The Membrane Mutation Effect Classifier (MMEC): A Novel-Structure Based Approach to Predicting the Functional Effects of Mutations in Membrane Proteins

Rebecca Alford

Part I

I always loved career day as an elementary student because I was able to share that my dad was a rocket scientist. Maybe he was not the astronaut flying into space or sitting in the control room, but I believed he had the 'coolest' job because he was the engineer – designing new space cameras. The process mesmerized me: my mind could not yet even wrap itself around the idea of creating something new. Therefore as I reached the high school level, I fell in love with innovation – the idea of creating something new in order to solve some real world problem.

My passion for innovation was somewhat out of the ordinary because I was facing a challenge that was very real for me. At age 5, I was diagnosed with a rare genetic condition that results in severe visual impairment. Through my various Google and WebMD queries, I found that there were limited answers relating to the diagnosis and treatment of my condition. However, as I matured I realized that I did not need to wait for other scientists to find the answers: I could find them myself.

From there, I jumped in to research with the help of the science research program as well as teachers Mr. Kurtz and Ms. Collette to engage in my first research project involving computational biology. I was particularly interested in computational biology because I had programed with my dad before and became very interested in intertwining the fields of computer science and biomedical research. In this project, I tried to manipulate protein structures involved in my particular visual impairment by making mutations on the computer and trying to measure

the effects on their structure. The project was somewhat successful; however, at my first science fair a judge asked an interesting question. She asked “Couldn’t you apply this idea to other proteins?” Suddenly, I had realized that I was on the right path to understanding one of the fundamental questions in biomedical research: connecting protein structures with function.

After several all-night brainstorming sessions, I approached my research teacher Mr. Kurtz one morning to present my idea. I said that I wanted to invent a computer program that could predict the effects of mutations in disease. Staring at me wide-eyed and confused, my teacher flat out told me I was crazy, yet believed in me enough to encourage me to follow my goal and start building.

For the first year and a half, I was essentially on my own: where my version of a ‘high-tech’ laboratory extended to my bedroom walls. However, though I did not have a professional mentor available, I was able to use my resources and collaborate with various scientists from around the United States and around the World. Through this experience I gained various insights. In this sense, instead of just having a mentor in one specific field, I was able to get the best of an interdisciplinary experience by having various mentors from various different fields. By the end of this first period, I had a working prototype of a program that would use mutation information in order to infer its functional effects.

At this point, I realized that my prototype was ready to become a finished product that the scientific community could ultimately benefit from. However, I needed some assistance to make this vision become a reality. I had been collaborating with five different labs from across the United States and ultimately decided to team up with the Bonneau Lab at NYU as my new mentor, Dr. Richard Bonneau, had extensive experience in software development for biomedical applications. Together, along with various graduate students, Post-Doctoral Fellows, and

undergraduate students, we worked together to improve the science behind my program. Finally, after almost four years of hard work my goal was realized and my program can make successful predictions of the functional effects of mutations at 87% accuracy.

Part II

The connection between protein structure, function, and disease is critical in improving both laboratory research and current approaches to diagnosis and treatment. For instance, in engineering new proteins for biomedical and environmental applications, it is necessary to make mutations that will modify and optimize the structure of a protein without deteriorating function. Further, connecting mutations in genes to phenotypes is a valuable tool in the diagnosis and treatment of rare genetic disorders whose associated genes and proteins are not yet characterized. Therefore, it is critical to link mutations in proteins with their resulting phenotypes. However, the study of protein function in the laboratory is complicated by the absence of tools that predict protein structure to function. Thus, researchers are often left testing specific sites in very large proteins which is time consuming, expensive, and provides little information regarding the function of that protein.

Computational models have successfully been applied to soluble protein structure models toward predicting the effects of mutations on protein function given the ease in accessing soluble protein structures in a publicly available database known as the Protein Databank (PDB). However, experimental determination and computational modeling of membrane protein structures is complicated by their complex interdependence on the membrane environment. Nonetheless, understand membrane protein structure and function is essential as these proteins comprise over 30% of proteins in the human body and are associated with several genetic

disorders. It is thus essential to design a system that considers these membrane-protein specific features to yield accurate predictions of functional effects due to mutations in membrane proteins. The ultimate goal of my research was to build a tool that could clarify this picture for researchers where the effects of mutating specific sites in a protein are known (fig. 1).

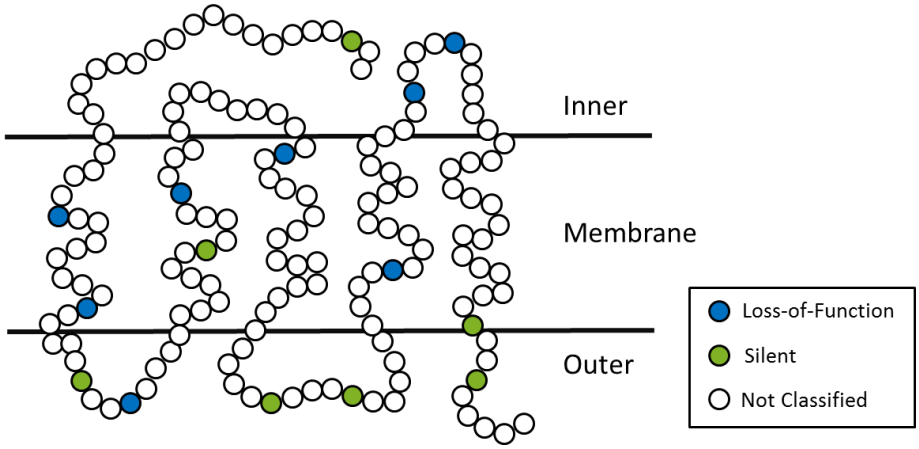

Figure 1 - The search for loss-of-function and silent mutations often involves scanning multiple positions (circles) and testing for the desired effect (loss of function or silent).

Through research and collaborating with other scientists, I found that applying machine learning to this problem would be a suitable approach to building this tool. Machine learning provides several benefits in the scope of biology by 'learning' complex relationships within data sets. For example, one method used machine learning to train a classifier that can predict whether a site on a protein structure is sensitive to temperature changes. Similarly, a method used machine-learning based classifiers to predict the functional effects of mutations in soluble (non-membrane) proteins. More so, machine learning provides the ability to adapt and change classifier training based on specific features in the problem such as biochemical changes due to the mutation, stability of the protein structure, as well as membrane environment features.

To train classifiers, I gathered a set of 85 different protein structure and sequence based features that could describe over 200 known mutation. I trained six different classifiers and here I represent the performance of the top three performing classifiers (called LIN-all, RBF-seq, and RBF-str). Each classifier was evaluated using four metrics: accuracy, precision, receiver operating characteristic (ROC) curve and area under the ROC curve (AUROC). Percentage of instances correctly classified (accuracy in prediction) were evaluated using aggregate output statistics from training each model and assessed for each class. (Fig. 2).

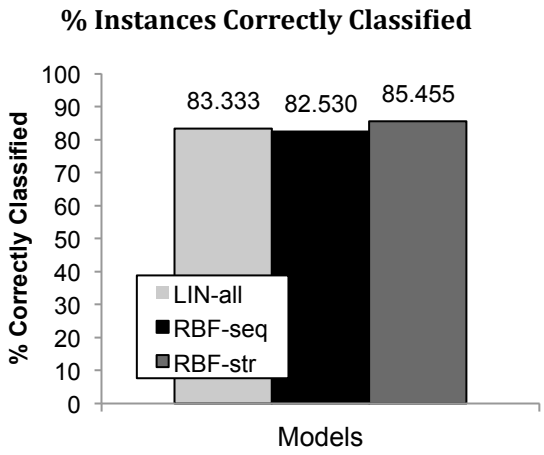

Figure 2: Classifier Accuracy
Classifier accuracy based on the percentage of samples correctly classified during 5x leave-out CV

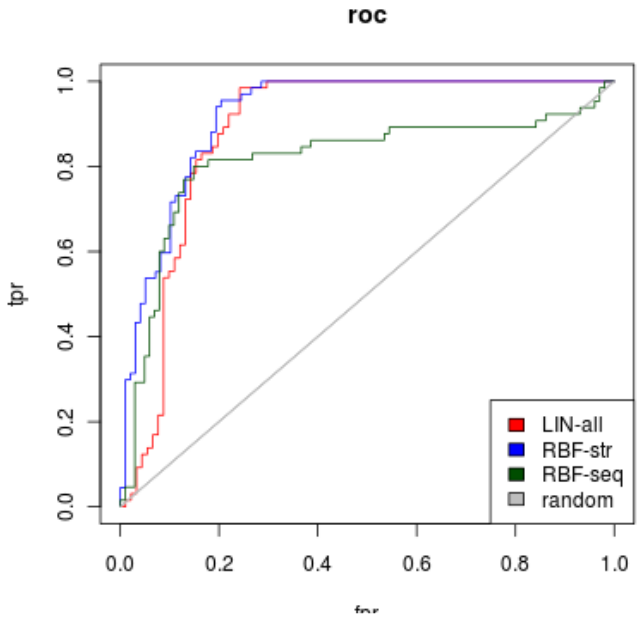

Figure 3: ROC Analysis
ROC analysis for the top 3 SVMs. Steep curves reflect a high true positive rate, illustrating a high probability of correct classification for each new sample.

The measures for accuracy and precision reflect overall performance of both classes (not shown) and total model performance. However an essential characteristic of SVM training is to understand how individual samples contribute to overall performance. This is achieved using an ROC curve and AUROC. The Receiver Operator Characteristic evaluates models based on the probability of correct vs. incorrect classification using true

positive and false positive rates from 5x leave-out CV. Thus, a greater AUROC value reflects a higher probability of correct classification; supporting better performance (AUROC for random binary predictions equals 0.5). ROC and AUROC analysis was completed for all SVM models, as shown in Figure 3.

The main objective in developing this method is to not only validate that each classifier performs well, but to also build a system that can generate highly-accurate predictions on new samples not included in the original training set. A set of 20 new proteins corresponding to 100 different loss-of-function and neutral mutations was obtained from various literature sources. Each sample was then tested with the classifier to test if the classifier's prediction matched with the known functional effect of the given mutation.

Predictions on new test proteins exceeded the original expected significance for each prediction. Based on the original testing, the classifier was expected to make predictions for new samples at 0.860 – however, the mean significance of the new test proteins was 0.958 with a standard deviation of 0.070. This is critical as it illustrates that the classifier can make accurate predictions to samples similar to those in the training set, as well as new samples with different characteristics.

By integrating a combination of sequence- and structure- based features, the best classifier (LIN-all) was selected for use with the program. Using the various methods described here, it was demonstrated that the protocol with this classifier successfully classifies mutations with a top accuracy at 95.8%. More importantly, testing, training, validation, and individual protein analysis are all indicators that LIN-all will not only perform well in classifying new samples that are similar to those in the training set, but also to those that expand those original features.

To date, function-prediction methods optimized for soluble proteins performed with 81% accuracy at best. Therefore, this new function-prediction method for membrane proteins performs 15% better than top performing methods for soluble proteins. This not only illustrates the strength of the machine-learning based algorithm in accurately predicting functional effects, but also exhibits that specific structural information as well as selected sequence-based features are critical in the prediction of mutations in membrane proteins. Furthermore, despite restrictions posed by the membrane environment in general structure prediction, this novel method is highly successful in predicting loss-of-function phenotypes.

The applications of this protocol extend to various research areas, from understanding membrane protein structures to interpreting genetic variation in a wide array of genes and proteins associated with disease. This protocol can significantly minimize search space for loss-of-function mutations, thus increasing efficiency in protein engineering, identifying new drug targets and studying membrane proteins *in vitro* and/or *in vivo*. Further, the ability, with this new protocol to classify mutations allows researchers to connect molecular-level changes to the wide spectrum of phenotypes observed in genetic disorders, providing new opportunities to improve diagnosis and treatment. By providing access to this newly developed protocol, it is hoped that this resource can contribute to current research and open many new opportunities by providing novel information regarding the functional effects of mutations in membrane proteins.

References

1. Barth, P., Schonbrun, J., & Baker, D. (2007). Toward high-resolution prediction and design of transmembrane helical protein structures. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(40), 15682-7.

2. Binda, C., Newton-Vinson, P., Hubálek, F., Edmondson, D. E., & Mattevi, A. (2002). Structure of human monoamine oxidase B, a drug target for the treatment of neurological disorders. *Nature structural biology*, *9*(1), 22-6.

3. Sanders, C. R., & Myers, J. K. (2004). Disease-related misassembly of membrane proteins. *Annual review of biophysics and biomolecular structure*, *33*, 25-51.

4. Hopkins, A. L., & Groom, C. R. (2002). The druggable genome. *Nature reviews. Drug discovery*, *1*(9), 727-30. doi:10.1038/nrd892

5. Bowie, J. U. (2005). Solving the membrane protein folding problem. *Nature*, *438*(7068), 581-9. doi:10.1038/nature04395

6. Engelman, D. M., Chen, Y., Chin, C.-N., Curran, A. R., Dixon, A. M., Dupuy, A. D., Lee, A. S., et al. (2003). Membrane protein folding: beyond the two stage model. *FEBS Letters*, *555*(1), 122-125.

7. Skach, W. R. (2009). Cellular mechanisms of membrane protein folding. *Nature structural & molecular biology*, *16*(6), 606-12. doi:10.1038/nsmb.1600

8. Ferguson, S. S. (2001). Evolving concepts in G protein-coupled receptor endocytosis: the role in receptor desensitization and signaling. *Pharmacological reviews*, *53*(1), 1-24. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11171937

9. Senes, A., Engel, D. E., & DeGrado, W. F. (2004). Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Current opinion in structural biology*, *14*(4), 465-79

10. Von Heijne, G. (2006). Membrane-protein topology. *Nature reviews. Molecular cell biology*, *7*(12), 909-18.

11. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) The Protein Data Bank, *Nucleic Acids Research, 28*:235-242.

12. Joosten, R. P., te Beek, T. a H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Sander, C., et al. (2011). A series of PDB related databases for everyday needs. *Nucleic acids research*, *39*(Database issue), D411-9.

13. Krieger, E., Joo K., Lee J., Lee J., Raman S., Thompson J., et al. (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins*. 77 Suppl9, 114-22.

14. Bonneau, R., Tsai J., Ruczinski I., Chivian D., Rohl C., Strauss C. E., et al. (2001). Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins*. Suppl 5, 119-26.

15. Leaver-fay, A., Tyka, M., Lewis, S. M., Lange, F., Thompson, J., Jacak, R., Kaufman, K., et al. (2011). ROSETTA 3 : An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Cancer Research*, *487*(11), 545-574. doi:10.1016/S0076-6879(11)87019-9

16. Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry*, *49*(14), 2987-98.

17. Yarov-Yarovoy, V., Schonbrun, J., & Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. *Proteins*, *62*(4), 1010-25. doi:10.1002/prot.20817

18. Bromberg, Y., & Rost, B. (2009). Correlating protein function and stability through the analysis of single amino acid substitutions. *BMC bioinformatics*, *10 Suppl 8*, S8.

19. Kellogg, E. H., Leaver-Fay, A., & Baker, D. (2010). Role of conformational sampling in computing mutation induced changes in protein structure and stability. *Proteins: Structure, Function, and Bioinformatics*, n/a-n/a.

20. Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, *24*(12), 1565-7.

21. Poultney, C. S., Butterfoss, G. L., Gutwein, M. R., Drew, K., Gresham, D., Gunsalus, K. C., Shasha, D. E., et al. (2011). Rational Design of Temperature-Sensitive Alleles Using Computational Structure Prediction. (V. N. Uversky, Ed.)*PLoS ONE*, *6*(9), e23947.

22. Ng, P. C., & Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annual review of genomics and human genetics*, *7*, 61-80.

23. T.Kawabata, M.Ota, and K.Nishikawa (1999). The Protein Mutation Database. *Nucleic Acids Research,* 27:355-357.

24. M. Preising (2011). Retina International Mutations Database. http://www.retina-international.org/index.php?menuid=59

25. Magrane M. and the UniProt consortium (2011) UniProt Knowledgebase: a hub of integrated protein data

26. N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali. Comparative Protein Structure Modeling With MODELLER. *Current Protocols in Bioinformatics,* John Wiley & Sons, Inc., Supplement 15, 5.6.1-5.6.30, 2006.

27. TOPCONS: consensus prediction of membrane protein topology. Andreas Bernsel, Håkan Viklund, Aron Hennerdal and Arne Elofsson (2009) *Nucleic Acids Research* 37(Webserver issue), W465-8

28. Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, *33*(Database issue), D501-4

29. Hall, M. A., & Witten, I. H. (2010). WEKA — Experiences with a Java Open-Source Project. *Journal of Machine Learning Research*, *11*, 2533-2541.

30. Burges, C. J. C. (1997). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *43*, 1-43.

31. Reddy, C. S., Vijayasarathy, K., Srinivas, E., Sastry, G. M., & Sastry, G. N. (2006). Homology modeling of membrane proteins: a critical assessment. *Computational biology and chemistry*, *30*(2), 120-6. doi:10.1016/j.compbiolchem.2005.12.002