

Annotation and Homology Modeling of the Multidrug Transport Protein P-Glycoprotein (ABCB1) of Equus Caballus

Barachel Butler¹ and Karobi Moitra[#]

¹Trinity Washington University, Washington, DC, USA

[#]Advisor

ABSTRACT

The function of ABC transporter proteins, such as ABCB1, is to transport substrates across cell membranes in many organs. ABCB1 can be found in multiple species, however, the sequence we annotated is derived from *Equus caballus* (horse). The objective of this research was to annotate and construct a structural homology model of the horse-derived P-glycoprotein. We classified the protein as a transmembrane ATP-binding cassette (ABC) protein ABCB1. Annotation of the sequence (which is not yet manually curated in NCBI at the time of writing) was carried out using various tools including BLAST, TMHMM, PFAM, HMM Logo etc. and supports its designation as ABCB1 or P-glycoprotein of horse. The homology model was constructed using SWISS-MODEL based on the structure of human P-glycoprotein (PDB code: 6c0v). Three structures that share high sequence-identity were chosen for analysis. Two homology models were prepared using the cryo-EM structures of human ABCB1 (PDB code: 6qex) and human-mouse chimeric P-glycoprotein (PDB code 6qee). The model using the structure of human P-glycoprotein (PDB code: 6c0v) as a template had the highest QMEAN score, Ramachandran favorability, and the most favorable MolProbity score. That model was evaluated using ProCheck and energy minimized using Chiron to give rise to the final model. Energy minimization resolved an unmodeled loop from Q625 to V691 and corrected other minor distortions. The clash ratio of the energy minimized model indicated that there are few clashes in the structure. Based on analysis of the structure validation parameters, the best homology model of equine P-glycoprotein was derived using the human P-glycoprotein (PDB: 6c0v) as a template.

Introduction

The function of ABC transporters is to transport substrates across cell membranes using ATP-dependent processes (Kopcho, et al., 2019). ABCB1 transports substrates (such as drugs) in many organs and may impair drug absorption in the digestive tract (Finch and Pillans, 2014). The ABC protein subfamily B member 1 (ABCB1) is a P-glycoprotein that is involved in toxin-protection in the liver and blood-brain barrier (Vore, 2019). ABCB1 proteins are of particular interest to pharmacologists and biochemists since they are overexpressed in drug resistant and tumor cells, hence the protein also is also known as the multi-drug resistant (MDR1) protein (Vore, 2019). This protein can be found in multiple species including donkey (*Equus asinus*) and walrus (*O. rosmarus*) (Camacho, et al., 2008). However, the sequence we annotated is derived from *Equus caballus* (Camacho, et al., 2008, O'Leary, et al., 2016). ABCB1 may have a role in drug transport in horses, especially for those with pathways through the digestive tracts, such as analgesics like methadone (Linardi and Natalini, 2006) (Linardi, et al., 2012). ABCB1 has also been linked to the expulsion of chemotherapeutic drugs from cells in other species, such as humans (Goodsell, 2010, Alam, et al., 2019). Homology modeling is a process by which similar sequences to the target query sequence are identified and used as a template for the development of a novel model (Bordoli, et al., 2008). We constructed a novel model of horse P-glycoprotein using homology modeling. *E. caballus* horse-derived P-glycoprotein (Reference Sequence: XP_014594657.1) is a

predicted multi-drug resistant ABCB1 P-glycoprotein, (O'Leary, et al., 2016). The purpose of this study is to use various bioinformatics programs to collect important structural information on the horse-derived P-glycoprotein and to apply that information to construct an appropriate model through homology modeling.

Materials and Methods

Annotation

First, the sequence was annotated using several web-based servers and tools. A basic local alignment search tool (BLAST) search was conducted to identify and analyze similar structures across species (Camacho, et al., 2008). Then, the number of transmembrane helices on the protein and the cellular location of the protein was predicted by several programs, including TMHMM (Sonhammer, et al., 2001); WolfPSORT (Horton, et al., 2007); Phobius (Kall, et al., 2007); TMPred (Hoffman and Stoffel, 1993); and HMMTop (Tsunady and Simon, 2001). The structure of the protein, including its domains, was analyzed through the Pfam (El-Gebali, et al., 2019), which was then visualized using Skyline (Wheeler, et al., 2014). Finally, the protein was given a name based on the information from this analysis.

Homology Modeling

The methods for homology modeling were derived from a protocol developed by Berdoli, et al. (2008). The three templates with the highest sequence identities were selected for analysis. Homology models were created for these templates using SwissModel (Waterhouse, et al., 2018) and compared to each other using their Molprobit scores, QMEAN scores, and Ramachandran plots (Benkert, et al., 2009; Bienert, et al., 2017). The top two models were compared for the selection of one final model that was used to construct the equine P-glycoprotein model. The model was evaluated for proper stereochemistry and geometry through the PDBSUM (ProCheck) (Lakowski, et al., 1993) and SAVES (SAVES v.5 is another model-validation software that includes WhatCheck, verify 3D, Errat, PROVE and PROCHECK programs) (<https://servicesn.mbi.ucla.edu/SAVES/>). The final structure was energy-minimized by Chiron (Ramachandran, et al., 2011) and visualized on the NCBI Structure Viewer (iCN3D) (Wang, et al., 2020).

Results

Annotation

Basic Information

E. caballus (horse)-derived p-Glycoprotein, which is labelled as a multi-drug resistant ABCB1 P-glycoprotein by the NCBI (Reference Sequence: XP_014594657.1) has a sequence, which is 1275 amino acids in length, the sequence is attached in the supplementary material and was derived from NCBI (O'Leary, et al., 2016).

Cellular Localization

Five programs were used to determine the structure of the sequence (TMHMM, WolfPSORT, Phobius, TMPred, SACS HMMTOP). TMHMM predicted the presence of 10 transmembrane alpha helices, each with 244 amino acids (see Figure 1A). However, this data is inconsistent with both the known structure of ABC proteins, as well as the number of amino acids in the sequence (1275 amino acids). Therefore, other programs had to be used to obtain a more

accurate prediction. WolfPSORT was able to predict that the localization of the protein is consistent across species. Based on the WolfPSORT search, one can conclude that the protein is an integral membrane protein, with the highest-identity results coming specifically from MDR1. The phobius program suggested that parts of the P-glycoprotein are in the cytoplasm, while other parts are not. This program, in contrast to the WolfPSORT and TMHMM, predicts the presence of 12 alpha helices (see Figure 1B). Given the fact that the WolfPSORT predicts that the protein is an integral membrane protein, it is reasonable to hypothesize that the P-glycoprotein has parts that are localized in the membrane as well. Because this data differs from the two previous sets, more localization analyses were required. The TMPred results showed that from inside helices to outside helices, 12 helices were found, but when counted from outside to inside, 14 helices were counted. The strongly-preferred model (which, according to the TMPred description of the data, is highly speculative) showed 11 strong transmembrane helices (see Figure 1.C). The final program of analysis, SACS HMMTOP, aligns the amino acid sequence such that 12 transmembrane helices exist on the protein, as shown below in the graphic produced (see Figure 1D). This analysis was supported by the structural evidence explored further in the annotation.

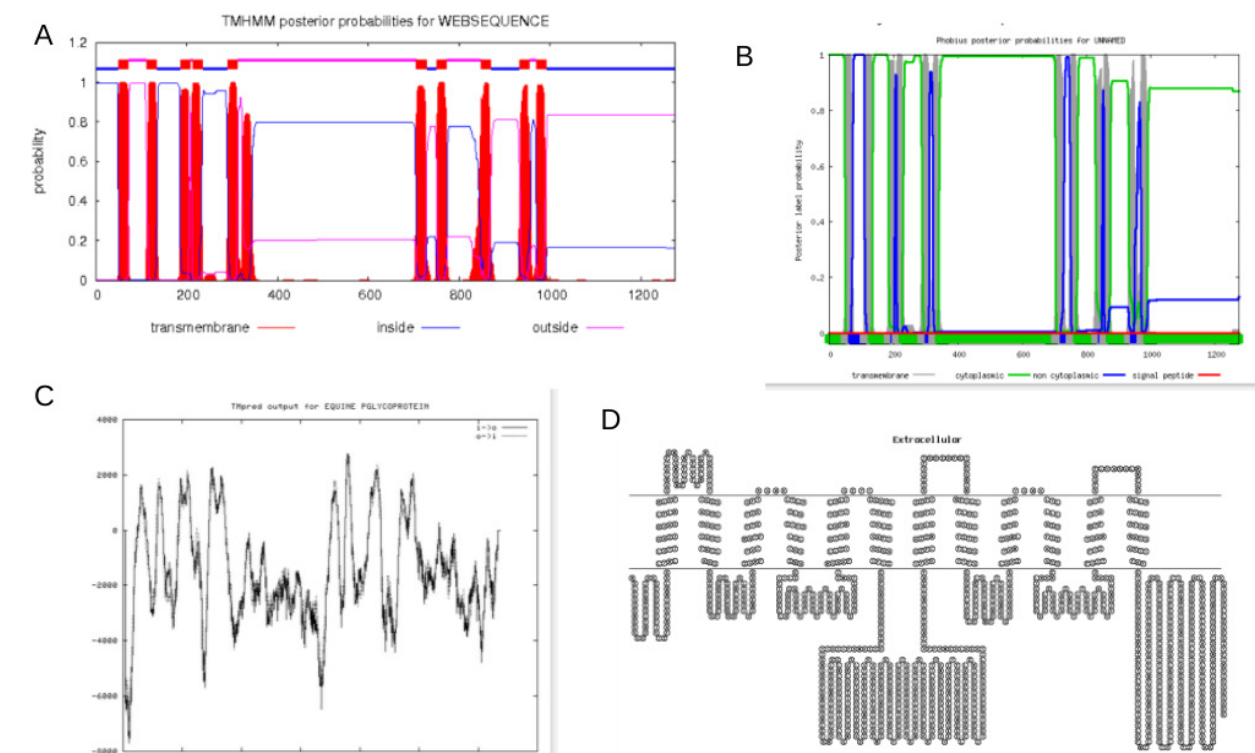


Figure 1. Cellular Localization Predictions. The TMHMM (A) predicted that the p-glycoprotein had 10 transmembrane helices, each with 244 amino acids. The Phobius (B) results predicted 12 alpha helices. TMPred (C) predicted that 11 strong transmembrane helices existed. The SACS HMMTOP (D) alignment showed 12 transmembrane helices, and was most supported by structure-based evidence.

Structure-Based Evidence

The amino acid sequence was entered into the Pfam database for domain analysis. According to Pfam, there are four significant domains in this protein structure, including two ABC Transporter Transmembrane domains, each of which contain six transmembrane alpha helices, and two ABC Transporter domains (El-Gebali, et al., 2019) or nucleotide

binding domain (NBD's). This structure is consistent with that of other ABC proteins, and supports the cellular localization prediction of 12 transmembrane alpha helices. Transmembrane domains function as the main structural units of ABC proteins, while NBD's are water-soluble ATP-binding domains that provide the power for the pump. ABC transporters typically have two nucleotide-binding domains (NBDs) and two transmembrane domains (TMDs). In Equine P-glycoprotein, the two NBD's contain Walker A (GNSGCGKS, GSSGCGKS) motifs between amino acids 425-432 (NBD1) and between amino acids 1066-1073 (NBD2). Both NBD's each have an LSGGQ amino acid sequence (amino acids 526-630, and 1171-1175, respectively). Pfam was used to generate an HMM logo, which was visualized by Skylign (Figures 2A and 2B)

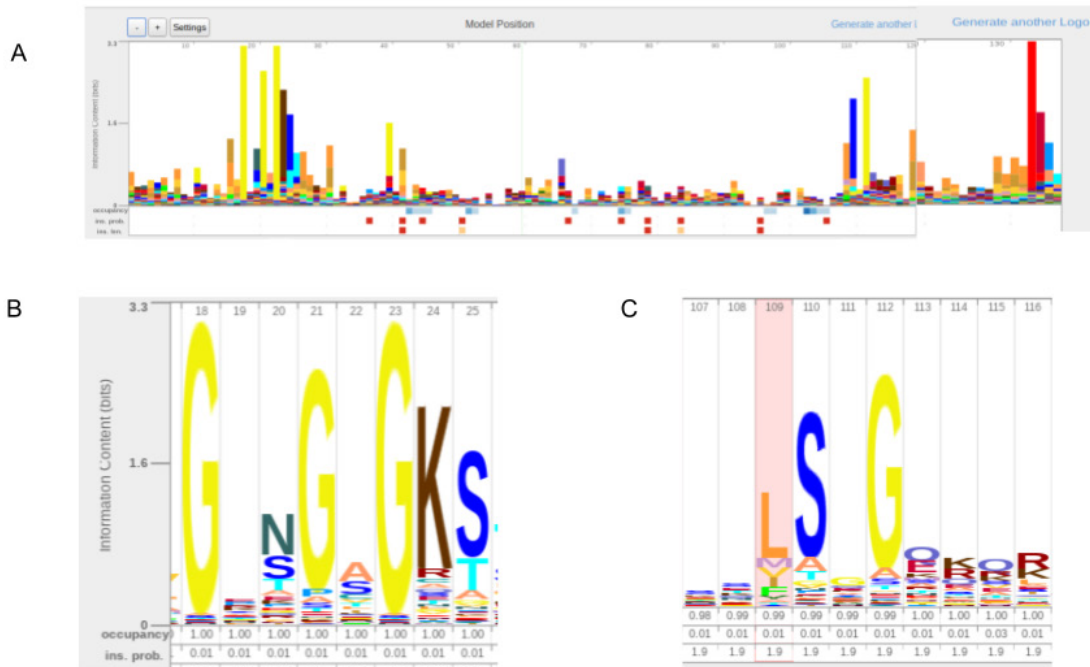


Figure 2. HMM logo and motifs found within it. The HMM logo (A) has the Walker A (B) and LSGGQ (C) motifs found on ABC Tran domains. ABC transporters' ABC Tran domains each have a Walker A and LSGGQ motif, therefore, the HMM logo of the sequence is supported by the structural information available.

The annotation supports the designation of the protein sequence as belonging to Equine P-glycoprotein. Homology modeling was then used to generate a predicted protein model.

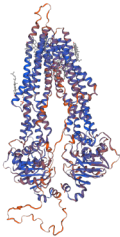
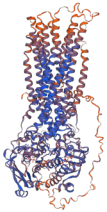
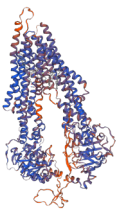
Homology Modeling

Phases of Data

- 1) Three templates were selected from the SWISS-MODEL search.
- 2) These three templates (Template 1: PDB 6qex., Template 2: PDB 6c0v, Template 3: PDB 6qee) had the highest sequence similarities (0.58, 0.58, 0.57) and, with the exception of Template 2, the highest identity scores (91.5%, 91.2%; Template 3 had 88.1 % but was greater in similarity than the HHblits version of that template). Models were derived from these templates (Table 1).

- 3) Based on the structure analyses for all three models, the two selected for further analysis and comparison were 6qex and 6c0v.
- 4) The model based on the template 6c0v (model 2) was selected as the final model based on having a higher Ramachandran favorability than the 6qex (95.7% vs. 94.4%, respectively), having a more favorable torsion score (-2.00 vs. -2.67) and QMEAN score (-2.54 vs. -2.83), all of which indicate a less distorted and more native model based on the amino acid sequence.
- 5) According to the ProCheck data, the model from template 6c0v.1.A. is of “good quality,” based upon its Ramachandran favorability being above 90% (92.3% in most favored regions) and a G-factor score (optimally above -0.5) that indicates that the model is not an unusual structure (0.01 average score).
- 6) The final model after energy minimization had a clash score of 0.0087654 (Figure 3).

Table 1. Template Data and Structure Comparison

<p>Template PDB ID: 6qex. % Sequence identity: 91.5 Found by: BLAST Resolution: N/A Sequence similarity: 0.58 Coverage: 0.97 Description/Name: Multidrug Resistance Protein 1 Method: Electron Microscopy Oligo state: hetero-1-1-1-mer</p> 	<p>Template PDB ID: 6c0v. % Sequence identity: 91.2 Found by: BLAST Resolution: 3.40 angstroms Sequence similarity: 0.58 Coverage: 0.97 Description/Name: Multidrug Resistance Protein 1 Method: Electron Microscopy Oligo state: monomer</p> 	<p>Template PDB ID: 6qee. % Sequence identity: 88.1 Found by: BLAST Resolution: N/A Sequence similarity: 0.57 Coverage: 0.97 Description/Name: ABCB1HMEQ Method: Electron Microscopy Oligo state: monomer</p> 
---	---	---

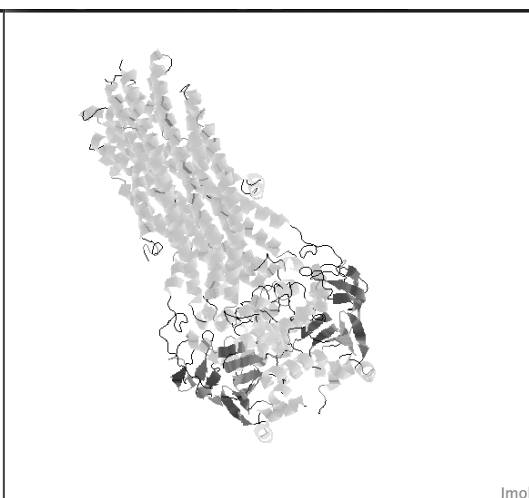
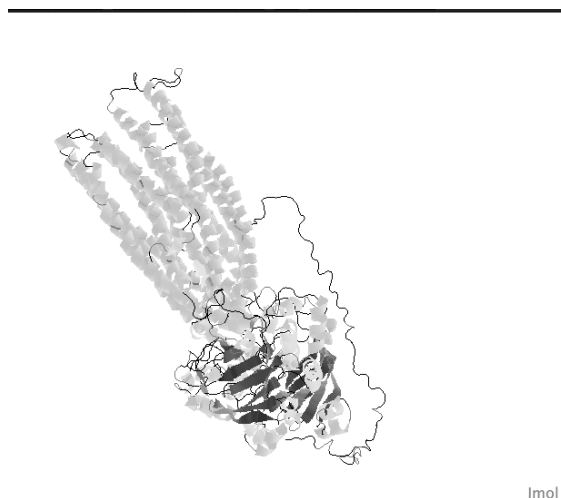


Figure 3. Final energy-minimized model. Original template-based model (left, PDB: 6c0v) beside the final Energy-Minimized model (right). Note that the long unresolved loop (Q625-V691, shown by the arrow) is resolved by the energy-minimization.

Discussion

The production of a 3D model of a horse P-glycoprotein can contribute to a growing body of case studies that can be used for educational purposes. This case study can be used like that of Haddad, Adam, and Heger (2020), in which a model was constructed of the MAM1 domain in ALK receptor to demonstrate a set of instructions for optimizing one's homology modeling. This would be an important tool in college-level courses related to biochemistry, bioinformatics, organic chemistry, etc. We intend to also make a set of instructions that students can easily follow on a set of Powerpoint slides so that they can learn how to find the biochemical properties of proteins of interest and connect these properties to the way the proteins function at the cellular level. In the future, we expect to express the protein and validate its structure through CryoEM, and to explore inhibitory strategies for equine P-glycoprotein such that veterinary pharmaceuticals could effectively be administered to horses. Increasing the understanding of the function of P-glycoprotein in more species by homology modeling and CryoEM is another important endeavor that can further contribute to proteomic literature.

Conclusion

The model based on template (pdb code: 6c0v) had the highest QMEAN score of -2.54 and other favorable factors. The quality evaluations showed that the model selected was accurate for equine P-glycoprotein, but there were some unresolved factors due to 91.2% of sequence alignment with the template and sequence similarity of 58%. There existed a disorganized loop of discrepancies in the sequence (Q625-V691) in the initial template-derived model. Energy minimization resolved some of the distortion of the model. Misalignment of the amino acid sequence will cause disorganized loops in the model, such as the loop in the original model (Q625-V691), however, since this structure was also unresolved in the CryoEM template we hypothesize that this loop could be a highly flexible region. The clash ratio of 8.77×10^{-3} indicated that there are few clashes in the structure. The data supports the final structure of the model as being the most accurate.

Limitations

The annotation and homology modeling was carried out entirely online, and the structural analysis of the protein was based on pre-existing structural information and various software programs that analyzed an entered amino acid sequence. There are general and inherent limitations to using homology modelling to construct novel models of protein structures. The annotation information is limited to proteins that were of interest to other annotators. It is also important to note that even among similar protein structures, environmental factors, and even intramolecular conditions, such as the shifting of protein domains, can cause the functions of these proteins to vary in ways that may impact the reliability of using homologous structures to construct new models (Launay and Simonson, 2008). These limitations are exacerbated when homology modeling is used to study protein complexes (Launay and Simonson, 2008). Also, by depending solely on online tools rather than experimental structure-determination methods, such as CryoEM, of isolated horse P-glycoprotein, it is possible that the structure could be slightly different in nature than the expected structure produced by homology modeling. More information about its interactions with specific molecules can be found if this study is extended to include biochemical assays involving the protein, such as transport assays and substrate binding assays.

Acknowledgements

We would like to thank Trinity Washington University, the DC NASA Space Grant Consortium for the opportunity to conduct and present this research and Dr. George Lountos for expert advice.

This work was prepared while Dr. Karobi Moitra was employed at Trinity Washington University. The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

References

Alam, A., Kowal, J., Broude, E., Roninson, I., & Locher, K. P. (2019). Structural insight into substrate and inhibitor discrimination by human P-glycoprotein. *Science (New York, N.Y.)*, 363(6428), 753–756. <https://doi.org/10.1126/science.aav7102>.

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., Schwede, T. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Scientific Reports* 7, 10480. <https://doi.org/10.1038/s41598-017-09654-8>.

Benkert, P., Künzli, M., & Schwede, T. (2009). QMEAN server for protein model quality estimation. *Nucleic acids research*, 37(Web Server issue), W510–W514. <https://doi.org/10.1093/nar/gkp322>.

Bienert, S., Waterhouse, A., de Beer, T. A., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2017). The SWISS-MODEL Repository-new features and functionality. *Nucleic acids research*, 45(D1), D313–D319. <https://doi.org/10.1093/nar/gkw1132>.

Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., & Schwede, T. (2009). Protein structure homology modeling using SWISS-MODEL workspace. *Nature protocols*, 4(1), 1–13. <https://doi.org/10.1038/nprot.2008.197>

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L. BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421-430 (2009). <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-421>

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic acids research*, 47(D1), D427–D432. <https://doi.org/10.1093/nar/gky995>

Finch, A. and Pillans, P. (2014). P-glycoprotein and its role in drug-drug interactions. *Australian Prescriber*, 37, 4 (Aug 2014). doi:10.18773/austprescr.2014.050

Goodsell, David. (2010). PDB-101: Molecule of the Month: P-glycoprotein. PBD. doi:10.2210/rcsb_pdb/mom_2010_3

Guex, N., Peitsch, M.C., Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* 30, S162-S173 .

- Haddad Y., Adam V., Heger Z. (2020) Ten quick tips for homology modeling of high-resolution protein 3D structures. *PLoS Comput Biol* 16(4): e1007449. <https://doi.org/10.1371/journal.pcbi.1007449>
- Hofmann, K. and Stoffel, W. (1993). TMBASE—A database of membrane spanning protein segments. *Biol. Chem. Hoppe-Seyler* 374 166. https://embnet.vital-it.ch/software/TMPRED_form.html
- Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic acids research*, 35(Web Server issue), W585–W587. <https://doi.org/10.1093/nar/gkm259>
- Käll, L., Krogh, A., & Sonnhammer, E. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic acids research*, 35(Web Server issue), W429–W432. <https://doi.org/10.1093/nar/gkm256>
- Kopcho, N., Chang, G., & Komives, E. A. (2019). Dynamics of ABC Transporter P-glycoprotein in Three Conformational States. *Scientific reports*, 9(1), 15092. <https://doi.org/10.1038/s41598-019-50578-2>
- Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26, 283-291. <https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/>
- Launay, G., Simonson, T. (2008). Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. *BMC Bioinformatics* 9, 427 <https://doi.org/10.1186/1471-2105-9-427>
- Linardi, R. L., Stokes, A. M., & Andrews, F. M. (2013). The effect of P-glycoprotein on methadone hydrochloride flux in equine intestinal mucosa. *Journal of veterinary pharmacology and therapeutics*, 36(1), 43–50. <https://doi.org/10.1111/j.1365-2885.2012.01390.x>
- Loo, T. W., & Clarke, D. M. (2013). Drug rescue distinguishes between different structural models of human P-glycoprotein. *Biochemistry*, 52(41), 7167–7169. <https://doi.org/10.1021/bi401269m>
- Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S., & Richardson, D. C. (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins*, 50(3), 437–450. <https://doi.org/10.1002/prot.10286>
- Mary Vore. (2019). ABCB subfamily (version 2019.4) in the IUPHAR/BPS Guide to Pharmacology Database. *IUPHAR/BPS Guide to Pharmacology*. DOI: <https://doi.org/10.2218/gtopdb/F152/2019.4>.
- Natalini, C.C., and Linardi, R.L. Identification of multi-drug resistance gene (MDR1) in equine ileum. (2006). *Ciencia Rural*, Santa Maria. 36, 1, 298-300.
- O'Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq)

database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* (2016 Jan 4);44(D1):D733-45 PubMed. <https://www.ncbi.nlm.nih.gov/refseq/>

Ramachandran, S., Kota, P., Ding, F., & Dokholyan, N. V. (2011). Automated minimization of steric clashes in protein structures. *Proteins*, 79(1), 261–270. <https://doi.org/10.1002/prot.22879>

Remmert, M., Biegert, A., Hauser, A., Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 9, 173-175 (2012). <https://www.nature.com/articles/nmeth.1818?message-global=remove>

Sonnhammer, E. L. L., von Heijne, G. and Krogh, A.. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, 175-182 (1998). <http://www.cbs.dtu.dk/services/TMHMM/>

Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., & Schwede, T. (2020). QMEANDisCo-distance constraints applied on model quality estimation. *Bioinformatics (Oxford, England)*, 36(6), 1765–1771. <https://doi.org/10.1093/bioinformatics/btz828>

Tusnády, G. E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics (Oxford, England)*, 17(9), 849–850. <https://doi.org/10.1093/bioinformatics/17.9.849>

Wang, J., Youkharibache, P., Zhang, D., Lanczycki, C. J., Geer, R. C., Madej, T., Phan, L., Ward, M., Lu, S., Marchler, G. H., Wang, Y., Bryant, S. H., Geer, L. Y., & Marchler-Bauer, A. (2020). iCn3D, a web-based 3D viewer for sharing 1D/2D/3D representations of biomolecular structures. *Bioinformatics (Oxford, England)*, 36(1), 131–135. <https://doi.org/10.1093/bioinformatics/btz502>

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46(W1), W296-W303. <https://swissmodel.expasy.org>

Wheeler, T.J., Clements, J. & Finn, R.D. (2014). Skylogn: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics* 15, 7 <https://doi.org/10.1186/1471-2105-15-7>