Homolog prediction using BLAST EMBOSS:

The Basic Local Alignment Search Tool: BLAST EMBOSS (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) [4] was used to perform homolog prediction and search for sequences of resemblance between that of the hypothetical protein by submission of the UniProt amino acid sequence. Not only was BLAST applicable for the findings of similar alignments, but it also could potentially provide the correct mRNA encoding sequences between such homologs. BLAST generated sufficient comparable measurements and presented the potential identities, positives, and gaps between the sequence alignments.

Further homolog conformation using Needle:

Needle (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) [5] was used as an additional homolog alignment tool. The sequence for the homolog of choice, *Spermatogenesis*, was first found by the usage of BLAST and confirmed by Uniprot. Both sequences for LOC685762 and the homologous protein were inserted into the tool and more specific homology findings were accessed and compared with results generated by BLAST.

Annotated Sequence features along with transmembrane classification using Protter:

Protter (https://wlab.ethz.ch/) [6] was used to perform a detailed analysis of the individual amino acids involved within LOC685762 as well as classify the protein. The UniProt protein accession or the amino acid sequence was inserted into the tool. Visualization of such outputs are specific to that of the protein that is submitted. Color coordination with individual details within pictogram could potentially be done after the sequence is analyzed and generated by the tool.

Further transmembrane classification of protein using SMART:

By submitting the UniProt accession name of the hypothetical protein (F1LVR6) SMART (http://smart.embl-heidelberg.de/) [7] generated geometric imagery that was able to further classify the hypothetical protein. The protein sequence could also potentially be used as an identifier when submitting a protein in SMART.

Protein secondary structure with Phyre^2:

To verify the complex secondary structure of LOC685762 and its relative homolog the tool Phyre^2 (http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index)[8] was applied. Sequence search was imported and a visual representation of what the potential 3D protein appears like in nature was generated. Not only was the portal able to provide an interactive model, but it was able to deliver the distinct sequences in which domains are visualized within the protein.

Mutational analysis with I-mutant:

To understand potential differences in the specific amino acid sequences I-mutant (https://folding.biofold.org/i-mutant/i-mutant2.0.html) [9] was used to generate potential outcomes of such scenarios. Specifically, a single amino acid mutation and the amino acid sequences were inserted into imutant and the results were generated.

Results and Discussions:

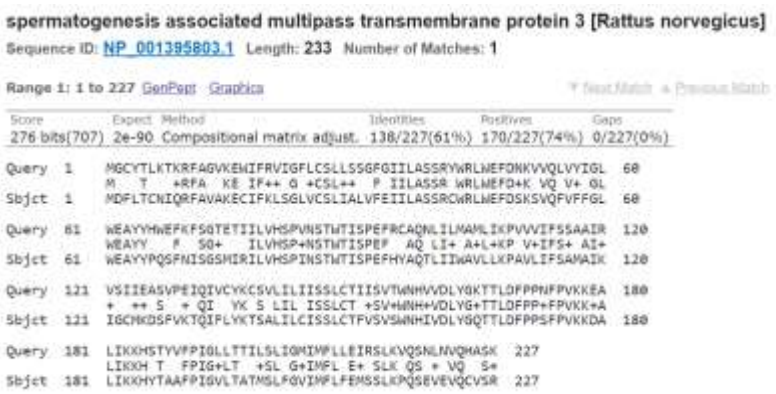
Experimental design/rational:

A protein's function is directly related to the protein's relative amino acid sequence, as well as the secondary structures involved within the protein, such as prominent domains, alpha helices, and beta sheets/turns. With reasoning, it can be assessed that similar proteins should exhibit similar amino acids sequences, secondary structures, and thus, similar functionality within biological systems, particularly within the cell, since it was determined that the hypothetical protein will function within the cell membrane. Mutations within an amino acid sequence can greatly differ the potential function of a protein. Mutational analysis was done computationally to envision how the stability of the protein could be differentiated from a single amino acid variation (SAV). LOC685762's potential role, function, and structure, was discovered by an *in silico* methodology to greater the knowledge available about the hypothetical protein of interest.

Homolog determination with BLAST and Needle:

BLAST was used to find potential homologs for LOC685762. By submitting the UniProt sequence of LOC685762, BLAST released a list of possible alignments that could be further explored. Hypothetical proteins have not been proven to exist in nature and in biological systems by lack of experimental procedures, so comparing them to authentic and experimentally proven proteins assists with the understanding of the potential amino acid sequences, domains, and secondary structures involved in the hypothetical protein. Spermatogenesis associated multiphases transmembrane protein 3, also found in the organism *Rattus norvegicus*, can be assigned as a homologous protein to that of the hypothetical protein studied in this research. As shown in Figure 2, 138/227 identities were classified when comparing these two homologs (*Spermatogenesis* being the subject and LOC685762 being the query). The results generated stated that the two individual protein's amino acid sequences were ~61% identical. There were 170/227 positives, stating that 74% of the amino acids were closely related within the sequences. There were 0 gaps generated, which means these amino acid sequences are proposed to be the same length. [4]

As shown in Figure 2 below, there are conservative substitutions within the amino acid sequences between the homolog and the hypothetical protein. A "+" shown represents similarity between two amino acids. For example, at position 227, there is a substitution between Lysine and Arginine. [4] Both of these amino acids have positively charged R groups. Since the R groups of both residues represent the same charge, BLAST considered them as "similar." Similar amino acid sequences can relate to similar functions between the two proteins. Not only are similar sequences relative to the function of the protein, but the secondary structures that the amino acid chains fold into may be similar as well. Further analysis using additional tools were used to potentially answer the question as to how similar the function and structure of LOC685762 is to that of *Spermatogenesis.*



spermatogenesis associated multipass transmembrane protein 3 [Rattus norvegicus]

Sequence ID: NP_001395803.1 Length: 233 Number of Matches: 1

Range 1: 1 to 227 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
276 bits(707)	2e-90	Compositional matrix adjust.	138/227(61%)	170/227(74%)	0/227(0%)

```
Query  1    MGCYTLKTKRFAGVKEWIFRVIGFLCSLLSSGFGIILASSRYWRLWEFDNKVVQLVYIGL  60
            M   T   +RFA  K E IF++ G +CSL++  F IILASSR WRLWEFD+K VQ V+ GL
Sbjct  1    MDFLTCNIQRFAVAKECIFKLSGLVCSLIALVFEIILASSRCWRLWEFDSKSVQFVFFGL  60

Query  61   WEAYYHWEFKFSGTETIILVHSPVNSTWTISPEFRCAQNLILMAMLIKPVVVIFSSAAIR  120
            WEAYY  F  SG+   ILVHSP+NSTWTISPEF  AQ LI+ A+L+KP V+IFS+ AI+
Sbjct  61   WEAYYPQSFNISGSMIRILVHSPINSTWTISPEFHYAQTLIIWAVLLKPAVLIFSAMAIK  120

Query  121  VSIIEASVPEIQIVCYKCSVLILIISSLCTIISVTWNHVVDLYGKTTLDFPPNFPVKKEA  180
            +  ++ S  + QI  YK S LIL ISSLCT +SV+WNH+VDLYG+TTLDFPP+FPVKK+A
Sbjct  121  IGCMKDSFVKTQIFLYKTSALILCISSLCTFVSVSWNHIVDLYGQTTLDFPPSFPVKKDA  180

Query  181  LIKKHSTYVFPIGLLTTILSLIGMIMFLLEIRSLKVQSNLNVQHASK  227
            LIKKH T  FPIG+LT  +SL G+IMFL E+ SLK QS + VQ  S+
Sbjct  181  LIKKHYTAAFPIGVLTATMSLFGVIMFLFEMSSLKPQSEVEVQCVSR  227
```

Figure 2: BLAST Homolog Result
Offers the homologous comparison provided by BLAST
between hypothetical protein LOC685762, and authentic protein, *Spermatogenesis.*

Other homologous tools can be used to compare similar proteins. Needle gave similar results as Protein BLAST but a few differences were noticed. Because the protein in question is hypothetical, the bioinformatic tools seem to alter the results by a small marginal factor, as seen in Figure 3. Needle states that there is a gap score of 2.9%. LOC685762 being about 7 amino acids longer than *Spermatogenesis.* A statistically significant difference is within 5%, so this can be neglected. The real understanding here is that there are in fact authentically biological proteins that have been studied that are similar to LOC685762. [5]

```
                        # Score: 728.0
EMBOSS_001         1 MGCYTLKTKRFAGVKEWIFRVIGFLCSLLSSGFGIILASSKYWRLWEFDN     50
                     |...|...:|||..||.||::.|.:|||::..|.||||||:.|||||||:
EMBOSS_001         1 MDFLTCNIQRFAVAKECIFKLSGLVCSLIALVFEIILASSRCWRLWEFDS     50

EMBOSS_001        51 TVVQLVYIGLWEAYYHWEFKFSGTETIILVHSPVNSTWTISPEFRCAQNL    100
                     ..||.|:.||||||||...|..||:...||||||:||||||||||..||.|
EMBOSS_001        51 KSVQFVFFGLWEAYYPQSFNISGSMIRILVHSPINSTWTISPEFHYAQTL    100

EMBOSS_001       101 ILMAMLIKPVVVIFSSAAIRVSIIKASVSEIQIVCYKCSVLILIISSLCT    150
                     |:.|:|:||.|:||||:.||::..:|.|..:.||..||.|.|||.||||||
EMBOSS_001       101 IIWAVLLKPAVLIFSAMAIKIGCMKDSFVKTQIFLYKTSALILCISSLCT    150

EMBOSS_001       151 IISVTWNHVVDLYGKTTLDFPPNFPVKKEALIKKHSTYVFPIGLLTTTLS    200
                     .:||:|||:||||||:||||||:|||||:|||||||.|..||||:||.|:|
EMBOSS_001       151 FVSVSWNHIVDLYGQTTLDFPPSFPVKKDALIKKHYTAAFPIGVLTATMS    200

EMBOSS_001       201 LFGVIMFLFEIKSLKVQSNLNVQHASKQADDMYINESAVA    240
                     ||||||||||:.|||.||.:.||..|:.     ||:.|
EMBOSS_001       201 LFGVIMFLFEMSSLKPQSEVEVQCVSRP-----INQKA--    233

# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 240
# Identity:     144/240 (60.0%)
# Similarity:   176/240 (73.3%)
# Gaps:           7/240 ( 2.9%)
```

Figure 3: Needle Homolog Result
The additional homologous comparison provided by Needle
between hypothetical protein LOC685762, and authentic protein, *Spermatogenesis.*

Secondary Structure analysis with Phyre[2]:

After analyzing the primary structure of the two proteins, it came to interest what specific secondary structures were involved in both homologous proteins. The tool Phyre[2] was used to generate detailed 3D figures of both the homologous protein and the hypothetical protein as well as their amino acid alignment corresponding to such secondary structures. Figure 4A shows the homologous protein's (*Spermatogenesis*) secondary structure. Different colors help differentiate between the unique secondary structure involved.[8]

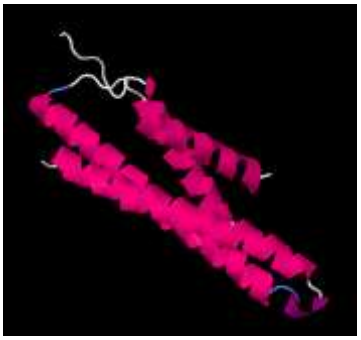


Figure 4A: *Spermatogenesis* 3D Structure
Gives the complex secondary structure of *Spermatogenesis*, provided by Phyre[2]. Pink alpha helical structures and purple beta sheets are present.

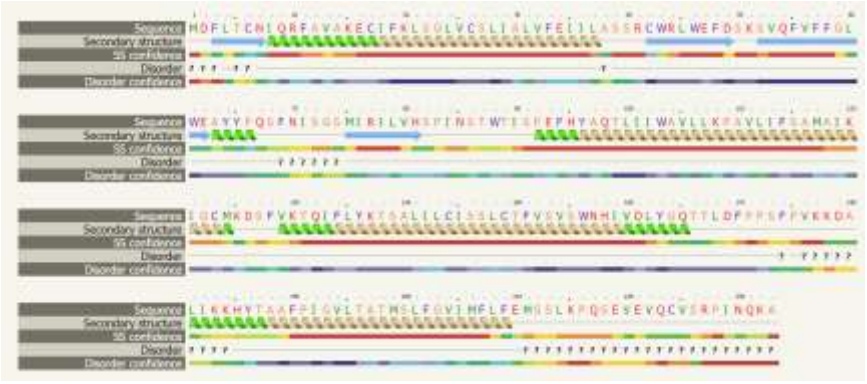


Figure 4B: *Spermatogenesis* alignment
Amino acid alignment of *Spermatogenesis*, with the addition of secondary structures provided by Phyre[2].

As shown in figure 4A *Spermatogenesis* contains 4 alpha helical structures, which are colored pink for differentiation between beta sheets (purple), the other secondary structure that is involved within the protein. As shown above, in figure 4B, *Spermatogenesis* amino acid sequence shows many hydrophobic amino acids where there is an alpha helix present in the structure. For example, at position 96 there is a Tyrosine present. Tyrosine's structure is nonpolar because its R-group is a hydroxyl substituted aromatic ring, which contains a lot of hydrogens and carbons. Alpha helixes are a common secondary structure, and they could

potentially be observed on x-ray's experimentality. Without actually conducting an experiment, Phyre[2] was able to provide a sufficient *in silico* structure of the homologous protein.

As shown in figure 5A, LOC685762 also contains 4 alpha helices. However, it does show more beta sheets and beta turns than the authentic rat protein, as it shows 4 sheets which are colored yellow.[8]

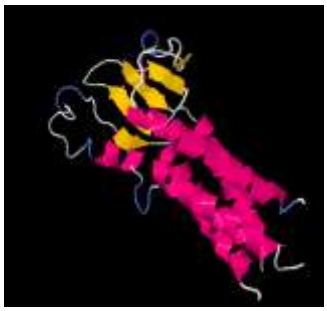


Figure 5A: LOC685762 3D Structure
Provides the complex secondary structure of LOC685762 provided by Phyre[2]. Pink alpha helical structures, blue beta turns, and yellow beta sheets are present.

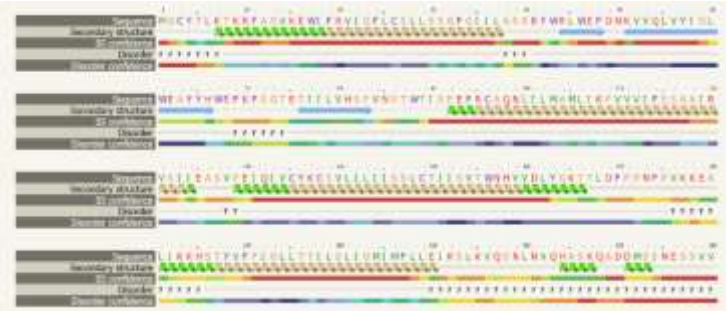


Figure 5B: LOC685762 alignment
Amino acid alignment of LOC685762, with the addition of secondary structures provided by Phyre[2].

Beta sheets are also stabilized by hydrogen bonding in polypeptides. The shape is folded in a pleated way, with more of a zigzag formation rather than a spiral like the alpha helix. The R-groups involved in the zigzag bulge out in opposite directions. B-sheets can be parallel or antiparallel. Shown in figure 5A, the hypothetical protein contains antiparallel sheets because they have opposite amino-to-carboxyl positionings. B-turns are a common structure contained within a protein in which they connect the antiparallel B-sheets together in a continuous fashion. The 180̊ turn includes four crucial amino acid residues. The carbonyl carbon of the first amino acid residue participates in the intermolecular force of hydrogen bonding with the amino-group hydrogen on the fourth residue. The two innermost residues do not contribute to any hydrogen bonding.

When comparing the 2 structures, observationally they are similar but visualizing and indicating what part of the amino acid sequence is involved in each structure can be significant to

determining the role of that amino acid in the protein. For example, at positions 63- 66, LOC685762 shows an “AYYH sequence” as provided in figure 5B. At this same location, *Spermatogenesis* shows an AYYP sequence. As shown below in Figures 4A and 5A, *Spermatogenesis* shows a helical structure at this location whereas LOC685762 shows a beta sheet configuration. In proteins, everything relates back to the sequence of amino acids, otherwise known as the primary structure. A different sequence might mean there is a different function. In this case, the result is unique. Proline (P) has a challenging time creating a helical structure because it has no substituent hydrogen present to help with the weak hydrogen bond interaction that is needed in order to stabilize the structure and bond with neighboring amino acid residues. Proline also has a nitrogen atom that is included in a rough cyclic ring, as this does not allow for the rotation that is needed to form an alpha helix. So, with this knowledge, it is questionable to as why Proline is added as a residue in a helical configuration in *Spermatogenesis.*

Both LOC685762 and *Spermatogenesis* consist of strong hydrogen bonding within the protein because they involve classes of secondary structures such as beta sheets, turns, and alpha helices. Hydrogen bonding is common to many proteins, but the most stable hydrogen bonds are within these complex secondary structures because the hydrogen bonds are maximized and create a rigidity within the protein that strengthens the relative stability of the protein as a whole.

Classification of transmembrane protein: Protter and SMART

There are many types of proteins that have different functions within the cell. By using bioinformatics tools like Protter and SMART and inputting the UniProt amino acid sequence one can get a visual representation of what kind of protein a hypothetical protein would exist as. Classifying hypothetical proteins is not always definitive. Since these proteins are not to our scientific knowledge, real, these tools may have a harder time generating sufficient results. With that being stated Protter and Smart generated similar results where the probability of a mistake is minimal.

Figure 6A shows that the hypothetical protein is a transmembrane protein. As stated previously, transmembrane proteins extend throughout the entirety of the membrane. Their function is to help transport specific materials and biomolecules throughout the cell. They have high specificity, allowing passage of only the chemical substances that are needed in the cell at the relative moment. [7]As shown below in figure 6B, four transmembrane regions are present in the protein. This also shows where in the sequence the transmembrane regions begin and end. The residues that are involved in the region that cross the membrane first are from 21- 43. As shown in figure 5B, there is a helical structure present (shown in brown) along with many hydrophobic amino acids such as isoleucine (I) at position 22.[8] With this knowledge it is now understood that LOC685762 is a polytopic transmembrane protein as it crosses the membrane several times. Polytopic proteins have a high degree of hydrophobic amino acids, especially sequences that cross the membrane as alpha helical structure as shown above.

Figure 6A: SMART pictograph
Depicts that LOC685762 is a transmembrane polytopic protein.
The Blue rectangles represent the helical regions.

Name	Start ▲	End	E-value
transmembrane region	21	43	N/A
transmembrane region	100	122	N/A
transmembrane region	134	156	N/A
transmembrane region	187	209	N/A

Figure 6B: SMART data set
Provides confidently predicted domains, repeats, motifs, and features involved in LOC685762

Figure 7 shows the Protter generational pictogram of the hypothetical transmembrane protein. It shows the individual amino acids that are within the membrane. These amino acids will all have unique and measurable pKas. These pKas will be altered because they are within the cell and imbedded in the semi permeable plasma membrane. By seeing these results, the protein can be referred to as a transmembrane 1-4 protein. [6]

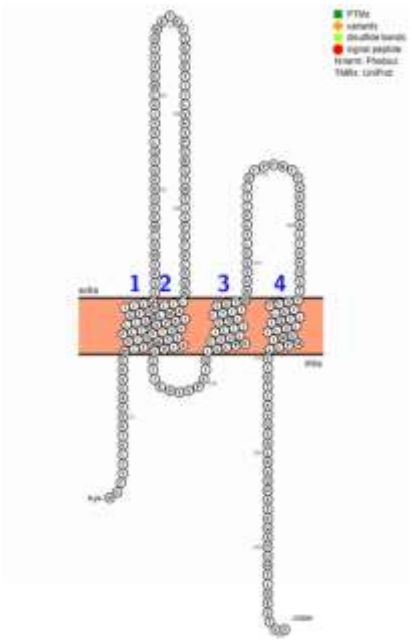


Figure 7: Protter,
LOC685762 structure
Transmembrane classification using protein with amino acid residues present.

Transmembrane classification builds on the previously known information from Phyre[2] that LOC685762 contains stabilized hydrogen bonding in secondary structures, such alpha helices. Not all transmembrane proteins exhibit a helical structure intertwined within the membrane. Some transmembrane proteins exist as beta-barrels instead. In some cases, transmembrane proteins can be channel proteins. Channel proteins allow smaller ions to pass through the semi-permeable membrane of the cell. Protter's generated structure could possibly be used to narrow down the identification of LOC685762 but not fully classify the type of transmembrane protein it could theoretically be. Channel proteins contain a hydrophilic core. [10] As shown in figure 7, the core of LOC685762 contains many hydrophobic amino acid residues. *Unofficially* it can be stated that the protein is not a channel protein due to the major hydrophobicity content of LOC685762. Protter's structure is detailed in a bioinformatics sense, but because this protein is hypothetical, computational analysis could be difficult to determine thus distinct class of transmembrane protein in which LOC685762 belongs too is still unknown.

Mutational Analysis with imutant:

A single amino acid substitution can alter the relative stability of a protein. Chemically, it is known that each amino acid serves their individual purpose in stabilizing and structurally supporting a protein. The primary sequence of the protein determines the structure and function. When classifying amino acids into to their respective categories, Lysine has a positively charged R group and Threonine has a polar, uncharged R group. Since these amino acids have R groups of different natures, determining the strength of the hypothetical protein (wildtype) compared to a mutant was performed.

I-mutant was able to generate a potential scenario where such amino acid variation occurs. When the temperature and pH stay constant and just the mutation from K to T occurs, the stability of the protein decreases. This is because the Delta Delta G is less than zero as shown below in Table 1. This represents the change in energy between the folded and the unfolded states and the change of free energy when this mutation is brought upon by is a single amino acid variation. A protein is more stable when it is in the folded conformation rather than the unfolded. [9] When the protein was replaced with a Threonine at position 51, the protein started to unfold. As shown in figure 5B, position 51 shows that a beta sheet is present. [8] This hypothetical mutational example goes to show that point mutations with amino acids can be for the better or for the worse but analyzing the data can help with the understanding of what specific amino acids are better at their job. For example, beta sheets form best when amino acids contain aromatic groups, or amino acids that are branched. Since Threonine is a branched amino acid, it seems that it could both potentially form a beta sheet. In this particular case for LOC685762, the beta sheet is more stable when a Leucine is present rather than a Threonine at position 51.

Position	51
WT	K
NEW	T
DDG	-1.22
pH	7.0
T (˚K)	25

Table 1: Mutational Data Set

Position, wildtype (WT), new amino acid, delta delta G, pH and temperature were generated by imutant.

Concluding Statements:

It was determined that the hypothetical protein potentially exists as a transmembrane protein. Further experimental or *in silico* data would be needed to conclude the specific kind of multi-pass transmembrane protein LOC685762 potentially exists as. Establishing the knowledge that LOC685762 may be a tight junction helped determine its potential role within the cell, specifically the cell membrane, as it would help with the selective permeability, and provide adhesion between two adjacent cells. With the usage of bioinformatic tools, complex 3D structures, homologous alignment, and mutational analysis, were generated, and thus further knowledge about LOC685762 has been found. LOC685762 maximizes the hydrogen bonding by forming secondary structures such as beta sheets and alpha helices. *Spermatogenesis* was found to be~61% identical and~74% similar to LOC685762. Since the primary structure, or the amino acid sequence, relatively acts as a precursor for determining the function of the protein, it can be stated that LOC685762 and its relative homolog, *Spermatogenesis,* could potentially exhibit similar roles within the cell membrane as transmembrane, tight junction proteins. Mutational analysis can help determine how the stability of a protein can be altered- even with just one single amino acid variation. Although this protein is hypothesized to exist, computational analysis was able to further classify the protein and provide a better understanding of its role within the mammal, *Rattus norvegicus.*

References:

(1) Baeza-Delgado, C.; Marti-Renom, M. A.; Mingarro, I. Structure-Based Statistical Analysis of Transmembrane Helices. *Eur Biophys J* **2013**, *42* (2–3), 199–207. https://doi.org/10.1007/s00249-012-0813-9.

(2) Liu, B.; Wang, X.; Chen, Q.; Dong, Q.; Lan, X. Using Amino Acid Physicochemical Distance Transformation for Fast Protein Remote Homology Detection. PLoS ONE 2012, 7 (9), e46633. https://doi.org/10.1371/journal.pone.0046633.

(3) Otani, T.; Furuse, M. Tight Junction Structure and Function Revisited. *Trends in Cell Biology* **2020**, *30* (10), 805–817. https://doi.org/10.1016/j.tcb.2020.08.004.

(4) Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421. doi: 10.1186/1471-2105-10-421. PMID: 20003500; PMCID: PMC2803857.

(5) Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research. 2022 Apr:gkac240. DOI: 10.1093/nar/gkac240. PMID: 35412617; PMCID: PMC9252731

(6) Omasits, U.; Ahrens, C. H.; Müller, S.; Wollscheid, B. Protter: Interactive Protein Feature Visualization and Integration with Experimental Proteomic Data. *Bioinformatics* **2014**, *30* (6), 884–886. https://doi.org/10.1093/bioinformatics/btt607.

(7) Letunic, I.; Khedkar, S.; Bork, P. SMART: Recent Updates, New Developments and Status in 2020. *Nucleic Acids Research* **2021**, *49* (D1), D458–D460. https://doi.org/10.1093/nar/gkaa937.

(8) Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. E. The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis. *Nat Protoc* **2015**, *10* (6), 845–858. https://doi.org/10.1038/nprot.2015.053.

(9) Bava, K. A. ProTherm, Version 4.0: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Research* **2004**, *32* (90001), 120D – 121. https://doi.org/10.1093/nar/gkh082.

(10) Montal, M. Molecular Anatomy and Molecular Design of Channel Proteins. *FASEB j.* **1990**, *4* (9), 2623–2635. https://doi.org/10.1096/fasebj.4.9.1693348.