SCIENTIFIC REPORTS
nature research

OPEN

Covalently modified carboxyl side chains on cell surface leads to a novel method toward topology analysis of transmembrane proteins

Anna Müller[1], Tamás Langó[1], Lilla Turiák[2], András Ács[2], György Várady[1], Nóra Kucsma[1], László Drahos[2] & Gábor E. Tusnády[1*]

The research on transmembrane proteins (TMPs) is quite widespread due to their biological importance. Unfortunately, only a little amount of structural data is available of TMPs. Since technical difficulties arise during their high-resolution structure determination, bioinformatics and other experimental approaches are widely used to characterize their low-resolution structure, namely topology. Experimental and computational methods alone are still limited to determine TMP topology, but their combination becomes significant for the production of reliable structural data. By applying amino acid specific membrane-impermeable labelling agents, it is possible to identify the accessible surface of TMPs. Depending on the residue-specific modifications, new extracellular topology data is gathered, allowing the identification of more extracellular segments for TMPs. A new method has been developed for the experimental analysis of TMPs: covalent modification of the carboxyl groups on the accessible cell surface, followed by the isolation and digestion of these proteins. The labelled peptide fragments and their exact modification sites are identified by nanoLC-MS/MS. The determined peptides are mapped to the primary sequences of TMPs and the labelled sites are utilised as extracellular constraints in topology predictions that contribute to the refined low-resolution structure data of these proteins.

Transmembrane proteins (TMPs) having at least one transmembrane segment, are located in the phospholipid bilayers of the cells. TMPs are crucial in several biological pathways such as the extracellular or intracellular transporting or signalling[1–3]. Approximately 55% of the drugs authorised by the Food and Drug Administration interact with TMPs[4] so the 3D structure of these proteins could be necessary for computational drug design. While around 20–30% of the open reading frames of the human genome code TMPs[5–7], only 2% of the previously solved protein structures belong to TMPs[8–11].

The determination of high resolution TMP structures is still complicated, so topology became a tool for representing low resolution data instead of the exact 3D structure[12,13]. Topology is the most frequently used representation of a TMP structure, defining the relative location and orientation of the transmembrane (TM) regions, the extracellular and intracellular loops to the membrane itself and the number of the TM regions.

Topology predictions remain a necessary method for the structure analysis of TMP structures[14]. Earlier predictions are built upon the secondary structure based on the chemical characteristic of the amino acids, searching for hydrophobic regions[15]. As bioinformatics developed, machine learning also appeared in the research toward topology of TMPs[13,16–19]. As supervised machine learning methods always utilises a learning set of data, we cannot avoid miscalculations on unknown protein families[20,21]. Machine learning itself has increased the accuracy of the predictions up to a point but with utilising the experimental topology data, the accuracy of the predictions increases remarkably[12,13,22].

[1]Institute of Enzymology, RCNS, Hungarian Academy of Sciences, Magyar Tudósok krt 2, Budapest, H-1117, Hungary. [2]Institute of Organic Chemistry, RCNS, Hungarian Academy of Sciences, Magyar Tudósok krt 2, Budapest, H-1117, Hungary. *email: tusnady.gabor@ttk.mta.hu

There are plenty of examples how topological data of TMPs can be experimentally determined with or without the modification of the coding sequence of the protein of interest. Several experimental methods were developed for the analysis of the topology of modified proteins such as protein fusion, epitope insertions, glycosylation motif insertion, C- or N-terminal tagging, single amino acid mutagenesis screening combined with different crosslinking agents[23–28], however these procedures are time-consuming and the interpretation of the results are sometimes not straightforward[24]. Moreover, the function of the modified protein is sometimes different compared to the wild-type protein, for example the function of the modified human folate transporter 1 (SLC19A1, S19A1_HUMAN) changed in certain cases[29]. The *in vivo* or *in vitro* translation of the modified TMPs might be extremely difficult[30] so we mainly focus on experiments without modification of the coding sequence. Among others, examining the extracellular region of the native protein through glycosylation sites contributes to the known topology of TMPs[31], especially by creating high-throughput glycosylation data banks (e.g. Cell Surface Protein Atlas)[32]. Partial proteolysis also provides small resolution data based on the known cleavage sites of the applied proteases and these sites can be detected even by the fragments of the examined proteins via SDS-PAGE[33,34]. The locations of endogenous epitopes are also able to provide low-resolution topology data of TMPs according to the applied antibody as for example in the case of wheat Aluminum-activated malate transporter 1 (ALMT1_WHEAT)[35].

In particular, the chemical modifications on the reactive side chains of accessible amino acids make the examination of their relative location to the membrane in a native TMP[36] possible. There are plenty of labelling agents available on the market and most of them are specific for several functional groups. Many crosslinking reactions have already provided distance constraints for the 3D structure determination of proteins based on the length of these spacer arms[37]. Besides intramolecular interactions, intermolecular crosslinking is also available this way[38]. In certain cases, by applying membrane-impermeable agents, it is also possible to provide topology information on TMPs[27,32,39].

The most popular amino acids for these covalent modifications are cysteins and lysines because sulphydrils and primary amines are reactive enough for a one-step modification by an appropriate chemical agent[40,41]. For the modification of sulphydrils, maleimides or pyridyl disulfides are mostly applied and for the primary amines, imido esters or N-hydroxysuccinimide esters are typically utilised[42]. Beside the most reactive side chains, there are two more amino acids whose reactivity is quite satisfying so it is also possible to modify the side chains of aspartic and glutamic acids[43,44] although these carboxyl groups are mostly modified in a two-step reaction[45]. As during the artificial peptide synthesis, the carboxyl group has to be activated before adding the free primary amine to the reaction. Using carbodiimides combined with succinimides in an acidic environment is a popular method for the activation step[46,47], which is followed by the labelling step. The formation of the amide group can only occur at a slightly alkaline pH because the amine group has to be deprotonated (Supplementary Fig. 1). On the other hand, the stability of the activated carboxyl groups incredibly decreases at higher pH[48].

Recently, we have developed an experimental method for the determination of extracellular lysine side chains of TMPs to provide topology data of them that can be utilised by the CCTOP prediction algorithm in order to achieve better prediction accuracies[39]. The experiments of the workflow are based on a method that allows the high throughput and accurate identification of extracellular lysine side chains that were modified with a membrane-impermeable labelling agent. This way, partial labelling of TMPs generated sufficient constraints to significantly increase the reliability and accuracy of topology predictions[39].

On the other hand, labelling the extracellular lysine side chains also has disadvantages. In order to produce the most detectable fragments for the nano liquid chromatography tandem-mass spectrometry (nanoLC-MS/MS), tryptic digestion is applied in most of the proteomic experiments[49–51]. Unfortunately, trypsin enzyme does not recognize the covalently modified lysine side chains[50,52] so a great amount of missed cleavages appear. Furthermore, peptides containing the dedicated covalent modifications cannot be straightforwardly sequenced this way due to their length and the fact that the labelled lysine side chains do not carry the positive charge which is crucial for MS/MS fragmentation[39]. To summarise, the digestion and the labelling are limiting each other so a need arose for a different labelling method.

The aim of this study was to provide an alternative labelling method to avoid the above described difficulties for generating further experimental topology data for the CCTOP algorithm and an accurate labelling method where the targeted amino acid side chains are modified only on one side of the membrane. Here, we targeted the extracellular carboxyl side chains of the TMPs by covalently modifying them with membrane-impermeable activating agents combined with a biotinylating reagent. To the best of our knowledge, only a few publications are connected with the modification of these amino acids[53,54] so the optimisation of these experiments seems to be a novel project for the development of the research toward the better understanding of the topology of TMPs.

Results

Detection of the covalent modification on model protein. We have successfully modified the carboxyl side chains of a model protein (Bovine Serum Albumin, BSA) with a biotin-containing agent (Fig. 1). The semi-quantitative results prove the importance of the activation step before biotinylation. Although we have observed the presence of the modification in the non-activated sample, the reaction was much more effective on the activated carboxyl groups (Fig. 2). Since BSA usually forms homodimers and we applied 98% pure BSA, some nonspecific bands also appeared.

The applied labelling and isolation steps lead to a known covalent modification on the carboxyl side chains. The successfully alkylated groups are modified by +116.040819 Da while the non-alkylated groups by +59.019355 Da. The nanoLC-MS/MS method followed by data analysis allowed the detection of these modifications (Supplementary Tables 5 and 6).

The presence of the Y_7, B_2, B_3, B_4 and B_5 fragments prove the successful modification of the highlighted aspartic acid (E_{226}) by +116 Da (Fig. 3). Altogether 143 modified BSA peptides resulted 29 modified aspartic or glutamic

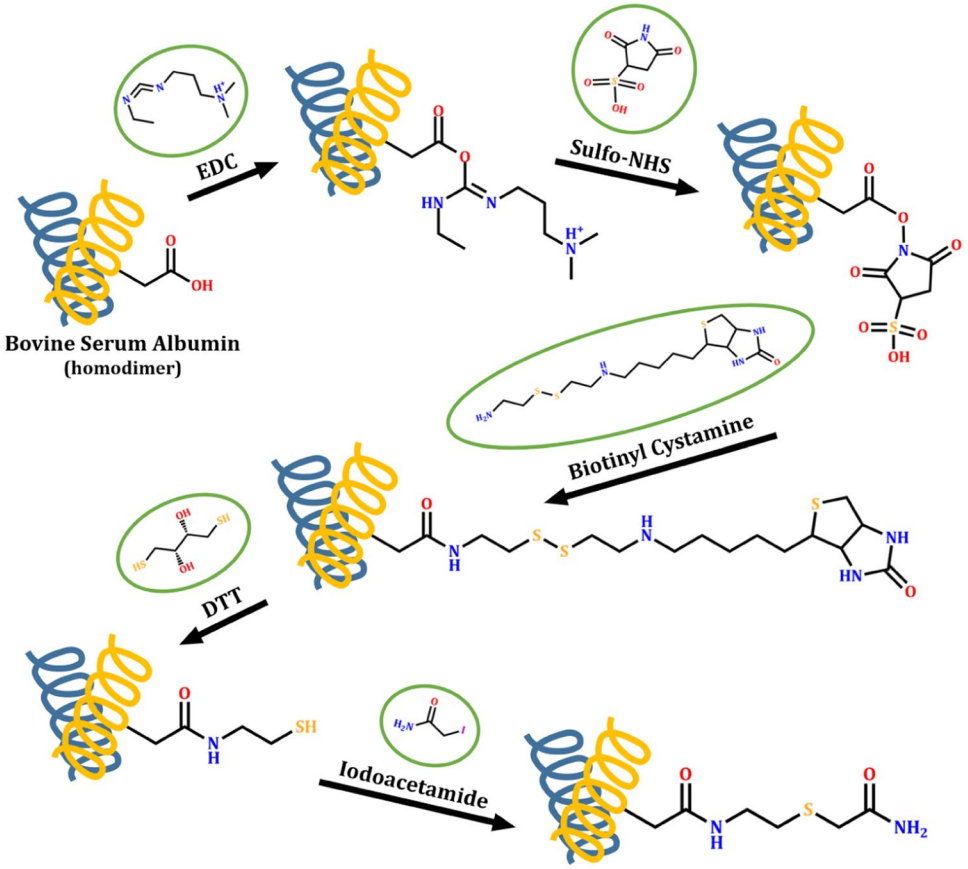

Figure 1. The reaction flow of the applied labelling method. The carboxyl amino acid side chains of Bovine Serum Albumin model protein were activated by EDC and Sulfo-NHS, and then biotinylated by Biotinyl Cystamine. During the nanoLC-MS/MS sample preparation, the disulphide bridges were cleaved by DTT and alkylated by Iodoacetamide reagents.

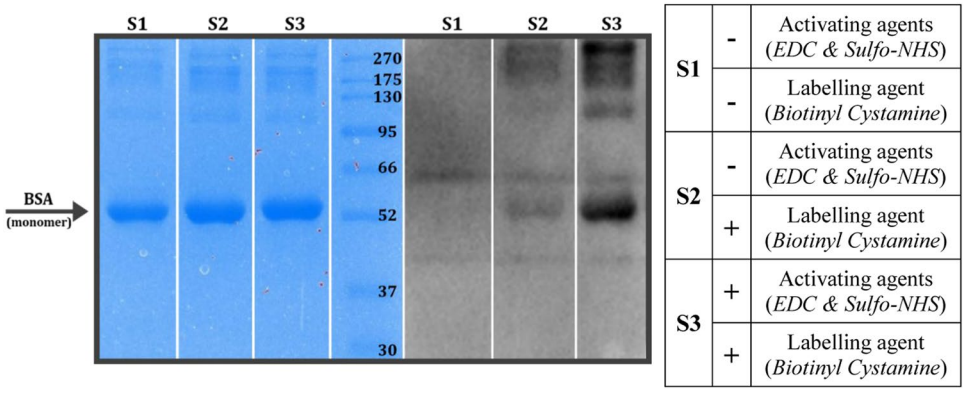

Figure 2. SDA-PAGE and Western Blot analysis of the biotinylated BSA model protein. The protein successfully bound the biotin-residues in both experiments when the labelling agent was applied but the activation step resulted in a higher biotinylated BSA yield. Images were captured by a Bio-Rad ChemiDoc XRS+ Imaging system. The full-length images are presented in Supplementary Fig. 2.

acid positions in the native model protein. The list of the labelled BSA peptides is shown in Supplementary Table 1.

FACS measurement. While extending the biotinylation method for labelling the extracellular carboxyl side chains of TMPs on the surface of living cells, it was essential to detect the effect of the reaction mixtures on living cells. We applied different concentrations of the activating agents then measured biotinylation and cell death by flow cytometry. Both cell death by Propidium Iodide uptake and successful surface labelling by CF488A

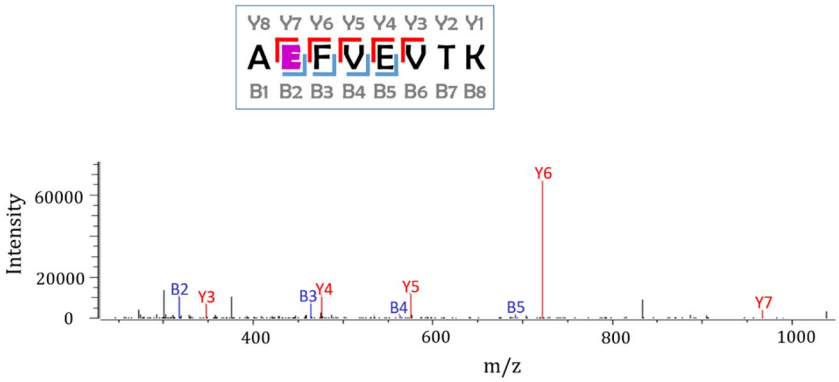

Figure 3. The Intensities of the fragments of a successfully modified BSA peptide. During the applied nanoLC-MS/MS analysis, we were looking for the known fix covalent modifications on the carboxyl amino acid side chains of BSA model protein. Based on peptide sequencing, here we present the +116 Da modification on the highlighted aspartic acid (E). The image was created by Byonic 2.15.7 (Protein Metrics Inc., Cupertino, CA, USA)[67].

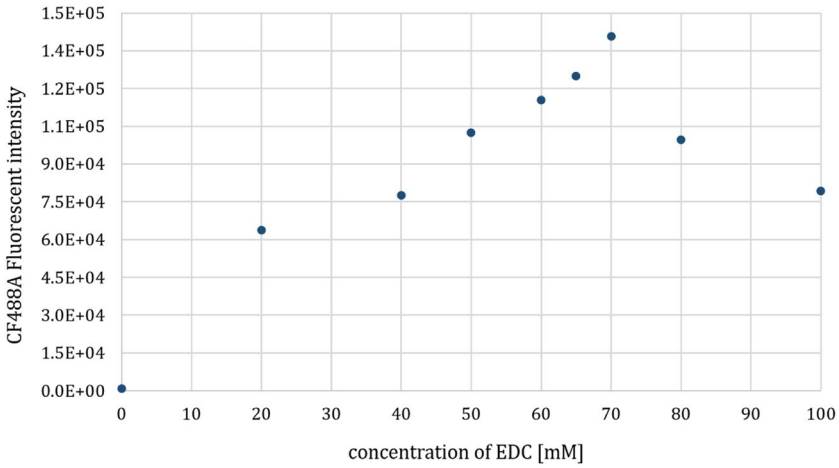

Figure 4. The effect of different activating agent concentrations on cell surface biotinylation. EDC and Sulfo-NHS were applied in a 1:2 molar ratio. According to the fluorescent intensities emitted by CF488A anti-biotin antibody, we found that cell surface biotinylation did not increase when applying the activating agents over the concentrations 70 mM EDC and 140 mM Sulfo-NHS.

fluorescent anti-biotin antibody were measured. First, we examined a wide range of EDC/Sulfo-NHS concentrations and we found a maximal fluorescent intensity of the applied biotin-binding antibody at 70 mM EDC and 140 mM Sulfo-NHS concentrations (Fig. 4 and Supplementary Fig. 4).

According to the Propidium Iodide uptake, the rate of cell death grows by the increase of the concentration of the activating agents but more than 99% of the examined cells were alive even at the highest concentration (Supplementary Fig. 4).

Confocal microscopy. The suspected maximal intensity of cell surface labelling at 70 mM EDC and 140 mM Sulfo-NHS was further examined by confocal microscopy. The CF488A fluorescent labelling of the cell surface was homogenous and there was no signal in the cytoplasm which indicated that the applied reagents in the above described parameters do not impair the integrity of the cells (Fig. 5). The results of the control experiments are presented in Supplementary Fig. 5.

Enrichment of extracellular protein segments. Labelled cells were solubilised, which was followed by the digestion of the membrane preparations. We enriched the biotin-containing peptides via affinity chromatography. The quality of the peptides was tested by blotting the samples onto a PVDF membrane.

The amount of the used neutravidin beads did not restrict the isolation of the total amount of the biotinylated peptides. Comparing the biotin content of the solutions that were taken before and after affinity column, it was clear that all the biotinylated components stayed on the neutravidin beads which means that we used sufficient amount of the neutravidin beads (Supplementary Fig. 6).

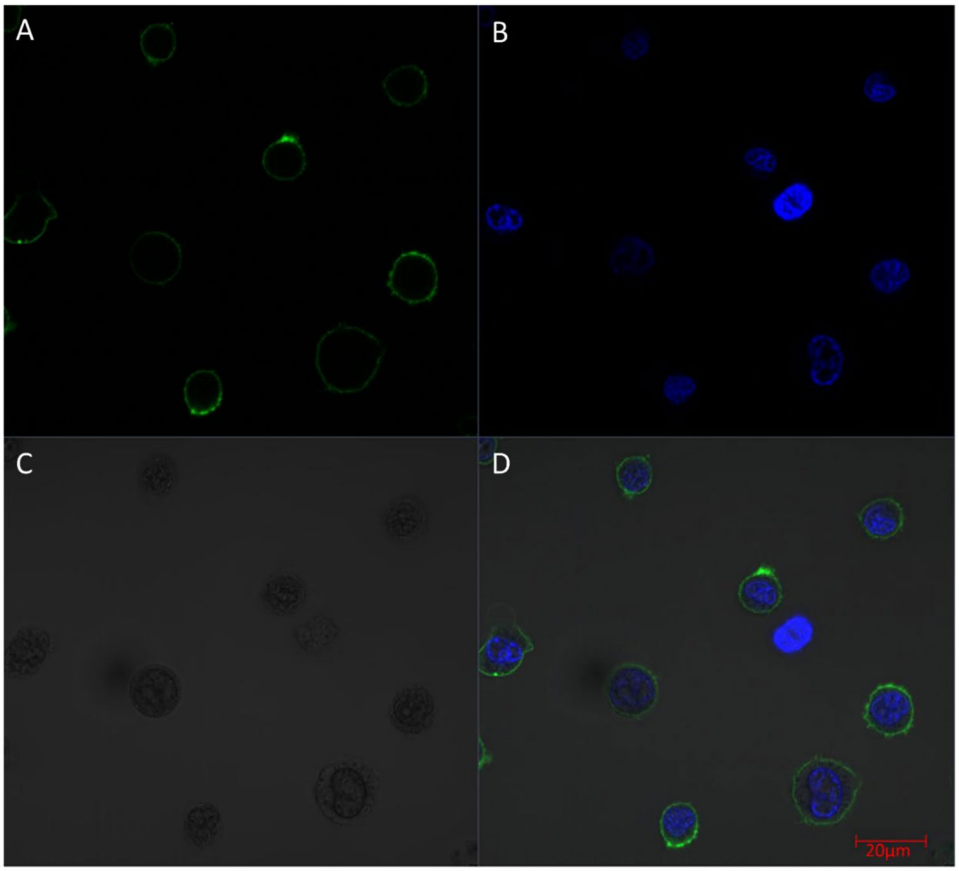

Figure 5. Examining the surface biotinylation of the cells by applying the observed maximal labelling concentrations. The fluorescent signal of the applied dyes and the HL60 cells were detected by confocal microscopy (A: Alexa Fluor 488 conjugated anti-biotin antibody fluorescence, B: Hoechst 33342 DNA dye fluorescence; C: Differential Interference Contrast; D: Merged picture). Scale bar: 20 µm. The images were created by Zeiss ZEN lite software (Carl Zeiss, Oberkochen, Germany).

Covalently labelled peptides were eluted by DTT reducing agent from affinity column then we applied iodoacetamide alkylating agent in order to avoid the aggregation of peptides through disulphide bridges. The method resulted in the covalent modification of extracellular carboxyl groups by +116.040819 or +59.019355 Da.

Identification of the labelled peptides by tandem mass spectrometry and validation of the experimental results. After setting the labelling parameters on BSA model protein and finding the best concentrations for the labelling agents by FACS and confocal microscopy that resulted in high fluorescence intensity on cell surface but left cells intact, we labelled human HL60 cells. The biotinylated cells were lysed, and the membrane preparations were solubilized and digested. Finally, the labelled peptides were isolated on neutravidin column. The list of the tandem mass spectrometry identified peptides carrying the modified carboxyl groups is shown in Supplementary Table 2 (raw ms data tab).

Altogether 1096 peptides containing covalently modified aspartic acid or glutamic acid were detected by nanoLC-MS/MS. The labelled peptides of HL60 cell line from the different Byonic searches were linked to individual proteins, and the sites of modifications in unique proteins were counted. We account only those aspartic or glutamic acids that were identified in at least three independent labelling experiments. This resulted in 135 positions in 38 TMPs (Supplementary Table 2, re-mapped peptides tab).

Discussion

Labelling the extracellular amino acid carboxyl side chains of TMPs in order to generate applicable data for their more accurate topology prediction and structure modelling is a novel tool. We utilised two membrane-impermeable activating agents and one biotinylating reagent in a two-step reaction. At first, the reaction was verified on a single protein (BSA) then extended to the surface of living cells. Both flow cytometry and confocal microscopy measurements confirmed the integrity of the cells. During the experiments, 135 new topological positions were identified for 38 TMPs in HL60 cell line by 16 nanoLC-MS/MS runs and from BSA digestions, 29 amino acid carboxyl side chains were detected from the accessible surface of the protein.

The majority of the identified TMPs (34 out of 38 labelled TMPs) has only one TM segment and a large extracellular domain. 30 of the 34 proteins are Single-pass type I proteins and the other 4 are Single-pass type II proteins. Their extracellular domain contains many glutamic and aspartic acids that are available for the chemical

agents during the cell surface labelling procedure. The existing topology predictions are already accurate for these proteins due to this low structural diversity and the number of already existing experimental results (such as signal peptides, labelled lysine residues or identified glycosylation sites)[32,39,55].

On the other hand, our protocol was also able to label multi-pass TMPs. There is no surprise that the predictions of multi-pass TMPs are often controversial because membrane-embedded parts are more complex, therefore re-entrant loops and interfacial helixes also appear, which could sometimes also mislead the prediction algorithm.

According to the HTP database, the predictions of the two TM-segment Cell cycle control protein 50A (CC50A_HUMAN) are quite synchronous. Unfortunately, there are only a few experimental positions available that could verify these topology predictions[31,56,57] (for details see data in HTP database; HTP id: 002623). Here, we successfully labelled E_{224} aspartic acid that is localised between two transmembrane segments of this protein. The extracellular location of their interconnecting loop has already been confirmed by three independent experiments. Thus, previous experimentally determined extracellular data and our labelled position validate each other.

Leukocyte surface antigen CD47 protein (CD47_HUMAN) is consistently predicted to have five TM segments. On the other hand, the extracellular and intracellular segments are controversial based on existing predictions as shown in Supplementary Fig. 7. Previous experiments have already provided some extracellular positions in the first extracellular domain on the protein[31,56,58] so the extracellular location of the labelled E_{122} aspartic acid is verified based on other experiments (HTP id: 002114). Using all the available information contributes to a more accurate topology prediction of this protein.

In case of the Neutral amino acid transporter B(0) protein (AAAT_HUMAN), different topology prediction algorithms provided various topology for this transporter (HTP id: 001260). Even the number of predicted membrane segments is different in some protein regions by different predictions. Our carboxyl-labelling method identified three different positions that prove the extracellular location of a loop region of the protein, which is consistent with the previously determined N-linked glycosylation site from this part[56]. Additionally, the 3D structure of this protein was solved by cryo-electron microscopy[59] lately so the results of our experiments were also examined by using this data. The structure of this protein (PDB code: 6GCT) suggested that all three labelled positions are localised in the extracellular region (Supplementary Fig. 8) so the correctness of our data was confirmed for the second time regarding this protein.

Considering Equilibrative nucleoside transporter 1 (S29A1_HUMAN) protein, our experiments resulted in one labelled small extracellular loop, where the modified E_{325} amino acid is relatively close to the membrane region. This way, we were also able to label small extracellular loops close to the membrane despite of the possible spherical effects and to the best of our knowledge, this extracellular segment has not been experimentally verified yet (HTP id: 002245). Furthermore, this modified amino acid confirms that it is also possible to identify labelled positions from smaller extracellular loop regions as opposed to previous examples.

32 out of the 38 identified TMPs have been investigated experimentally so far. According to these experiments, all the labelled carboxyl groups are localised in the extra-cytosolic region. Clustering proteins that contain at least one peptide with modified carboxyl group and identified at least three times independently resulted in 95 TMPs altogether.

Today, many of the available high-throughput method identified glycosylation sites are based on the Cell Surface Protein Atlas (CSPA) where extracellular glycosylation positions were determined by LC-MS/MS[32]. The CSPA dataset also contains information about HL60 cell line that was used in our experiments too. 194 proteins were identified totally from HL60 cell line by their cell surface capture technology. According to the CCTOP algorithm, 172 out of them are TMPs. Considering all these TMPs, only 21 of our carboxyl-labelled TMPs are already shown in CSPA. The extracellular regions of the 17 new TMPs contain much more aspartic and glutamic acids than glycosylation sites so our labelling method might have detected them easier.

Additionally, the +0.984 Da modification resulted on the asparagine amino acids by deglycosylation with PNGase F enzyme sometimes cannot be unambiguously detected by LC-MS/MS because the deamidation can also spontaneously occur on asparagine residues[60]. This is the reason why the identified extracellular positions are less reliable in CSPA compared to our carboxyl-labelling experiment.

Regarding the already existing topology predictions and experiments, all of the labelled carboxyl groups are located in the extracellular region, therefore the data produced by carboxyl group labelling is 100% accurate. Supplementary Table 3 contains the predicted topology of the modified TMPs including the modified aspartic and glutamic acids from our experiments and also the already existing other experimental results. Interestingly, the rate of labelled aspartic and glutamic acids is different. Glutamic acid side chains were modified around twice as often as aspartic acid side chains because the longer amino acid side chains could be more accessible to the applied chemical agents.

Characterising the Custom modifications of +59.019355 and +116.040819 Da, we have to evaluate that while 2% of the modified HL60 peptides contained the non-alkylated +59.019355 Da modification, applying the "labelled position per 3 peptide" filter, 100% of the analytically accepted modifications on TMPs are +116.040819 Da.

Here, we would like to take the opportunity to compare the labelling yield of the carboxyl side chain targeted method to the lysine-labelling procedure. Our previous publication was built upon the analysis of three cell lines including HL60[39]. In case of this acute myeloid leukemia cell line, 371 positions were detected in 114 TMPs via the lysine-labelling protocol by nanoLC-MS/MS on a Bruker Maxis II ETD Q-TOF mass spectrometer in 12 measurements. There are 13 TMPs that were detected by both protocols and the carboxyl-labelling method resulted in 25 new TMPs. Although 24 out of these 25 TMPs show homology with the previously lysine-detected TMPs, Equilibrative nucleoside transporter 1 (S29A1_HUMAN) protein was only captured by the carboxyl-labelling method. According to CCTOP algorithm, this protein contains 3 extracellular lysine residues beside 11 aspartic or glutamic acids. Even if there is a quite reasonable difference between the TMP-yield of the two experimental

protocols, less false labelling appeared in our newly presented carboxyl targeted experiments due to the selective labelling and the several washing steps.

Beside 38 TMPs, we also labelled 5 non-TM proteins. According to the Peripheral Membrane Protein database[61] that contains proteins attached to the plasma membrane, 2 out of these 5 proteins are in the database, furthermore they are located on the extracellular region of the cell (PDIA1_HUMAN & B2MG_HUMAN) so our protocol made it possible to label peripheral membrane proteins on the cell surface too.

To summarise the advance of the optimised carboxyl-labelling experiments described here, we can state that providing topology data for TMPs require 3–4 weeks which is a shorter period than by previously existing techniques that mainly characterised a single protein or a smaller protein family. Moreover, culturing parallel cell lines from different origin could contribute to topology data of many hundreds of TMPs which could exponentially increase the number of known topologies.

Beside labelling and isolating living cell surface, the covalent modification of the carboxyl side chains has other possibilities. Among others, our developed labelling protocol will make protein carboxyl residues available targets for other conjugation techniques. For example, antibody-drug conjugations based on antibody-antigen interactions also require free functional groups on the heavy chain of the antibody[62]. In this field of study, cysteines are mostly targeted through their sulphydril-reactions, but it also has several disadvantages. The native protein structure is quite sensitive and the complex dissociates easily because the heavy and light chains of the antibody bind each other through disulphide brides of the cysteine residues[63]. The modification of amino acid residues in novel proteins usually target cysteine or lysine residues[64]. Here, we prove that aspartic and glutamic acids could also become a satisfying target for several conjugation experiments.

We would like to highlight the possibility that the carboxyl-labelling protocol will soon become a tool for cell surface isolation either used individually or in combination with the lysine-labelling method. Although carboxyl-activating reagents were previously used for other techniques[36], it is clear that protonated and covalently modified lysine side chains do not bind the activated carboxyl groups. The applied washing steps and pH values allowed our selective conjugation reaction. In consideration of the above described advantages, this optimised carboxyl-labelling protocol could be the basis for a new cell surface-isolation kit in the near future.

Methods

For the unity of the text, the details of the applied materials are indicated in Supplementary Table 7 and the applied instruments and machines in Supplementary Table 8.

Labelling the carboxyl groups of a model protein. In preliminary experiments, we have applied a model serum protein, Bovine Serum Albumin (100 µg BSA). At first, the carboxyl groups were modified either in a one-step or in a two-step reaction. The one-step reaction contained only biotinylation by 1 mM Biotinyl Cystamine. During the two-step reaction, 3 mM 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) and 6 mM N-hydroxysulfosuccinimide (Sulfo-NHS) were applied in acidic environment (200 mM MES, pH = 5.0) for the activation followed by the biotinylation step with 1 mM Biotinyl Cystamine.

Detection of the covalent modification on model protein. For detecting the labelling efficiencies, we applied SDS-PAGE and Western Blot. The model protein samples (5 µg) were loaded on a 12% SDS-PAGE and stained with Coomassie Brilliant Blue. For the semi-quantitative Western Blot, HRP-conjugated Streptavidin and Immobilon Western Chemiluminescent HRP Substrate were applied.

The biotinylated protein samples (50 µg BSA) were incubated at 37 °C for 16 hours with MS-grade trypsin in a 1:100 (w/w) protease:protein mass ratio in the presence of 0.1% (w/v) Rapigest surfactant. The digestion was inactivated by heat (95 °C for 10 min) and 1 mM TLCK inhibitor (room temperature for 30 min). For the cleavage of the disulphide bound, the samples were incubated with 50 mM NH_4HCO_3 (pH = 8.0) buffer containing 10 mM 1,4-dithiothreitol (DTT) for 1 hour at 37 °C. In order to avoid further disulphide-bridge formation, free sulfhydryls were alkylated with 22 mM iodoacetamide.

The peptides were then purified on C18 spin columns and diluted in 20 µl loading buffer containing 2% acetonitrile and 0,1% formic acid. 6 µl of the solution was injected onto the nanoLC-MS/MS.

Cell cultures and isolation. HL60 – which is an acute promyelocytic leukemia cell line - cells were obtained from American Type Culture Collection and were cultured in RPMI supplemented with 50 µg/ml Penicillin-Streptomycin and 10% Fetal Bovine Serum (FBS) in a humidified 37°C incubator with 5% CO_2 atmosphere. All media were sterile filtered by bottle-top vacuum filter systems (0.2 µm pore size). HL60 cells were collected by centrifugation at 300 g for 5 minutes at 4 °C and washed with PBS (pH = 7.3; 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4 and 1.8 mM KH_2PO_4) three times. During the last washing step, we applied 4 mM iodoacetamide alkylating agent in order to avoid the production of "piggyback" peptides on free sulfhydryl groups ("piggyback" peptides are cysteine containing peptides that can bind to each other through disulphide-bridge)[65]. For MS analysis 10^8 HL60 cells were used in each experiment.

Labelling the carboxyl groups on the surface of living cells. For each sample, we used $2*10^6$ HL60 cells that were cultured and isolated as previously described. For the activation step, we applied the agents in a 1:2 molar ratio in a range of 0–150 mM for EDC and 0–300 mM for Sulfo-NHS in an activating buffer (pH = 5; 100 mM MES, 150 mM NaCl) at 4 °C for 15 minutes. It was followed by a washing step, where the activating buffer was applied at a higher pH (pH = 6.5; 100 mM MES 150 mM NaCl). The biotinylation step was performed using 1 mM Biotinyl Cystamine in PBS (pH = 8.0). The labelling reaction was stopped by adding 100 mM glycine (in PBS, pH = 7.3). The labelling efficiency was first detected by Dot Blot technique (Supplementary Fig. 3).

FACS measurement. The labelled cells were washed with PBS (pH = 7.3) and incubated with PBS containing 2% (w/v %) BSA before applying CF488A fluorescent anti-biotin antibody (200x diluted) and Propidium Iodide DNA-dye (1000x diluted). The measurements were conducted by FACS Attune Acoustic Focusing Cytometer.

Confocal microscopy. For the maximal fluorescence anti-biotin antibody intensity, we applied 70 mM EDC and 140 mM Sulfo-NHS during the activation step. Each sample was prepared as described in the flow cytometry protocol except for the DNA dying step. Instead of Propidium Iodide, Hoechst 33342 DNA-dye (10000x diluted) was applied. The samples were measured by Zeiss LSCM 710 Confocal Microscope (objective: 63x NA = 1.4 Plan Apo).

Preparation of the cells for MS analysis. For mass spectrometry analysis, cell surface biotinylation was performed using the optimal 70 mM EDC and 140 mM Sulfo-NHS concentrations under the same activation conditions as mentioned above. The washing step, biotinylation and quenching were applied as previously described. Here, none of the fluorescent antibodies were utilised.

Membrane preparation. We used the protocol described in our previous work (Lango *et al.*)[39] with a few modifications. Shortly, a hypotonic lysis buffer (pH = 7.4, supplemented with 10 mM iodoacetamide) was applied for the lysis of the carboxyl-labelled HL60 cells a 4 °C for 10 minutes. Then the cells were disrupted by micro-pestle 40 times and we passed the samples 20 times through a needle (26 gauge). Cell lysate was centrifuged at 1700 g for 10 minutes at 4 °C for separating the cell debris and the nuclei. The supernatant was further centrifuged at 40000 rpm (in a Beckmann 70.1 Ti rotor) for 1 hour at 4 °C (by a Beckman Ultracentrifuge) for collecting biotinylated membrane fraction. Pellets were purified with washing buffer (pH = 7.7) and were centrifuged at 40000 rpm for a further 1 hour at 4 °C then resuspended in the washing buffer. We applied the method of Lowry *et al.*[66] for the determination of the protein concentration in the membrane fraction using a standard stock solution of BSA.

Membrane protein solubilisation and digestion. The solubilisation and digestion of the membrane proteins were similar to Lango *et al.*[39]. We applied a slightly alkaline buffer (100 mM NH_4HCO_3; pH = 8.0) supplemented with 0.1% (w/v) Rapigest surfactant and 1.2 mM iodoacetamide. The solution also contained 1.2 mM 2,2′-thiodiethanol in order to prevent overalkylation of proteins and peptides during the digestion process. We also applied sonication to assist the solubilisation before incubating the samples for 30 minutes on ice. The suspension was treated with 500 units of PNGaseF and 60 units of α2-3,6,8,9 Neuraminidase A for 2 hours at 37 °C before adding trypsin in a 1:100 (w/w) protease:protein mass ratio. The samples were incubated at 37 °C for 16 hours then heat inactivation (at 95 °C for 10 min) was applied, finally 1 mM TLCK trypsin inhibitor was added to the solution for stopping the enzymatic digestion.

Biotinylated peptide isolation. For the precipitation of the biotinylated peptides, high capacity neutravidin agarose resin was used as described in Lango *et al.*[39]. First, we monitored the binding capacity of the neutravidin columns by Dot Blot technique in case of the carboxyl-labelled samples (Supplementary Fig. 3). The biotinylated components of the solution were isolated on equilibrated neutravidin agarose resin (150 µl, 1 hour, room temperature). In order to reduce the number of nonspecific peptides or contaminants, we washed the columns extensively: 3 ml of each buffer was applied: first 50 mM NH_4HCO_3 (pH = 8.0), then 5 M NaCl in PBS, followed by 50 mM NH_4HCO_3 (pH = 8.0), 100 mM NaHCO3 (pH = 10.0) and hot (60 °C) 50 mM NH_4HCO_3 (pH = 8.0). Before the last washing step, we transferred the agarose resin into a new spin column. 50 mM NH_4HCO_3 (pH = 8.0) buffer supplemented with 10 mM DTT was applied for the elution of the biotinylated peptides from the immobilised neutravidin column. The samples were incubated for 1 hour at 37 °C before adding 22 mM iodoacetamide alkylating agent for preventing the formation of further disulphide bonds between the free sulfhydryl groups.

Mass spectrometry analysis and peptide identification. The desalted samples were analysed by nanoLC-MS/MS similarly to Lango *et al.*[39]. The mass spectrometer was a Bruker Maxis II ETD Q-TOF, the ionization source was a CaptiveSpray nanoBooster source connected to a nanoLC (Dionex Ultimate 3000 NanoLC System). For peptide trapping, an Acclaim PepMap100 C18 Nano-Trap column (5 µm, 100 Å, 100 µm × 20 mm) was applied before separating them online by a Waters Acquity UPLC M-Class Peptide BEH C18 column (25 cm, 1.7 µm particle size). Components of the samples were analysed during 90 min gradient elution (4–50% eluent made of acetonitrile and 0.1% formic acid). The time of the MS measurements was fixed in a cycle of 2.5 sec and for generating the MS spectra, 3 Hz was set in the 150–2200 m/z mass range. For abundant precursors, CID was performed at 16 Hz and for low abundance, at 4 Hz. We used Bruker Compass DataAnalysis software 4.3 for recalibration of the data for the internal standard. MS/MS peak list was generated by ProteinScape software 3.1 and the labelled peptides were identified by Byonic 2.15.7 software. Supplementary Table 4 shows the parameters of the Byonic search engine.

Processing of the MS results. The experiments on the nanoLC-MS/MS method identified carboxyl-labelled peptides belonging to distinct proteins. While evaluating the results of the Byonic Search Engine, we set the LogProb cut off value greater or equal to 2.00. This way, the failure rate of the results (FDR) is only 1%.

The identified peptides were searched again by blastp on the human proteome in order to identify all proteins that contained similar peptides and could not be differentiated based on nanoLC-MS/MS analysis alone. For further selection, only those proteins were considered that contained at least one peptide with modified carboxyl

group and identified at least three times independently. Proteins that share peptides with significant sequence homologies were grouped into clusters. Topologies of TMPs were predicted by CCTOP algorithm. For evaluating topological localisation of labelled carboxyl groups, we used former results of experiments collected in the TOPDB database.

Received: 31 July 2019; Accepted: 5 October 2019;

Published online: 31 October 2019

References

1. Neer, E. J. & Clapham, D. E. Roles of G protein subunits in transmembrane signalling. *Nature* **333**, 129–134 (1988).
2. Acquati, F. *et al*. The gene encoding DRAP (BACE2), a glycosylated transmembrane protein of the aspartic protease family, maps to the Down critical region. *FEBS Lett.* **468**, 59–64 (2000).
3. Saier, M. H., Tran, C. V. & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.* **34**, D181–6 (2006).
4. Uhlén, M. *et al*. Tissue-based map of the human proteome, https://doi.org/10.1126/science.1260419 (2015).
5. Wallin, E. & von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7**, 1029–38 (1998).
6. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *J. Mol. Biol.* **305**, 567–580 (2001).
7. Dobson, L., Reményi, I. & Tusnády, G. E. The human transmembrane proteome. *Biol. Direct* **10**, 31 (2015).
8. Sussman, J. L. *et al*. Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54**, 1078–1084 (1998).
9. Tusnady, G. E., Dosztanyi, Z. & Simon, I. Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics* **20**, 2964–2972 (2004).
10. Kozma, D., Simon, I. & Tusnády, G. E. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.* **41**, D524–D529 (2012).
11. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* **40**, D370–D376 (2012).
12. Dobson, L., Langó, T., Reményi, I. & Tusnády, G. E. Expediting topology data gathering for the TOPDB database. *Nucleic Acids Res.* **43**, D283–9 (2015).
13. Dobson, L., Reményi, I. & Tusnády, G. E. CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.* **43**, W408–W412 (2015).
14. Wang, H., He, Z., Zhang, C., Zhang, L. & Xu, D. Transmembrane Protein Alignment and Fold Recognition Based on Predicted Topology. *PLoS One* **8**, e69744 (2013).
15. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–32 (1982).
16. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–80 (2001).
17. Tusnady, G. E. & Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17**, 849–850 (2001).
18. Kall, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. *Nucleic Acids Res.* **35**, W429–W432 (2007).
19. Viklund, H. & Elofsson, A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* **24**, 1662–1668 (2008).
20. Tusnády, G. E. & Simon, I. Topology prediction of helical transmembrane proteins: how far have we reached? *Curr. Protein Pept. Sci.* **11**, 550–61 (2010).
21. Rost, B., Sander, C., Casadio, R. & Fariselli, P. Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**, 521–533 (2008).
22. Rapp, M. *et al*. Experimentally based topology models for E. coli inner membrane proteins. *Protein Sci.* **13**, 937–945 (2004).
23. Manoil, C. & Beckwith, J. A genetic approach to analyzing membrane protein topology. *Science (80-.).* **233**, 1403–1408 (1986).
24. Salaün, C., Rodrigues, P. & Heard, J. M. Transmembrane topology of PiT-2, a phosphate transporter-retrovirus receptor. *J. Virol.* **75**, 5584–92 (2001).
25. Lorenz, H., Hailey, D. W. & Lippincott-Schwartz, J. Fluorescence protease protection of GFP chimeras to reveal protein topology and subcellular localization. *Nat. Methods* **3**, 205–10 (2006).
26. Wang, H. *et al*. Membrane topology of the human breast cancer resistance protein (BCRP/ABCG2) determined by epitope insertion and immunofluorescence. *Biochemistry* **47**, 13778–87 (2008).
27. Feramisco, J. D., Goldstein, J. L. & Brown, M. S. Membrane topology of human insig-1, a protein regulator of lipid synthesis. *J. Biol. Chem.* **279**, 8487–96 (2004).
28. Hong, M., Tanaka, K., Pan, Z., Ma, J. & You, G. Determination of the external loops and the cellular orientation of the N- and the C-termini of the human organic anion transporter hOAT1. *Biochem. J.* **401**, 515–20 (2007).
29. Liu, X. Y. & Matherly, L. H. Analysis of membrane topology of the human reduced folate carrier protein by hemagglutinin epitope insertion and scanning glycosylation insertion mutagenesis. *Biochim. Biophys. Acta* **1564**, 333–42 (2002).
30. Tate, C. G. Overexpression of mammalian integral membrane proteins for structural studies. *FEBS Lett.* **504**, 94–98 (2001).
31. Chen, R. *et al*. Glycoproteomics analysis of human liver tissue by combination of multiple enzyme digestion and hydrazide chemistry. *J. Proteome Res.* **8**, 651–61 (2009).
32. Bausch-Fluck, D. *et al*. A mass spectrometric-derived cell surface protein atlas. *PLoS One* **10**, e0121314 (2015).
33. Levy, S., Nguyen, V. Q., Andria, M. L. & Takahashi, S. Structure and membrane topology of TAPA-1. *J. Biol. Chem.* **266**, 14597–602 (1991).
34. Blodgett, D. M., Graybill, C. & Carruthers, A. Analysis of glucose transporter topology and structural dynamics. *J. Biol. Chem.* **283**, 36416–24 (2008).
35. Motoda, H. *et al*. The Membrane Topology of ALMT1, an Aluminum-Activated Malate Transport Protein in Wheat (Triticum aestivum). *Plant Signal. Behav.* **2**, 467–72 (2007).
36. Debelyy, M. O., Waridel, P., Quadroni, M., Schneiter, R. & Conzelmann, A. Chemical crosslinking and mass spectrometry to elucidate the topology of integral membrane proteins. *PLoS One* **12**, e0186840 (2017).
37. Leitner, A. *et al*. Chemical cross-linking/mass spectrometry targeting acidic residues in proteins and protein complexes. *Proc. Natl. Acad. Sci.* **111**, 9455–9460 (2014).
38. Mendoza, V. L. & Vachet, R. W. Probing protein structure by amino acid-specific covalent labeling and mass spectrometry. *Mass Spectrom. Rev.* **28**, 785–815.
39. Langó, T. *et al*. Identification of Extracellular Segments by Mass Spectrometry Improves Topology Prediction of Transmembrane Proteins. *Sci. Rep.* **7**, 42610 (2017).

40. Bai, X.-Y. *et al*. Membrane topology structure of human high-affinity, sodium-dependent dicarboxylate transporter. *FASEB J.* **21**, 2409–2417 (2007).
41. Roesli, C., Mumprecht, V., Neri, D. & Detmar, M. Identification of the surface-accessible, lineage-specific vascular proteome by two-dimensional peptide mapping. *FASEB J.* **22**, 1933–1944 (2008).
42. Back, J. W., de Jong, L., Muijsers, A. O. & de Koster, C. G. Chemical cross-linking and mass spectrometry for protein structural modeling. *J. Mol. Biol.* **331**, 303–13 (2003).
43. McGrath, N. A., Andersen, K. A., Davis, A. K. F., Lomax, J. E. & Raines, R. T. Diazo compounds for the bioreversible esterification of proteins. *Chem. Sci.* **6**, 752–755 (2015).
44. Díaz-Rodríguez, A. & Davis, B. G. Chemical modification in the creation of novel biocatalysts. *Curr. Opin. Chem. Biol.* **15**, 211–219 (2011).
45. Schanté, C. E., Zuber, G., Herlin, C. & Vandamme, T. F. Chemical modifications of hyaluronic acid for the synthesis of derivatives for a broad range of biomedical applications. *Carbohydr. Polym.* **85**, 469–489 (2011).
46. Sinz, A. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *J. Mass Spectrom.* **38**, 1225–1237 (2003).
47. Akhshabi, S., Biazar, E., Singh, V., Heidari Keshel, S. & Geetha, N. The effect of the carbodiimide cross-linker on the structural and biocompatibility properties of collagen-chondroitin sulfate electrospun mat. *Int. J. Nanomedicine* **13**, 4405–4416 (2018).
48. Lim, C. Y. *et al*. Succinimidyl Ester Surface Chemistry: Implications of the Competition between Aminolysis and Hydrolysis on Covalent Protein Immobilization. *Langmuir* **30**, 12868–12878 (2014).
49. Olsen, J. V., Ong, S.-E. & Mann, M. Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. *Mol. Cell. Proteomics* **3**, 608–614 (2004).
50. Rauh, M. LC–MS/MS for protein and peptide quantification in clinical chemistry. *J. Chromatogr. B* **883–884**, 59–67 (2012).
51. Atacan, K., Çakıroğlu, B. & Özacar, M. Efficient protein digestion using immobilized trypsin onto tannin modified Fe_3O_4 magnetic nanoparticles. *Colloids Surfaces B Biointerfaces* **156**, 9–18 (2017).
52. Zee, B. M. & Garcia, B. A. Discovery of lysine post-translational modifications through mass spectrometric detection. *Essays Biochem.* **52**, 147–63 (2012).
53. Bronfman, F. C., Tcherpakov, M., Jovin, T. M. & Fainzilber, M. Ligand-induced internalization of the p75 neurotrophin receptor: a slow route to the signaling endosome. *J. Neurosci.* **23**, 3209–20 (2003).
54. Ezgi, N., Küçük, Ö., Tan, E., Mitchison, T. & Özlü, N. Labeling Carboxyl Groups of Surface Exposed Proteins Provides an Orthogonal Approach for Cell Surface Isolation (2018).
55. Li, Y. *et al*. Sensitive profiling of cell surface proteome by using an optimized biotinylation method. *J. Proteomics* **196**, 33–41 (2019).
56. Wollscheid, B. *et al*. Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat. Biotechnol.* **27**, 378–86 (2009).
57. Vakhrushev, S. Y. *et al*. Enhanced mass spectrometric mapping of the human GalNAc-type O-glycoproteome with SimpleCells. *Mol. Cell. Proteomics* **12**, 932–44 (2013).
58. Hatherley, D. *et al*. Paired Receptor Specificity Explained by Structures of Signal Regulatory Proteins Alone and Complexed with CD47. *Mol. Cell* **31**, 266–277 (2008).
59. Garaeva, A. A. *et al*. Cryo-EM structure of the human neutral amino acid transporter ASCT2. *Nat. Struct. Mol. Biol.* **25**, 515–521 (2018).
60. Palmisano, G., Melo-Braga, M. N., Engholm-Keller, K., Parker, B. L. & Larsen, M. R. Chemical Deamidation: A Common Pitfall in Large-Scale N-Linked Glycoproteomic Mass Spectrometry-Based Analyses. *J. Proteome Res.* **11**, 1949–1957 (2012).
61. Nastou, K. C., Tsaousis, G. N., Hamodrakas, S. J. & Iconomidou, V. A. PerMemDB: a database for eukaryotic peripheral membrane proteins. *bioRxiv* 531541, https://doi.org/10.1101/531541 (2019).
62. Tsuchikama, K. & An, Z. Antibody-drug conjugates: recent advances in conjugation and linker chemistries. *Protein Cell* **9**, 33–46 (2018).
63. Sun, M. M. C. *et al*. Reduction−Alkylation Strategies for the Modification of Specific Monoclonal Antibody Disulfides. *Bioconjug. Chem.* **16**, 1282–1290 (2005).
64. Lewis Phillips, G. D. *et al*. Targeting HER2-Positive Breast Cancer with Trastuzumab-DM1, an Antibody-Cytotoxic Drug Conjugate. *Cancer Res.* **68**, 9280–9290 (2008).
65. Hofmann, A. *et al*. Proteomic cell surface phenotyping of differentiating acute myeloid leukemia cells. *Blood* **116**, e26–e34 (2010).
66. Lowry, O. H., Rosebrough, N. J., Farr, A. L. & Randall, R. J. Protein measurement with the Folin phenol reagent. *J. Biol. Chem.* **193**, 265–75 (1951).
67. Bern, M., Kil, Y. J. & Becker, C. Byonic: Advanced Peptide and Protein Identification Software. in *Current Protocols in Bioinformatics*, https://doi.org/10.1002/0471250953.bi1320s40 (John Wiley & Sons, Inc., 2012).

Acknowledgements

This work was supported by grants from Hungarian Research and Developments Fund [OTKA K119287 and K125607] and from Research and Technology Innovation Fund [VKSZ-12–1–2013–0001]. G.E.T. [LP2012-35] was supported by the Momentum Grant of the Hungarian Academy of Sciences.

Author contributions

A.M.: performed all experiments, wrote the manuscript; T.L.: performed experiments, wrote the manuscript; L.T.: performed the MS analysis on Bruker Maxis II Q-TOF, processed MS data and wrote the manuscript; A.Á.: performed the MS analysis on Bruker Maxis II Q-TOF, processed MS data; G.V.: performed flow cytometry measurements; N.K.: performed confocal microscopy measurements; L.D.: supervised MS analysis; G.E.T.: supervised experiments, performed computational biology studies and wrote the manuscripts. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-52188-4.

Correspondence and requests for materials should be addressed to G.E.T.

Reprints and permissions information is available at www.nature.com/reprints.