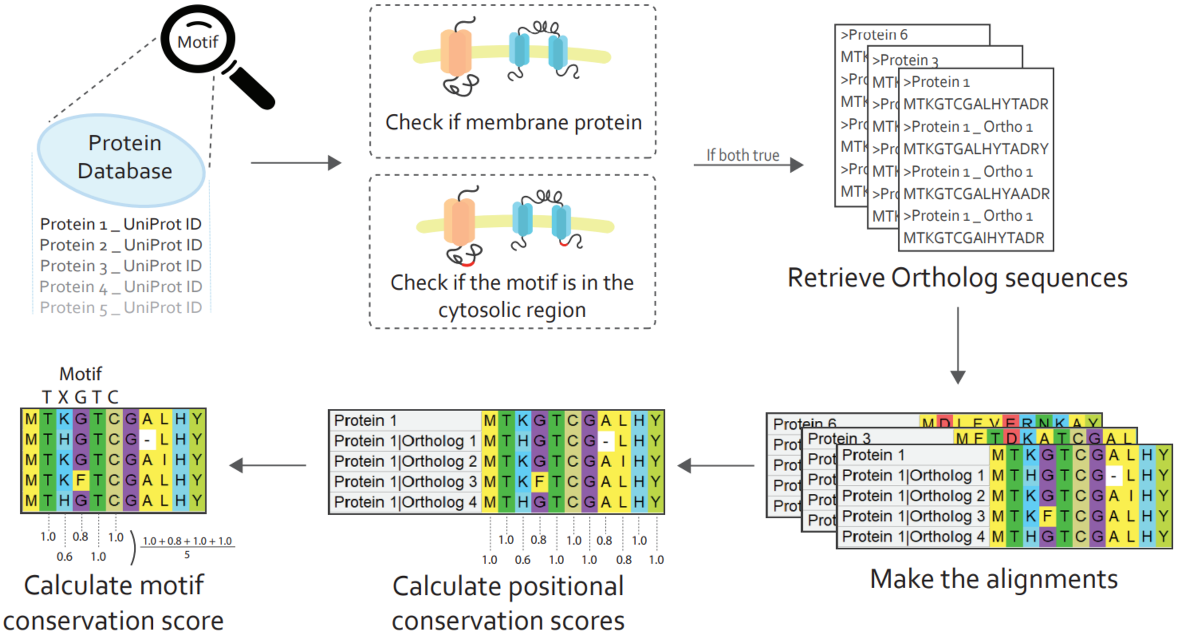
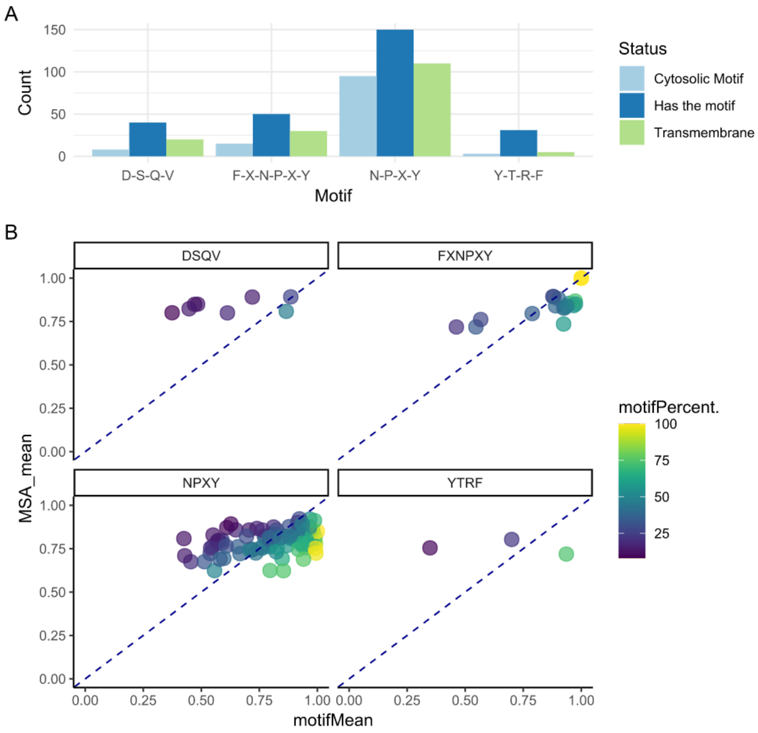
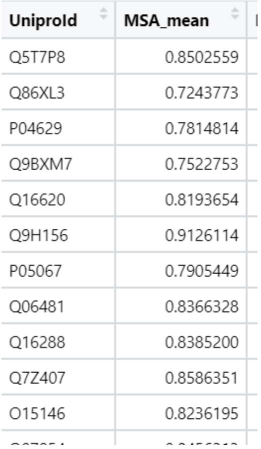
Identification of candidate transcytosis proteins through sequence analysis.

A set of motifs has been identified from the literature, which has been demonstrated to play a crucial role in the transcytosis process (as detailed in Table 1). To further investigate these motifs, the REST (Representational State Transfer) web service of the PrositeScan tool (de Castro et al., 2006) provided by the ExPASY portal was employed to search for proteome sequences that possess these motifs. The UniProt proteome database was searched for each motif, with isoforms filtered out and the search restricted to the Homo Sapiens organism. Metadata is generated using UniProt REST API. The metadata includes the location of the motifs on the sequence, topology information, and database reference IDs of each protein. Through the analysis of topology information, proteins without transmembrane domains are filtered out. The proteins that contain the motif in their cytoplasmic domains were selected from the remaining protein sequences. These proteins have been designated as preliminary candidates. The orthologous protein sequences of the preliminary candidate proteins are obtained from the OMA (Orthologous Matrix) (Altenhoff vd., 2021) database. Using the orthologous sequences, multiple sequence alignments (MSAs) were built for each pre-candidate protein using the Clustal Omega (Sievers vd., 2011) alignment method.



A custom Python script was used to calculate the positional conservation scores, which are determined by the ratio of the most common amino acid at that position to the total number of sequences in the alignment. The average conservation score of each MSA has been calculated by taking the mean conservation score of the positions for the alignment. While taking this average, the positions where the most observed character is a gap have been noted as significant contributors to the mean of the MSA, specifically for the sequences with more distant common ancestors. Hence, two mean MSA conservation scores have been generated, both with and without the positions primarily composed of gaps. The mean motif score has been calculated by averaging the conservation scores of the motif positions. For the motifs which have an X character, the X characters conservation score has been neglected for the calculation of the mean motif score. The percentile of the motif's conservation score is determined from the data by a custom Python script in order to assess the degree of preservation of the motif region in comparison to the rest of the MSA. The percentile of the motifall positions





UniproId	MSA_mean
Q5T7P8	0.8502559
Q86XL3	0.7243773
P04629	0.7814814
Q9BXM7	0.7522753
Q16620	0.8193654
Q9H156	0.9126114
P05067	0.7905449
Q06481	0.8366328
Q16288	0.8385200
Q7Z407	0.8586351
O15146	0.8236195

Figure X:

After the filtration of non-transmembrane proteins and motif location, the pre-candidate counts for each motif have been found as; 3 for YTRF, 109 for NPXY, 19 for FXNPXY, and 8 for DSQV. 82 of the NPXY protein showed a higher motif mean than the mean of their MSA. Out of these proteins, 10 of them have >98% conservation score.

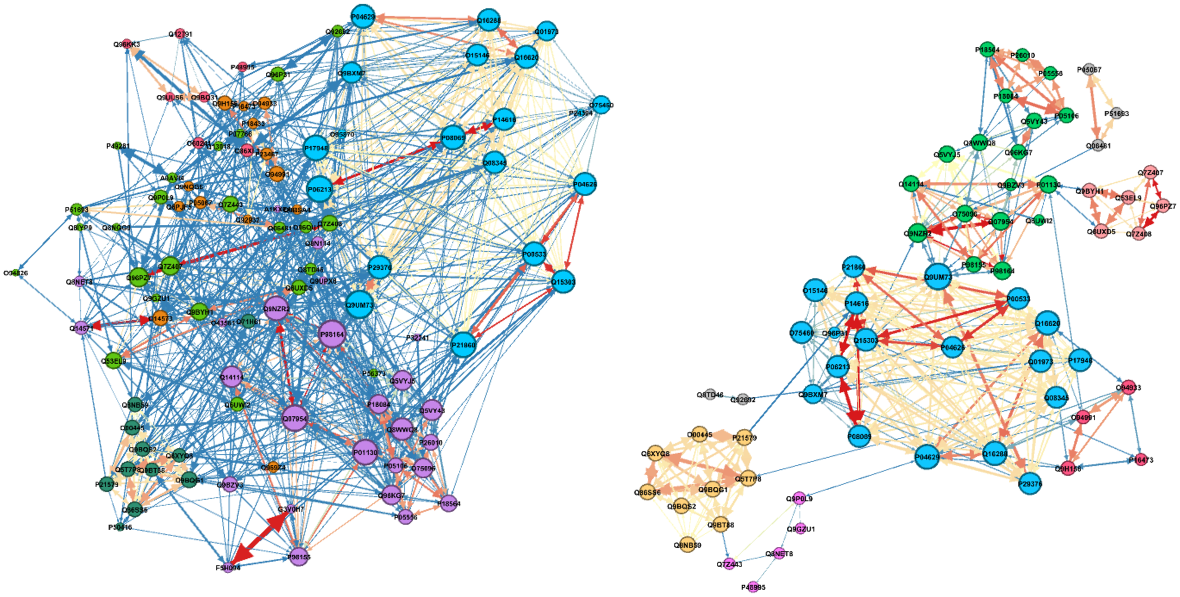
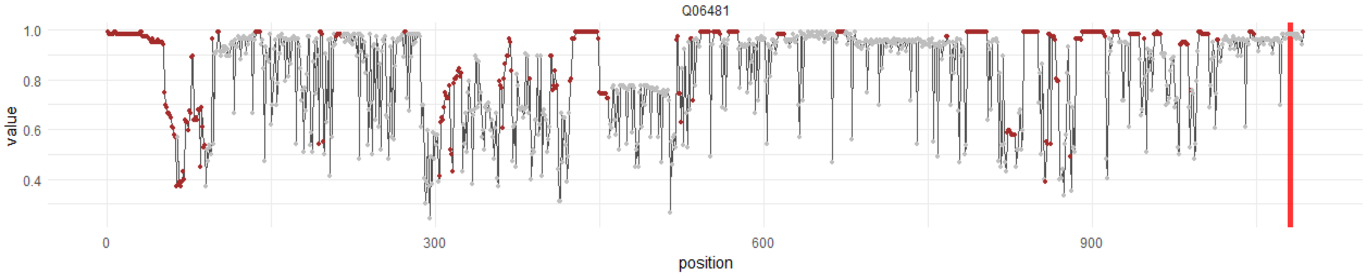
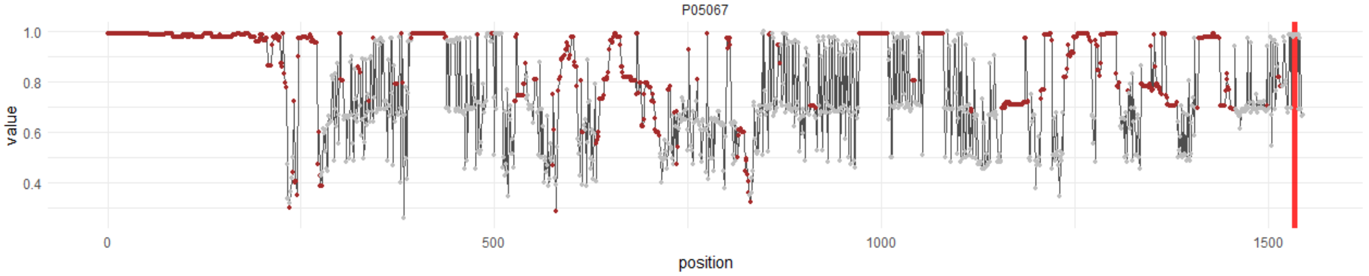
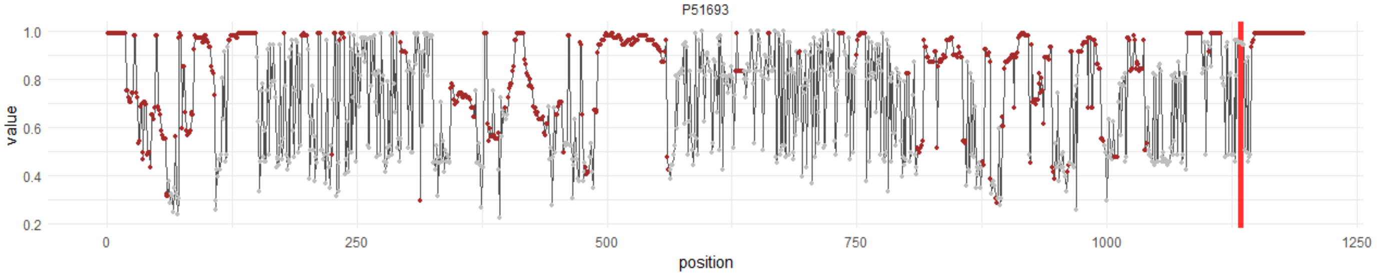
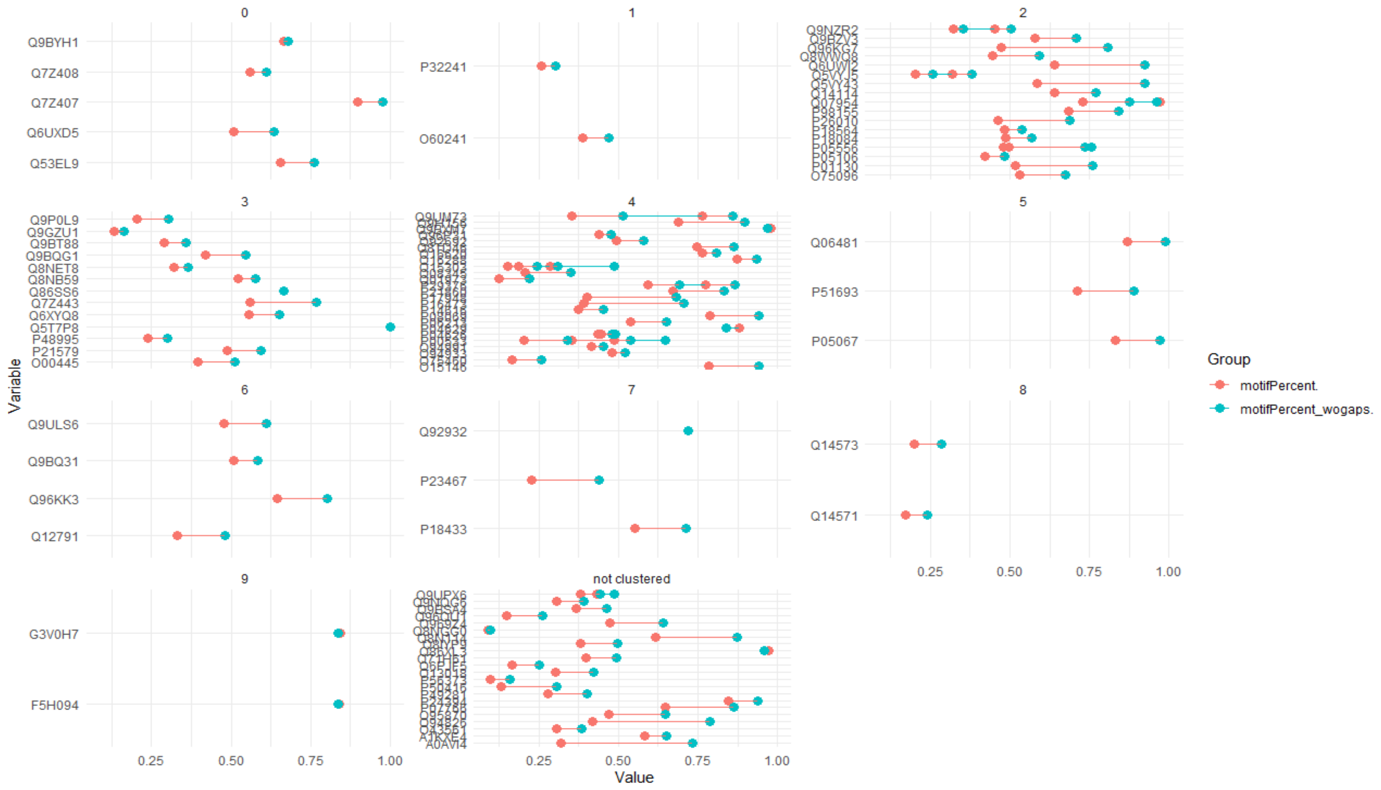

P04629
Q16288
Q01973
Q92692
Q96KK3
Q12791
P48995
Q96P31
Q9BXM7
Q15146
Q16620
Q9UL55
Q9BQ31
Q9H156
P16473
O94933
O75460
P24394
P18433
P14616
P07766
Q95870
P17948
P08069
P49281
O60241
Q13018
Q86XL3
P23467
O94991
Q08345
A0AVI4
Q9NQG6
P06213
P04626
Q9P0L9
Q6PJF5
P05067
Q7Z443
A1KXE4
Q8BSA4
P51693
Q92932
Q8IYP9
Q8NGG0
Q06481
Q96QU1
Q7Z408
Q8N114
P00533
O94826
Q7Z407
Q8TD46
P29376
Q15303
Q8NET8
Q96PZ7
Q9UPX8
Q6UXD5
Q9GZU1
Q9UM73
Q9BYH1
Q9NZR2
Q14573
Q14571
O43561
Q71H61
P98164
P32241
P21860
Q53EL9
P56373
Q14114
Q5VYJ5
Q8NB59
Q6UWI2
P18084
Q07954
Q5VY43
O00445
Q8WWQ8
Q9BQS2
P26010
Q6XYQ8
P01130
Q5T7P8
Q9BT88
P05106
O75096
P21579
Q9BQG1
Q969Z4
Q9BZV3
Q96KG7
P18564
Q86SS6
P05556
G3V0H7
P50416
P98155
F5H094
P18564
P26010
P05556
P05067
P18084
P51693
P05106
Q5VY43
Q8WWQ8
Q5VYJ5
Q96KG7
Q06481
Q7Z407
Q14114
Q9BZV3
P01130
Q9BYH1
Q53EL9
Q96PZ7
O75096
Q07954
Q6UWI2
Q9NZR2
Q6UXD5
Q7Z408
P98155
P98164
P21860
Q9UM73
Q15146
P14616
P00533
Q16620
O75460
Q96P31
Q15303
P04626
P06213
Q01973
P17948
Q8TD46
Q92692
Q9BXM7
Q08345
O94933
O94991
P08069
O00445
P21579
Q6XYQ8
Q5T7P8
P04629
Q16288
Q9H156
P16473
Q86SS6
Q9BQG1
P29376
Q9BQS2
Q9P0L9
Q9BT88
Q9GZU1
Q8NB59
Q8NET8
Q7Z443
P48995

Figure X :


Q06481
value
1.0
0.8
0.6
0.4
0
300
600
900
position


P05067
value
1.0
0.8
0.6
0.4
0
500
1000
1500
position


P51693
value
1.0
0.8
0.6
0.4
0.2
0
250
500
750
1000
1250
position


0
Q9BYH1
Q7Z408
Q7Z407
Q6UXD5
Q53EL9
1
P32241
O60241
2
Q9NZR2
Q9BZV3
Q96KG7
Q8WWQ8
Q6UWI2
Q5VYJ5
Q5VY43
Q14114
Q07954
P98155
P26010
P18564
P18084
P05556
P05106
P01130
O75096
3
Q9P0L9
Q9GZU1
Q9BT88
Q9BQG1
Q8NET8
Q8NB59
Q86SS6
Q7Z443
Q6XYQ8
Q5T7P8
P48995
P21579
O00445
4
Q9UM73
O15146
5
Q06481
P51693
P05067
6
Q9ULS6
Q9BQ31
Q96KK3
Q12791
7
Q92932
P23467
P18433
8
Q14573
Q14571
0.25
0.50
0.75
1.00
9
G3V0H7
F5H094
0.25
0.50
0.75
1.00
not clustered
Q9UPX6
Q9NQG6
Q9BSA4
Q13018
P50416
P49281
O94826
O43561
A1KXE4
A0AVI4
0.25
0.50
0.75
1.00
Value
Variable
Group
● motifPercent.
● motifPercent_wogaps.