates barrier properties and E-Cadherin processing. *Biochem. Biophys. Res. Commun.*, 475, 342–349.

Farr,G.A. *et al.* (2009) Membrane proteins follow multiple pathways to the basolateral cell surface in polarized epithelial cells. *J. Cell Biol.*, 186, 269–282.

Garcia-Gonzalo,F.R. and Reiter,J.F. (2012) Scoring a backstage pass: mechanisms of ciliogenesis and ciliary access. *J. Cell Biol.*, 197, 697–709.

Gfeller,D. *(*2012) Uncovering new aspects of protein interactions through analysis of specificity landscapes in peptide recognition domains. *FEBS Lett.*, **586**, 2764–2772.

Goutham,S. *et al.* (2020) Mutually exclusive locales for N-linked glycans and disorder in human glycoproteins. *Sci. Rep.*, 10, 6040.

Hanukoglu,I. *et al.* (2017) Expression of epithelial sodium channel (ENaC) and CFTR in the human epidermis and epidermal appendages. *Histochem. Cell Biol.*, 147, 733–748.

Hegedűs,T. *et al.* (2015) Inconsistencies in the red blood cell membrane proteome analysis: generation of a database for research and diagnostic applications. *Database* , 2015, bav056.

Hertz,E.P.T. *et al.* (2016) A Conserved Motif Provides Binding Specificity to the PP2A-B56 Phosphatase. *Mol. Cell*, 63, 686–695.

Hunziker,W. and Fumey,C. (1994) A di-leucine motif mediates endocytosis and basolateral sorting of macrophage IgG Fc receptors in MDCK cells. *The EMBO Journal*, 13, 2963–2969.

Inukai,K. *et al.* (2004) Carboxy terminus of glucose transporter 3 contains an apical membrane targeting domain. *Mol. Endocrinol.*, 18, 339–349.

Jarnot,P. *et al.* (2020) PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. *Nucleic Acids Res.*, 48, W77–W84.

Kumar,M. et al. (2019) ELM—the eukaryotic linear motif resource in 2020. Nucleic Acids Res., 48, D296–D306.

Laird,V. and Spiess,M. (2000) A novel assay to demonstrate an intersection of the exocytic and endocytic pathways at early endosomes. *Exp. Cell Res.*, 260, 340–345.

Langó,T. *et al.* (2017) Identification of Extracellular Segments by Mass Spectrometry Improves Topology Prediction of Transmembrane Proteins. *Sci. Rep.*, 7, 42610.

Langó,T. *et al.* (2020) Partial proteolysis improves the identification of the extracellular segments of transmembrane proteins by surface biotinylation. *Sci. Rep.*, 10, 8880.

de Las Rivas, M. et al. (2019) Polypeptide GalNAc-Ts: from redundancy to specificity. *Curr. Opin. Struct. Biol.*, **56**, 87–96.

Le Bivic,A. *et al.* (1991) An internal deletion in the cytoplasmic tail reverses the apical localization of human NGF receptor in transfected MDCK cells. *J. Cell Biol.*, 115, 607–618.

Martín,M. *et al.* (2019) A Carboxy-Terminal Monoleucine-Based Motif Participates in the Basolateral Targeting of the Na+/I- Symporter. *Endocrinology*, 160, 156–168.

Mészáros,B. *et al.* (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.*, 46, W329–W337.

Müller,A. *et al.* (2019) Covalently modified carboxyl side chains on cell surface leads to a novel method toward topology analysis of transmembrane proteins. *Sci. Rep.*, 9, 15729.

Nishikawa,I. et al. (2010) Computational prediction of O-linked glycosylation sites that preferentially map on intrinsically disordered regions of extracellular proteins. *Int. J. Mol. Sci.*, **11**, 4991–5008.

Park,S.Y. and Guo,X. (2014) Adaptor protein complexes and intracellular transport. *Biosci. Rep.*, **34**.

Revoredo,L. et al. *(2016)* Mucin-type O-glycosylation is controlled by short- and long-range glycopeptide substrate recognition that varies among members of the polypeptide GalNAc transferase family. *Glycobiology*, **26**, 360–376.

Riga,A. *et al.* (2020) New insights into apical-basal polarization in epithelia. *Curr. Opin. Cell Biol.*, 62, 1–8.

Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, 14, 55–67.

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7, 539.

Simons,K. and Ikonen,E. (1997) Functional rafts in cell membranes. *Nature*, 387, 569–572.

Stoops,E.H. and Caplan,M.J. (2014) Trafficking to the apical and basolateral membranes in polarized epithelial cells. *J. Am. Soc. Nephrol.*, 25, 1375–1386.

Tusnády,G.E. *et al.* (2015) Disordered regions in transmembrane proteins. *Biochim. Biophys. Acta*, 1848, 2839–2848.

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, 47, D506–D515.

Urquhart,P. *et al.* (2005) N-glycans as apical targeting signals in polarized epithelial cells. *Biochem. Soc. Symp.*, 39–45.

Van Roey,K. *et al.* (2014) Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.*, 114, 6733–6778.

Weisz,O.A. and Rodriguez-Boulan,E. (2009) Apical trafficking in epithelial cells: signals, clusters and motors. *J. Cell Sci.*, 122, 4253–4266.

Wohlgemuth,N. et al. (2018) Influenza A Virus M2 Protein Apical Targeting Is Required for Efficient Virus Replication. *J. Virol.*, **92**.

Yeaman,C. *et al.* (1997) The O-glycosylated stalk domain is required for apical sorting of neurotrophin receptors in polarized MDCK cells. *J. Cell Biol.*, 139, 929–940.

Zeke,A. *et al.* (2020) PolarProtDb: A Database of Transmembrane and Secreted Proteins showing Apical-Basal Polarity. *J. Mol. Biol.*, 166705.