Machine learning to accurately predict eukaryotic expression of chimeric membrane proteins

***Introduction***

The expression and proper localization of integral membrane proteins in eukaryotic systems is a significant hurdle in their study and engineering. Between co-translational membrane insertion, rigorous quality control, and multi-step trafficking, expression and localization of integral membrane proteins represent exceptionally complex sub-cellular processes.

1. The importance and difficulty of integral membrane protein expression and localization
   1. *Background:* Membrane protein expression is a multi-step, complex process.
   2. *Motivation:* The study of membrane proteins is vital. membrane protein engineering is under-developed. Simply expressing a membrane protein is often a limiting factor in its study and development.
      1. Introduce GPCRs and opsins (ChRs) as 7TM helix membrane proteins with applications as drug targets and optogenetic tools respectively
      2. If it does not express and localize it will not be functional
      3. As a first step to engineering ChR function we need to confidently predict whether a chimera will or will not localize from its primary sequence.
   3. *Goal:* To overcome this hurtle, we developed data-driven models to predict expression and localization of integral membrane channelrhodopsins in eukaryotic (human) cells. Despite our limited understanding of the exceptionally complex cellular processes involved, the models that we present here for channelrhodopsin are capable of predicting the likelihood of expression and plasma membrane localization from an amino-acid sequence.
2. Channelrhodopsin (ChR) background.
   1. *Background:* ChRs and other opsin proteins are light sensitive membrane proteins found in photosynthetic algae and microbes that are used in nature as either sensors (move to or away from light) or for energy production (pump ions against the natural gradient). ChRs are light gated, passive ion channels that were originally found in photosynthetic algae. Upon light activation these membrane channels open and allow predominantly positive ions flow into the cell with the natural ion gradient.
   2. *Application:* Neurobiologists have taken advantage of this light gated machinery to control neuronal activation with light. ChRs have been extensively applied *in vivo* in animal to interrogate neuronal circuitry.
   3. *Engineering:* These opsin tools have limitations that we would like to overcome through screening natural opsin diversity or through protein engineering.
   4. **One major limitation to the development of novel engineered opsin proteins or applying new opsins found in nature is that many of them come from algal or bacterial hosts and do not express well in eukaryotic cells of interest in neurobiology (e.g. mammalian cells).**
3. Making predictions using machine learning
   1. Here we have explored a number of different data driven models to predict membrane protein expression and localization simply from the primary amino acid sequence.
   2. *Limitations of existing methods:* Given the complexity of the expression and localization there is no physical model that can be used to calculate the probability of expression and localization. Such simulations would require extremely long simulations time (105 times the length of kinetic simulations) for the highly detailed and integrated physics.
   3. Machine learning is ideal for predicting expression and localization because it does not require prior knowledge of the complex biological process and instead can generate predictive models for an output (expression and localization data) based on a simple input (protein amino acid sequence).
   4. Figure 1: Schematic showing how machine learning allows us to skip the details of expression and localization processes in the cell.
   5. *Brief background on machine learning:* Machine learning techniques used include linear models, random field models, and partial order optimum likelihood.
   6. In this study we have generated a library of diverse sequences (the training set) and characterized them for a properties of interest. We can learn from these sequences and predict how other novel sequences will behave as long as the novel sequences have some identity with the training set.
4. Using recombination for diversification.
   1. We want our expression and localization models to be useful for a wide diversity of ChR variants. Thus we need the data set of input sequences used for training the model to be highly diverse, spreading across sequence space for opsin proteins. We can do this most effectively by taking advantage of natural diversity and using chimeragenesis to recombine parts of different parental sequences.
   2. *Previous opsin recombination:* Chimeragenesis of related proteins is a proven method to generate opsin proteins that express and localize, but with recombination of highly diverse sequences there is also greater potential for producing proteins incapable of insertion, folding, expression, and trafficking.
   3. *Introduce SCHEMA on soluble proteins:* We want to build libraries of chimeras where most of the chimeras fold. To build such libraries we have used structure-guided recombination methods previously developed and used for soluble proteins.
   4. *End with what we have built and tested:* We have designed and systematically tested two structure-guided recombination libraries that incorporated diversity from three ChR parental sequences each with less than 53% identity.

***Results***

1. Parental analysis
   1. Selecting the parents.
   2. Assay description and placement of the SpyTag on parents.
   3. Collection of expression and localization data in 96-well plate
   4. Figure 2: Parental analysis
   5. Supp Figure S1: Image processing for expression and localization data
   6. Parents show distinct expression/localization levels.
   7. Assay functions well
2. Chimera analysis
   1. Description of library design
   2. Figure 3: Block locations on C1C2 structure
   3. Supp Figure S2: Alignment with 2nd structure and color codes for 2 libraries
   4. Description of 1st generation test set, data collection, and analysis
   5. Supp Figure S3: Example images from 96-well plate
      1. Images of representative chimeras with three channels
   6. Figure 4: Assay data for Chimeras
      1. Trends via sorted chimeras colored by block
      2. Functional diversity shown via histograms
   7. Supp Table S1. 1st generation Chimera data
      1. Library; ID; block code; E; mutations;
      2. Measured values of expression; localization; localization efficiency
   8. 65% express and localize above threshold
3. Model building and validation
   1. Linear regression models
      1. Supp Table S2. Predictive quality of alternative models
         1. For each block and all blocks together
      2. Not good enough
   2. Lasso model
      1. To identify regions of interest for possible sequence determinates of localization/localization eff. Not expecting it to be as good as GPM.
      2. Model building
         1. About 5,000 parameters can be compressed into 400 that do not covary.
         2. About 24 of these have non-zero weighting.
   3. GP Model details
      1. GPM benefits
         1. Combines data from both libraries
         2. Takes cooperative effects into account
         3. Supp Figure S4. Visualization of contacts and kernel development
      2. Model development for classification and linear regression
      3. Optimal data entry (e.g. log\_GFP), kernel choice
         1. Add other models and kernels including all regression models to Supp Table S2. Predictive quality of alternative models
      4. Figure 5. Measured vs predicted values for classification
         1. For expression, localization, and localization efficiency
         2. High/low chimeras in two colors
         3. Cross validated
         4. Supp Figure S5. ROC curves
      5. Learning curves for GP classifiers
         1. Supp Figure S6. Learning curves for GP classifiers
            1. AUC vs number of chimeras used to train
         2. With only maximally informative sequences and with both max informative and monomeras
         3. Show that a smaller set of 1st generation sequences could have been used to train the classifier, but that maximally informative sequences were particularly important for training
      6. Compare Lasso predictions to GPM
         1. Which is more predictive?
         2. For what sequences do they show the most distinct differences in prediction?
            1. Are there any similarities in these sequences?
      7. From cross-validation, the GP classifier is highly predictive
   4. 2nd generation sequences
      1. Predicted by GP classifiers, but can be used to test GP regression and Lasso
         1. At least 16 mutations removed from any parent and ## mutations away from any chimera
      2. Add measured/predicted values to Figure 5.
         1. 2nd generation chimeras (not used to train) shown in third color
      3. Supp Table S3. 2nd generation Chimera data
         1. Library; ID; block code; E; mutations
         2. Predicted classification and values for expression; localization; localization eff
         3. Measured values for expression; localization; localization eff
      4. Those that were predicted to localize do
      5. Those that were predicted to localize with high efficacy do

***Discussion***

1. ***Through moderate-throughput data collection and advanced statistical modeling, we show that integral membrane protein expression and plasma membrane localization can be predicted***. With this powerful tool we can confidently identify expressing and localizing ChR variants from a library of nearly 120,000 chimeras.
   1. Given the complexity inherent in these cellular processes it is surprising and impressive that we can make accurate predictions based solely on amino acid sequence. Moreover, we can manage a comparable level of predictive power with a training set of only 50 chimeras.
2. We can also glean information on the sequence determinates of ChR chimera expression
   1. Monomeras
      1. When single blocks are swapped in what effect is seen with localization & loc. efficiency
   2. Different features of high and low localizing sequences
      1. From data and predictions
      2. Figure 6.
   3. Similarities and differences between sequences predicted to have high expression, high localization, and high localization efficiency
      1. Figure 6. 3D plot for 3 properties
      2. What chimeras fall along the correlation line?
      3. What is similar between the chimeras?
   4. From Lasso, weighted contacts and residues for localization/localization eff.
      1. Where are they? Figure 6. Heat map
      2. Possible reasons for this site’s importance
         1. Hypothesis testing
3. Applying the models
   1. In future engineering efforts of ChR, where low-throughput assays will be used to quantify a range of channel properties, we can use the models developed here to narrow our focus on well-localizing ChR chimeras.
   2. Furthermore, the modeling methodology that we develop here with a moderate number of ChR sequences can be applied other channel properties for which a more limited set of data will be available
      1. Subsequent work will model these properties for second-generation sequences with optimal channel properties and the stipulation that they also localize well.
   3. Augmenting the localization of ChRs outside the parental set via limited chimeragenesis
      1. Model expansion