Putative linear motifs mediate the trafficking to apical and basolateral membranes

Laszlo Dobson[1,*], András Zeke[1,*], Levente Szekeres[1], Tamás Langó[1], Gábor Tusnády[1]

[1]Research Centre for Natural Sciences, Magyar Tudósok Körútja 2, 1117 Budapest, Hungary
*Contributed equally

Abstract

Cell polarity refers to the asymmetric organisation of cellular components in various cells. Epithelial cells are the best known examples of polarized cells, featuring apical and basolateral membrane domains. Despite huge efforts, the exact rules governing the protein distribution in such domains are still elusive. In this study we examined linear motifs accumulating in these parts and based on the results we prepared 'Classical' and Convolutional Neural Networks to classify human transmembrane proteins localizing into apical/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 dataset may offer help to experimentalists to characterize novel molecular targets to regulate transport processes more precisely.

Introduction

Polarity is an essential feature of cells, especially in differentiated, multicellular organisms. In these cells, components (e.g. plasma membrane proteins, cytoskeletal components) 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, as well as to deliver information to different regions of an organism [1], for example they maintain ion homeostasis in the eccrine glands and ducts [2] or play a role in nutrient uptake [3]. Although we have an increasingly growing knowledge of the main determinants of apical and basolateral polarity networks, the exact composition of these membranes are still elusive for most tissues [4]. 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. Many times 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 [5]. During this journey to the cell surface proteins are tightly regulated via post-translational modifications and transient interactions with other molecules [6].

Many of these regulation processes are mediated via Short linear motifs (SLiMs), flexible protein segments composed of a restricted number of residues (usually between 3-10), that usually bind to ordered protein domains via coupled folding and binding. Their properties enable them to bind to a diverse range of partners with low micromolar affinity and establish transient interactions [7]. Besides mediating protein interactions, they also provide sites for post-translational notifications or proteolytic cleavage sites [8]. Recent decades provided 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 include multiple pathways [9,10] and often include cargo sorting [11]. The basolateral targeting of transmembrane proteins may rely on cytosolic tyrosine [12], mono- and dileucine motifs [13,14]. Localization may also be proteolytic processing and glycosylation dependent [15]. In contrast, the apical targeting can occur in the absence of basolateral signal and may also involve rafts [16]. Both N- and O-glycans play important roles in apical sorting [17,18], as well as interaction between transmembrane regions and their surrounding [19]. Apical trafficking is sometimes functionally redundant or multipart, meaning proteins own a set of motifs, and those may serve as replacements for each other, individually capable of proper targeting [6]. The divided nature of apical membranes adds further complication to trafficking [20].

Although we have a moderate understanding which regions/residues/modifications play a critical role in individual proteins to reach their destination, the dozens of possible pathways makes it hard to apply general rules to them. Here we propose a novel approach to classify alpha-helical transmembrane proteins in polarized cells, based on their topology and putative SLiMs driving their localization. We collected hundreds of proteins with reliable experimental evidence of their destination and used computational biologist approaches to characterize motifs responsible for their trafficking. We used the resulting dataset to predict which membrane proteins localize to apical/basolateral membranes. Our dataset can be used by experimentalists to find new molecular targets governing transport processes in polarized cells.

Results

Ser/Thr rich motifs are abundant at the extracellular loops of apical transmembrane proteins

We scanned 200 human transmembrane (TM) proteins (100 apical and 100 basolateral) proteins for linear motifs, and then compared their distribution in disordered extracellular and cytoplasmic regions. To reduce the possible false positive hits, we only accepted putative motifs when conservation was visible across orthologs. 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. We grouped each motif based on their localization, i.e. if they fall into a

cytoplasmic/extracellular disordered region, or into any other segments. Next we counted the occurrence of motifs in apical and basolateral TM proteins (Figure 1).

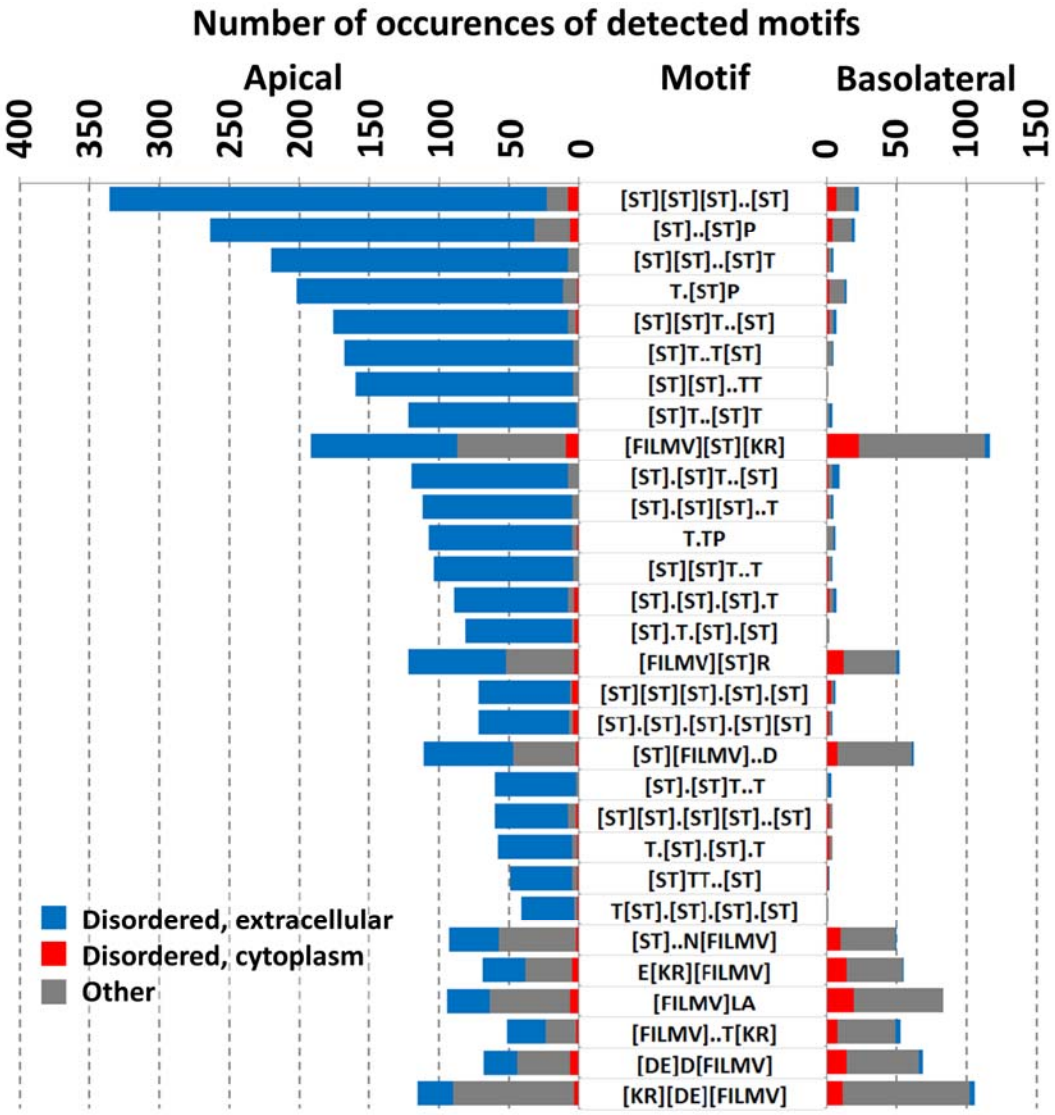

Figure 1. The first 30 most abundant putative linear motifs in apical and basolateral membrane proteins. We have to note, that these motifs highly overlap and they can be merged for a better consensus definition.

We found that motifs containing multiple Ser and Thr residues are highly abundant in the extracellular tail and loops of apical membrane proteins. Notably, ~90% of these motifs are distant from the membrane region (>10 amino acids). Naturally, most of the identified motifs can also be merged for a better consensus description. At the second half of our list, hydrophobic and charged residue containing motifs also appear, however their discriminative power is relatively meager, compared to S/T motifs. The biochemical implications of the latter motifs are still unclear.

Apical and basolateral membrane proteins can be classified based on the distribution of adjacent residue pairs

We prepared a 'classical' Neural Network (NN) to classify proteins based on their localization. Since apical/basolateral membranes can be considered as plasma membranes, we prepared four datasets containing apical, basolateral, plasma and other (Endoplasmic Reticulum, mitochondria, etc.) TM proteins. Input features included detected motifs and their extracellular/cytoplasmic localization, disordered and low complexity features, transmembrane topology and basic amino acid features. Although we achieved moderate success with this method (55% accuracy on multiclass classification), we concluded that there is still room for improvements.

We also prepared a Convolutional Neural Network (CNN), where protein sequences were converted into images. Each image contains 20x20 pixels, representing the 20 standard amino acids. Values in this matrix were calculated based on the distance of different residue pairs. Adjacent amino acid pairs increase the value of a point with a higher value compared to distant ones. The CNN achieved 61% accuracy as a multiclass classifier.

Last, we combined the output of the two predictors to classify proteins into 4 classes. By combining the output probabilities of the 'classical' and the convolutional NN, we increased the accuracy of the method to 66% (Note, as we have 4 classes, random prediction would achieve 25%). The most commonly occurring false cases were the cross-predictions of basolateral and plasma TM proteins, as apical and other TM proteins, respectively. Table 1 shows the binary evaluation of the method. In this case, there are 4 assessments for each group (apical vs other, basolateral vs other etc.). Using this evaluation the method performed quite well and achieved 78-85% accuracy.

Table 1: Binary evaluation of the proposed method.

	Training set				Validation set				Independent test set			
	Apical	Basolateral	Plasma	Other	Apical	Basolateral	Plasma	Other	Apical	Basolateral	Plasma	Other
MCC	0.56	0.72	0.65	0.78	0.56	0.68	0.6	0.75	0.51	0.47	0.38	0.58
Sensitivity	0.6	0.88	0.71	0.8	0.6	0.86	0.65	0.81	0.61	0.69	0.48	0.67
Specificity	0.94	0.87	0.93	0.95	0.93	0.85	0.93	0.94	0.89	0.81	0.88	0.9
Accuracy	0.88	0.87	0.88	0.91	0.86	0.86	0.86	0.9	0.85	0.78	0.78	0.84

Next we evaluated how the output probability (see methods) correlates with the accuracy (Figure 2). Predictions are sorted based on their probability value, the localization accuracies and reliability measured on the benchmark set are plotted against coverage. Reliability and accuracy correlates well. Half of the predictions have 70% or higher probability, where the prediction accuracy is 75%.

Dataset of apical and basolateral membrane proteins

We also run the prediction method on the human TM proteome. The classifier predicted 1285 proteins as apical, and 1475 proteins as basolateral. Considering the human TM proteome contains 5492 proteins, this means roughly 50% of these proteins may localize in the apical/basolateral membranes of epithelial cells during their lifetime. We have to note that cross-prediction from plasma/other to apical/basolateral TM proteins is more common than vice versa, thus the real proportion of such proteins is probably below 50%. We also measured the probability of the predictions on the full proteome and plotted it against accuracy and probability on the benchmark set (Figure 2). The probability distribution of the benchmark set and the proteome highly correlates, therefore it is safe to assume that the accuracy also shows a similar trend on the human TM proteome. Supplementary Table 1 contains the most reliable part of the prediction, i.e. predictions above 75% probability (381 apical and 649 basolateral membrane proteins).

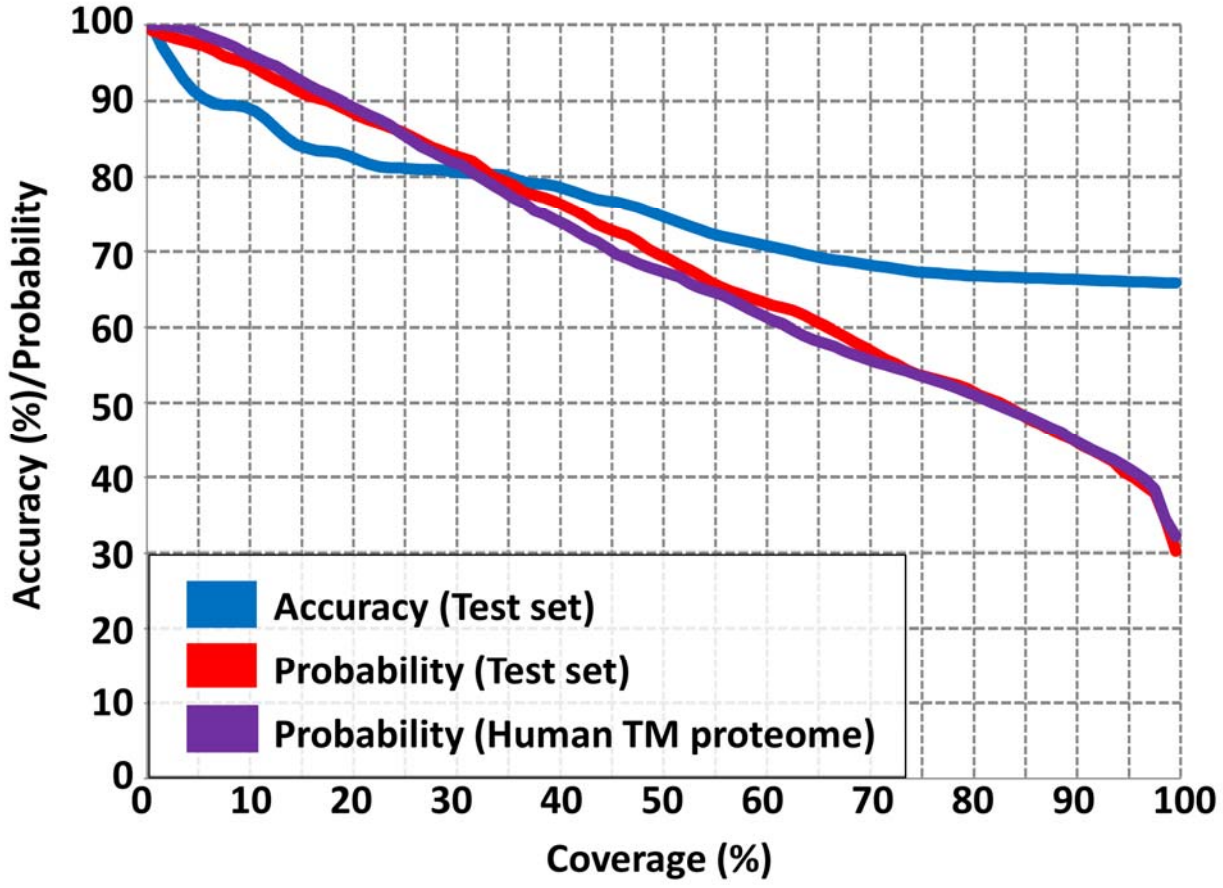

Figure 2. Correlation between the multiclass localization accuracy and probability. Predictions are sorted according to their probability values. Accuracies and the lowest probability measured on the subset of the benchmark set (blue and red, respectively) and probability values on the human transmembrane proteome (purple) are plotted against their rank in the sorted list. The Y-axis shows the coverage of the dataset (predictions above probability threshold divided by the number of the proteins in the benchmark set/TM proteome (coverage)).

Discussion

Examples with O-glycosylation regions governing sorting

While both apical and basolateral sorting can be complex and dependent on large inventory of different motifs, certain common features are already beginning to emerge. One of the most trivial features identified in our current study is the presence of O-glycosylation motifs or even complete regions in apically sorted proteins. For example, Sucrase-isomaltase (SI) is known to be sorted to the apical brush border in small intestine enterocytes, mostly governed by its O-glycosylated, membrane-proximal segment [21]. A similar, glycosylated “stalk” has also been found to govern the apical sorting of mucin1 in Madin-Darby Canine Kidney (MDCK) cells (representing the distal tubules of kidney) [22]. Such more-or-less extensive mucin-like regions also play a role in proper localization of the apical determinant podocalyxin, in addition to an NHERF-protein binding PDZ motif [23]. Although speculative, or results suggest that such regions might form the apical signal in many more proteins, such as maltase-glucoamylase.
Accumulation of S/T rich motifs is likely connected with mucin-type O-glycosylation phenomena intimately connected to a preferentially apical sorting [21,23-24]. These sequences (especially due to their prominent Thr and Pro content) closely resemble the target sites of O-GalNac transferases [25-26]. However, due to the complex, processive nature of GalNac-T enzymes, the exact sequence of O-glycosylation sites is impossible to predict [27-28]. A cautious alignment of these regions reveal 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 3).

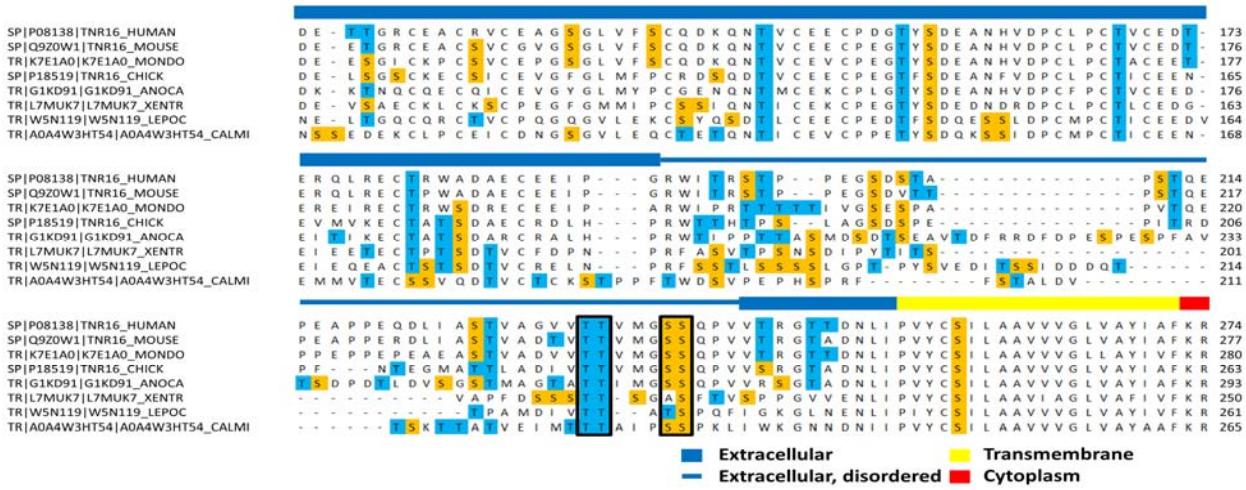

Figure 3: Sequence alignment of the O-glycosylated linker region between the extracellular domains (blue) and the transmembrane helix (yellow) of the low affinity Neurotrophin receptor (p75NTR or TNFR16). Despite excellent architectural homology between vertebrate receptors, only a few Ser-Thr rich repeats 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.

Accuracy of disordered prediction methods is limited in TM proteins [29]. Although extracellular disordered regions are depleted in TM proteins [29], they are definitely present and one important function of them is to serve sites for glycosylation [30] mediating sorting.

Limitation of our method

Despite the simplicity of our approach and its merits, this model also has many shortcomings. One of the most important problems 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 very 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 mildl (e.g. choroid plexus epithelium, with apical 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 become universally applicable.

Furthermore, current analytical methods provide limited information about polarized cells, as they usually characterize individual proteins with immunolocalization or with labeling amino acids selectively on one side of the polarized cell before proteomic analysis. All of these techniques provide a relative measure for the protein abundances in the different compartments of the polarized cell. The main bottleneck of immunolocalization is the limited availability of appropriate antibodies. For example, monoclonal antibodies can be highly specific and may only recognize one epitope, thus in this case we cannot safely assume that they are available in other regions of the polarized cell to the same degree. The most limiting factor in selective labeling that it uses primer amine specific reagents, enabling them to identify only those proteins that have such available regions. In accord, the coverage of TM proteins is relatively low in succeeding proteomic analyses, therefore differences between the two sides based on a limited number of peptides.

Other resources, similar approaches

Recent decades provided several experimentally derived datasets that utilized high-throughput methods to classify localization of human proteins [31]. There are also a number of prediction methods that predict localization information, either the presence of a signal peptide [32], or the exact localization of proteins [33]. Some of these methods were trained on data automatically downloaded from computationally annotated databases, thus any bias in their sources affected 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, this way we provided a very clean training set for our method.

Defining SLiMs is a rather laborious process, both by computationally and experimentally [34]. Dozens of pieces of evidence are required to confirm the functionality of a linear motif, as the information content of the peptide is very low and most of their occurrence in proteins are just

'random' false positive hits. There are several other efforts for the large scale computational identification and analysis of putative linear motifs in membrane proteins [35]. Here we demonstrated that such considerations have a strong predictive power, even if a significant proportion of the identified motifs is likely not functional.

The provided dataset can be directly utilized by experimentalists in the near future. We published a list of proteins that may have not been considered until now as potential targets and can be used to design selective inhibitors.

Methods

Training and testing data

We manually collected 171 apical and 125 basolateral membrane proteins. Although some of these proteins are available in high-throughput sets [31], each of these proteins was manually confirmed to be localized to the apical/basolateral membrane in kidney tissues. Furthermore, we collected 421 plasma membrane proteins from the RBCC database [36] and other sources [31], including our previous experimental pipelines [37–39]. We collected 521 proteins localizing to other membranes (Mitochondrial membrane, Endoplasmic reticulum, Lysosomal membrane etc) from UniProt [40].
We collected ortholog proteins from *Pan Troglodytes, Gorilla Gorilla, Pongo Abelii, Macaca Mulatta, Felis Catus, Canis Familiaris, Equus Caballus, Ovis Aries, Bos Taurus, Oryctolagus Cuniculus, Callithrix Jacchus, Mus Musculus, Rattus Norvegicus,* and *Sus Scrofa* from the OMA database [41]. We aligned the sequences with ClustalOmega [42]. We discarded those alignments, where we found discrepancies in the aligned transmembrane topologies.

Prediction

We randomly selected 100 proteins from each localization group and used them for training and validating the predictions using 10/1 jackknife. Another 25 proteins from each group was used as an independent test set to measure the accuracy of the prediction.
We used CCTOP [43] to predict transmembrane regions. We used IUPred [44] to detect disordered regions, however we masked out those segments that were also included in PFAM domains [45]. We used PlatoLoCo [46] to detect low complexity regions. Amino acid properties were derived from AAIndex [47]. Teiresias [48] 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 checked orthologs and discarded non-conserved hits. 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, similarly as described in the TOPDOM database [49]. Combinations of these features were used as an input to the classical NN. We used 16 input features, 8 hidden layers and stochastic gradient descent algorithm for the NN.

We also prepared a CNN, where each sequence was converted into a 20x20 matrix, and each point in the matrix represents a residue pair. The bottom triangle of the matrix represented extracellular disordered regions, while the top triangle represented cytoplasmic disordered regions. In each sequence, any neighbouring residue increased the point in the matrix by 2, while more distant pairs (up to 3 places) increased the point by 1. Each cell in the matrix was standardized using the same cells across the dataset. Ortholog proteins were also used during the training. The CNN consisted of 2 convolutional layers (filters: 32, dimensions: 3 and filters: 16, dimensions: 2) and a pooling layer (dimensions: 2) with relu activation functions and dropout (0.25). The final activation function was softmax with 4 output neurons.
To combine the results, first a vote was performed on the CNN prediction of different orthologs, using probabilities as weight. Finally, the normalized prediction accuracy of the classical NN and the CNN multiplied with their output probability was used as the final probability value.

Dataset of apical and basolateral transmembrane proteins

We ran the prediction method described in previous sections on the Human Transmembrane Proteome database [50]. We selected the most reliable part of the prediction (above 75% probability, where prediction accuracy is 80%).

References

1. Bryant DM, Mostov KE. From cells to organs: building polarized tissue. Nat Rev Mol Cell Biol. 2008;9: 887–901.

2. Hanukoglu I, Boggula VR, Vaknine H, Sharma S, Kleyman T, Hanukoglu A. Expression of epithelial sodium channel (ENaC) and CFTR in the human epidermis and epidermal appendages. Histochem Cell Biol. 2017;147: 733–748.

3. Inukai K, Shewan AM, Pascoe WS, Katayama S, James DE, Oka Y. Carboxy terminus of glucose transporter 3 contains an apical membrane targeting domain. Mol Endocrinol. 2004;18: 339–349.

4. Riga A, Castiglioni VG, Boxem M. New insights into apical-basal polarization in epithelia. Curr Opin Cell Biol. 2020;62: 1–8.

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

6. Stoops EH, Caplan MJ. Trafficking to the apical and basolateral membranes in polarized epithelial cells. J Am Soc Nephrol. 2014;25: 1375–1386.

7. Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, et al. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev. 2014;114: 6733–6778.

8. Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, et al. Attributes of short linear motifs. Mol Biosyst. 2012;8: 268–281.

9. Farr GA, Hull M, Mellman I, Caplan MJ. Membrane proteins follow multiple pathways to the basolateral cell surface in polarized epithelial cells. J Cell Biol. 2009;186: 269–282.

10. Weisz OA, Rodriguez-Boulan E. Apical trafficking in epithelial cells: signals, clusters and motors. J Cell Sci. 2009;122: 4253–4266.

11. Di Martino R, Sticco L, Luini A. Regulation of cargo export and sorting at the trans-Golgi network. FEBS Lett. 2019;593: 2306–2318.

12. Le Bivic A, Sambuy Y, Patzak A, Patil N, Chao M, Rodriguez-Boulan E. An internal deletion in the cytoplasmic tail reverses the apical localization of human NGF receptor in transfected MDCK cells. J Cell Biol. 1991;115: 607–618.

13. Hunziker W, Fumey C. A di-leucine motif mediates endocytosis and basolateral sorting of macrophage IgG Fc receptors in MDCK cells. EMBO J. 1994;13: 2963–2969.

14. Martín M, Modenutti CP, Peyret V, Geysels RC, Darrouzet E, Pourcher T, et al. A Carboxy-Terminal Monoleucine-Based Motif Participates in the Basolateral Targeting of the Na+/I-Symporter. Endocrinology. 2019;160: 156–168.

15. Evdokimov K, Biswas S, Schledzewski K, Winkler M, Gorzelanny C, Schneider SW, et al. 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. 2016;475: 342–349.

16. Simons K, Ikonen E. Functional rafts in cell membranes. Nature. 1997;387: 569–572.

17. Urquhart P, Pang S, Hooper NM. N-glycans as apical targeting signals in polarized epithelial cells. Biochem Soc Symp. 2005; 39–45.

18. Yeaman C, Le Gall AH, Baldwin AN, Monlauzeur L, Le Bivic A, Rodriguez-Boulan E. The O-glycosylated stalk domain is required for apical sorting of neurotrophin receptors in polarized MDCK cells. J Cell Biol. 1997;139: 929–940.

19. Dunbar LA, Aronson P, Caplan MJ. A transmembrane segment determines the steady-state localization of an ion-transporting adenosine triphosphatase. J Cell Biol. 2000;148: 769–778.

20. Garcia-Gonzalo FR, Reiter JF. Scoring a backstage pass: mechanisms of ciliogenesis and ciliary access. J Cell Biol. 2012;197: 697–709.

21. Jacob R, Alfalah M, Grünberg J, Obendorf M, Naim HY. Structural determinants required for apical sorting of an intestinal brush-border membrane protein. J Biol Chem. 2000;275: 6566–6572.

22. Kinlough CL, Poland PA, Gendler SJ, Mattila PE, Mo D, Weisz OA, et al. Core-glycosylated mucin-like repeats from MUC1 are an apical targeting signal. J Biol Chem. 2011;286: 39072–39081.

23. Yu C-Y, Chen J-Y, Lin Y-Y, Shen K-F, Lin W-L, Chien C-L, et al. A bipartite signal regulates

the faithful delivery of apical domain marker podocalyxin/Gp135. Mol Biol Cell. 2007;18: 1710–1722.

24. Breuza L, Garcia M, Delgrossi M-H, Le Bivic A. Role of the membrane-proximal O-glycosylation site in sorting of the human receptor for neurotrophins to the apical membrane of MDCK cells. Exp Cell Res. 2002;273: 178–186.

25. Fritz TA, Raman J, Tabak LA. Dynamic association between the catalytic and lectin domains of human UDP-GalNAc:polypeptide alpha-N-acetylgalactosaminyltransferase-2. J Biol Chem. 2006;281: 8613–8619.

26. Daniel EJP, Las Rivas M, Lira-Navarrete E, García-García A, Hurtado-Guerrero R, Clausen H, et al. Ser and Thr Acceptor Preferences of the GalNAc-Ts Vary Among Isoenzymes to Modulate Mucin Type O-Glycosylation. Glycobiology. 2020. doi:10.1093/glycob/cwaa036

27. Revoredo L, Wang S, Bennett EP, Clausen H, Moremen KW, Jarvis DL, et al. 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. 2016;26: 360–376.

28. de Las Rivas M, Lira-Navarrete E, Daniel EJP, Compañón I, Coelho H, Diniz A, et al. The interdomain flexible linker of the polypeptide GalNAc transferases dictates their long-range glycosylation preferences. Nat Commun. 2017;8: 1959.

29. Tusnády GE, Dobson L, Tompa P. Disordered regions in transmembrane proteins. Biochim Biophys Acta. 2015;1848: 2839–2848.

30. Goutham S, Kumari I, Pally D, Singh A, Ghosh S, Akhter Y, et al. Mutually exclusive locales for N-linked glycans and disorder in human glycoproteins. Sci Rep. 2020;10: 6040.

31. Caceres PS, Gravotta D, Zager PJ, Dephoure N, Rodriguez-Boulan E. 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. 2019;116: 11796–11805.

32. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019;37: 420–423.

33. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;33: 3387–3395.

34. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, et al. ELM-the eukaryotic linear motif resource in 2020. Nucleic Acids Res. 2020;48: D296–D306.

35. Stavropoulos I, Khaldi N, Davey NE, O'Brien K, Martin F, Shields DC. Protein disorder and short conserved motifs in disordered regions are enriched near the cytoplasmic side of single-pass transmembrane proteins. PLoS One. 2012;7: e44389.

36. Hegedűs T, Chaubey PM, Várady G, Szabó E, Sarankó H, Hofstetter L, et al. Inconsistencies in the red blood cell membrane proteome analysis: generation of a database for research and diagnostic applications. Database . 2015;2015: bav056.

37. Langó T, Pataki ZG, Turiák L, Ács A, Varga JK, Várady G, et al. Partial proteolysis improves the identification of the extracellular segments of transmembrane proteins by surface biotinylation. Sci Rep. 2020;10: 8880.

38. Langó T, Róna G, Hunyadi-Gulyás É, Turiák L, Varga J, Dobson L, et al. Identification of Extracellular Segments by Mass Spectrometry Improves Topology Prediction of Transmembrane Proteins. Sci Rep. 2017;7: 42610.

39. Müller A, Langó T, Turiák L, Ács A, Várady G, Kucsma N, et al. Covalently modified carboxyl side chains on cell surface leads to a novel method toward topology analysis of transmembrane proteins. Sci Rep. 2019;9: 15729.

40. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46: 2699.

41. Altenhoff AM, Glover NM, Train C-M, Kaleb K, Warwick Vesztrocy A, Dylus D, et al. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res. 2018;46: D477–D485.

42. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high□quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011. p. 539. doi:10.1038/msb.2011.75

43. Dobson L, Reményi I, Tusnády GE. CCTOP: a Consensus Constrained TOPology prediction web server. Nucleic Acids Res. 2015;43: W408–12.

44. Mészáros B, Erdos G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018;46: W329–W337.

45. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47: D427–D432.

46. Jarnot P, Ziemska-Legiecka J, Dobson L, Merski M, Mier P, Andrade-Navarro MA, et al. PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. Nucleic Acids Res. 2020;48: W77–W84.

47. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008;36: D202–5.

48. Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics. 1998;14: 55–67.

49. Varga J, Dobson L, Tusnády GE. TOPDOM: database of conservatively located domains and motifs in proteins. Bioinformatics. 2016;32: 2725–2726.

50. Dobson L, Reményi I, Tusnády GE. The human transmembrane proteome. Biology Direct. 2015. doi:10.1186/s13062-015-0061-x