Protein-Protein Interaction Detection Using Mixed Models

Andrew Best, Andrea Ekey, Alyssa Everding, Sarah Jermeland,
Jalen Marshall, Carrie N. Rider, and Grace Silaban

July 25, 2013

Summer Undergraduate Research Institute in Experimental Mathematics
(SURIEM)
Lyman Briggs College
Michigan State University

Abstract

Membrane protein-protein interactions (PPI) play an important role in biological processes; however, knowledge about membrane proteins is limited. Mating-based Split Ubiquitin System is a technique used to investigate interactions between proteins by utilizing yeast as a heterologous system. The observed fluorescence scores are a result of PPIs. The fluorescence scores may be affected by various fixed and random effects such as overall mean, test versus positive controls, plate effect, and PPI effect. We model these effects using a statistical mixed model and apply it to a simulated data set. The success of the simulation study implies that our mixed model may fit the actual PPI data.

1 Introduction

Interactions among proteins are important in all biological processes; however, there is little information known about these interactions. One of the primary functions of protein-protein interactions (PPI) is to allow organisms to adjust to changes in their environment. Due to this, developing a knowledge of these interactions is relevant to learning about the adaptation of organisms. For example, they control cell permeability and allow organisms to acclimate to changing conditions by coordinating internal transport and metabolism [4]. Identifying the PPI is especially important in plants since plants are sessile organisms that must be uniquely efficient in adapting to environmental changes due to their immobility. The data collected for this research comes from *Arabidopsis thaliana* proteins. The *Arabidopsis* genus is widely used in research due to its short genome. Thus, it is ideal for a lab to use this genus for a high-throughput method.

One of the benefits of using high-throughput methods is that a single lab can carry out a whole screen and generate a complete dataset collected under comparable conditions. There are, however, potential drawbacks in using high-throughput methods as well, such as the tendency for high noise. Thus, a large number of replications is necessary to account for noise [4].

In order to identify PPIs, plant mating-based Split-Ubiquitin System (mbSUS) has been developed and used to detect interactions of translocon complex at the outer chloroplast membrane. The system uses yeast as a heterologous system, which relies on the release of a transcription factor if there is an interaction among membrane proteins [1].

The process of mbSUS requires the use of a ubiquitin protein which has been split into two

halves. The N-terminal domain of ubiquitin (Nub) can reconstitute a functional ubiquitin when co-expressed with its C-terminal ubiquitin half (Cub). In this system, protein 'X' is fused to Nub, and protein 'Y' is fused to Cub. When 'X' interacts with 'Y', a Nub and Cub come together, and the ubiquitin molecule is reconstituted. This action triggers the release of an artificial transcription factor, causing the activation of transcription marker genes to produce yeast fluorescence [4].

Wild-type protein pairs are studied as a positive control. Positive controls are known to have true interactions. In mbSUS, mutant Nubs with lowered affinity to Cub reconstitute the full-length ubiquitin only when brought into close proximity via interaction of the two fusion partners. When proteins 'X' and 'Y' react using a mutant Nub, it is questionable whether there is a significant PPI. Although the PPI effect cannot be directly observed or measured, it can be examined through the fluorescence scores from the growth of yeast [4].

To provide a more accurate detection of the PPI, we propose a statistical mixed model to improve the modeling of yeast growth. For our purposes, a mixed model is useful since it considers several factors that affect the observed fluorescence of the yeast, including test versus positive controls, variation due to individual plates, and PPI effects. The proposed model is applied to detect interactions between thousands of *A. thaliana* membrane proteins. In our study, we simulate mbSUS data and apply our model to it. The results from our simulation suggest that our model is a reasonable fit for the actual data.

2 Experimental Design

The plant science group [1] divided the proteins of interest into two partially overlapping groups, one of size 778, the other of size $2,151$. We call a protein from the former group a *Cub protein*, since it is attached to the Cub half of the ubiquitin under the mbSUS regime. Similarly, we call a protein from the latter group a *Nub protein*. (These labels are merely convenient.) In this experiment, a Cub protein and a Nub protein comprise a PPI, whose strength is indirectly measured by the amount of fluorescence of the resultant yeast growth.

To perform the experiment, researchers replicated each of $1,550$ unique plate designs four times, for a total of $6,200$ microtiter plates. We use *plate design* to denote the map of all PPIs that are tested on a particular plate. Each plate contains $1,536$ wells, arranged in 48 rows and 32 columns. A plate design divides the plates into two regions - positive control and test. Wells in the control region, also called *control wells*, are located on the boundary of the plate, which is two wells wide. Protein pairs in control wells are known to interact strongly. Consequently, control wells tend to produce high fluorescence scores and provide a frame of reference for other wells. Conversely, *test wells* are located in the center of the plate, the test region.

On a given plate, the Nub protein generally varies across the wells, while the Cub protein is held constant across the plate. As the total number of Cub proteins is 778, many of the $1,550$ total plate designs employ the same Cub protein as other plate designs. Moreover, not all of the Nub proteins on a given plate design are unique. In some plate designs, there are up to 80 duplicated Nub proteins, that is, up to 80 duplicated PPIs. In any case, with this experimental design, the researchers tested $1,417,701$ unique PPIs and generated several

million fluorescence scores.

The resulting data is formatted in spreadsheets such that there are two files for each plate, one for the test region and one for the control region. Each file specifies which Nub protein is in which well position. An additional administrative file specifies which Cub protein is used across a given plate. The files are systematically named for ease of matching a plate-specific Cub protein to a well-specific Nub protein. Thus, barring noise and input errors, the data corresponds a PPI to a fluorescence score and a plate location.

3 Mixed Model for PPI Detection

Let Y_{ijk} be the fluorescence score for the k^{th} replicate of the j^{th} protein pair on the i^{th} plate. We consider sources of variation as follows. First, plate effect accounts for variation between plates. For example, experimental errors might result in contamination of plates or other errors that can significantly affect the fluorescence scores from a particular plate. To represent plate effect, a β_i is assigned to each plate i, such that $i = 1, 2, ..., 6200$. Second, the control effect, represented by λ, depends on whether a particular well contains a control protein pair or a test protein pair.

Finally, PPI effect accounts for the varying strengths of different PPIs. Since our goal is to determine whether two proteins can be said to interact, we are interested in this effect. In the model, we denote PPI effect by τ_j, where $j = 1, 2, ..., 1417701$, the total number of unique PPIs.

Consequently, we model the fluorescence score as follows:

$$Y_{ijk} = \mu + \lambda + \beta_i + \tau_j + e_{ijk}, \tag{1}$$

where μ is the overall mean and e_{ijk} is the pure error. In our model, plate effects and error are assumed to follow normal distributions: β_i is i.i.d. $N(0, \sigma_\beta^2)$ and e_{ijk} is i.i.d. $N(0, \sigma_e^2)$ respectively. If no PPI effect exists, then τ_j is i.i.d. $N(0, \sigma_\tau^2)$. If PPI effect does exist in a protein pair, then the expected value of that τ_j should be positive, that is, $E(\tau_j) > 0$. The two random effects are independent of each other; thus, there are no covariances between them.

4 Parameter Estimation

4.1 MME for Fixed Effects and Random Effects

Prediction theory [5] estimates properties of random variables through data that is sampled from a population with known variance-covariance structures. In this case, the data is sampled via the scores of the test plates. The theorem is regulated by the model

$$\mathbf{y} = \mathbf{X}\mathbf{u} + \mathbf{Z}\mathbf{b} + \mathbf{e}, \tag{2}$$

where $\mathbf{y}$ is an $N \times 1$ vector of observations, $\mathbf{u}$ is a $p \times 1$ vector of unknown constants, $\mathbf{b}$ is a $q \times 1$ vector of unknown effects of random variables, and $\mathbf{e}$ and is an $N \times 1$ vector of unknown residuals. Matrices $\mathbf{X}$ and $\mathbf{Z}$ are design matrices of order $N \times p$ and $N \times q$ respectively. They

relate the elements of $\mathbf{u}$ and $\mathbf{b}$ to the elements of $\mathbf{y}$. Therefore, $\mathbf{Xu}$ denotes the fixed effects, and $\mathbf{Zb}$ denotes the random effects.

For our model,

$$\mathbf{u} = \begin{bmatrix} \mu \\ \lambda \end{bmatrix}, \tag{3}$$

which represents the fixed effects, and

$$\mathbf{b} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\tau} \end{bmatrix}, \tag{4}$$

which represents the random effects.

The null hypothesis, H_o, is that there are no significant PPIs. Assuming H_o is true, the expectation values are then

$$E(\mathbf{e}) = \mathbf{0} \tag{5}$$

and

$$E(\mathbf{b}) = \mathbf{0}; \tag{6}$$

thus,

$$
\begin{aligned}
E(\mathbf{y}) &= E(\mathbf{Xu} + \mathbf{Zb} + \mathbf{e}) \\
&= E(\mathbf{Xu}) + E(\mathbf{Zb}) + E(\mathbf{e}) \\
&= \mathbf{X}E(\mathbf{u}) + \mathbf{Z}E(\mathbf{b}) + E(\mathbf{e}) \\
&= \mathbf{X}(\mathbf{u}) + \mathbf{Z}(\mathbf{0}) + \mathbf{0} \\
&= \mathbf{Xu}.
\end{aligned} \tag{7}
$$

In other words, since the expectation of the errors and the expectation of the random effects are assumed to be zero, the expectation of the observations will be equivalent to the fixed effects. In our model, this includes the overall mean and the control effect. However, when there are significant PPI effects and no plate effects, then the fixed effects will be positive. This shows that the observed scores are positive.

Covariance does not exist between the random effects, $\mathbf{b}$, and errors, $\mathbf{e}$. The covariance structure is then modeled by

$$
Var \begin{bmatrix} \mathbf{b} \\ \mathbf{e} \end{bmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}. \tag{8}
$$

The variables $\mathbf{G}$ and $\mathbf{R}$ are represented by the matrices

$$
\mathbf{G} = Var(\mathbf{b}) = \begin{pmatrix} Var(\boldsymbol{\beta}) & \mathbf{0} \\ \mathbf{0} & Var(\boldsymbol{\tau}) \end{pmatrix} \tag{9}
$$

and

$$\mathbf{R} = Var(\mathbf{e}) = \sigma^2\mathbf{I}. \tag{10}$$

Therefore, Henderson's Mixed Model Equation (MME),

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}, \tag{11}$$

can be used to solve for $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$, where $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$ are the estimates of vectors $\mathbf{b}$ and $\mathbf{u}$ respectively. Now that there are all of the components for the mixed model, (11) can be used. In the matrices $\mathbf{G}$ and $\mathbf{R}$, variance component values are estimated using REML, which is described in Section 4.2.

4.2 REML Estimation for Variance Component

We use Restricted Maximum Likelihood (REML) estimators to estimate the variance components in $\boldsymbol{G}$ and $\boldsymbol{R}$ of Henderson's Mixed Model Equation. Following the description of Corbeil and Searle [2], we rewrite the model as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}_1 u_1 + \cdots + \boldsymbol{Z}_{125} u_{125} + \boldsymbol{e}, \tag{12}$$

where

$\boldsymbol{y}$ is a vector of N observed scores, i.i.d. from $N(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{H}\sigma^2)$.

$\boldsymbol{b}$ is a vector of fixed effects. Here $\boldsymbol{b} = \left[\begin{smallmatrix}\mu \\ \lambda\end{smallmatrix}\right]$.

$\boldsymbol{u}$ is a vector of length 125 (see Section 5) of mutually independent random variables,

such that $\boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_{125} \end{bmatrix}$.

$\boldsymbol{X}$ is a design matrix referencing the vector of fixed effects $\boldsymbol{b}$ to number of observed scores.

$\boldsymbol{Z}$ is a design matrix with 125 number of columns, referencing the vector of random effects $\boldsymbol{u}$ to the number of observed scores.

$\boldsymbol{e}$ is a vector of error or residual effects, with i.i.d. $N(0, \sigma^2)$.

Also, $Var(\boldsymbol{y}) = \boldsymbol{H}\sigma^2$, and we write $\boldsymbol{H}$ as

$$\boldsymbol{H} = \sum_{\ell=1}^{125} \gamma_\ell \boldsymbol{Z}_\ell \boldsymbol{Z}_\ell' + \boldsymbol{I}_N, \gamma_\ell = \frac{\sigma_\ell^2}{\sigma^2}, \tag{13}$$

where $\sigma_\ell^2 = \sigma_\beta^2$ for $\ell = 1, \cdots, 6$ and $\sigma_\ell^2 = \sigma_\tau^2$ for $\ell = 7, \cdots, 125$. In REML, first consider the singular transformation $\boldsymbol{y'}[\boldsymbol{S}\vdots\boldsymbol{H^{-1}X}]$, where $\boldsymbol{S}$ is a symmetric matrix modeled as such:

$$\boldsymbol{S} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X}'. \tag{14}$$

Hence, $\boldsymbol{SX} = \boldsymbol{0}$ and $\boldsymbol{Sy} \sim N(\boldsymbol{0}, \boldsymbol{SHS}\sigma^2)$. In order to avoid singularity of **SHS**, we use an alternative **S** by omitting the n_1^{th}, $(n_1+n_2)^{th}$, $(n_1+n_2+n_3)^{th}, \cdots$, and $(n_1+n_2+\cdots+n_m)^{th}$ rows. This alternate **S** is the **T** matrix with order $(N-m) \times N$. The log likelihood of **Ty**, denoted by λ_1, is given by

$$\lambda_1 = -\frac{1}{2}(N-m)\log 2\pi - \frac{1}{2}(N-m)\log\sigma^2 - \frac{1}{2}\log|\boldsymbol{THT}'| - \frac{1}{2}\boldsymbol{y}'\boldsymbol{T}'(\boldsymbol{THT'})^{-1}\boldsymbol{Ty}/\sigma^2. \tag{15}$$

We find the values of σ^2 and γ_ℓ's by maximizing λ_1. The differentials of λ_1 with respect to σ^2 and γ_ℓ are as follows:

$$\frac{\partial \lambda_1}{\partial \sigma^2} = -\frac{1}{2}(N-m)/\sigma^2 + \frac{1}{2}\boldsymbol{y}'\boldsymbol{T}'(\boldsymbol{THT}')^{-1}\frac{\boldsymbol{Ty}}{\sigma^4} \tag{16}$$

and

$$\frac{\partial \lambda_1}{\partial \gamma_\ell} = -\frac{1}{2}tr[\boldsymbol{Z}_\ell'\boldsymbol{T}'(\boldsymbol{THT}')^{-1}\boldsymbol{TZ}_\ell] + \frac{1}{2}\boldsymbol{y}'\boldsymbol{T}'(\boldsymbol{THT}')^{-1}\boldsymbol{TZ}_\ell\boldsymbol{Z}_\ell'\boldsymbol{T}'(\boldsymbol{THT}')^{-1}\frac{\boldsymbol{Ty}}{\sigma^2}, \tag{17}$$

for $\ell = 1, 2, ..., 125$.

We equate (16) to zero, and find the values for σ^2 as follows:

$$\hat{\sigma}^2 = \boldsymbol{y}'\boldsymbol{T}(\boldsymbol{THT}')^{-1}\boldsymbol{Ty}/(N-m). \tag{18}$$

Since (17) has no analytic solution for γ_ℓ, a numerical method is necessary.

5 Simulated Data

To demonstrate the proposed mixed effect model, we performed a simulation study. The simulation includes 2 unique plate designs, each replicated 4 times, for a total of 8 plates. The dimension of each plate is 24 rows by 16 columns, and there are 4 replicates of each PPI per unique plate design. We arrange control wells on the boundary region, which is two wells wide.

The fluorescence scores $\boldsymbol{Y}$ for this simulation are calculated using our mixed model (1).

To do this, we rewrite the model in matrix form for easier computation:

$$\boldsymbol{Y} = \boldsymbol{X} \begin{bmatrix} \mu \\ \lambda \end{bmatrix} + \boldsymbol{Z} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\tau} \end{bmatrix} + \boldsymbol{e}, \tag{19}$$

where $\boldsymbol{Y}$ is a vector of all observations/fluorescence scores. There are a total of $3,072$ scores from our 8 simulated plates; therefore, $\boldsymbol{Y}$ is a vector of length $3,072$. Since our model assumes that the effects are normally distributed, we simulated β_j from $N(0, \sigma_\beta^2)$ with $\sigma_\beta^2 = 5,000^2$, and τ_k from $N(0, \sigma_\tau^2)$ with $\sigma_\tau^2 = 6,000^2$. Error values e_{ijk} are also sampled from normal distribution, i.e., $N(0, \sigma^2)$ with $\sigma^2 = 1,000^2$. We set the mean value, μ, at $5,000$, and the control effect, λ, at $15,000$.

The design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ have entries "0" and "1", as they indicate whether each score relates to a specific factor. A "0" indicates no relation; a "1" indicates that the score is related to that factor. The $\boldsymbol{X}$ matrix corresponds with the vector of the two fixed effects μ and λ; therefore, its dimensions are $3,072$ rows by 2 columns. The first column of $\boldsymbol{X}$ only has entries of "1" because all scores are affected by the overall mean. In the second column, only entries corresponding to control wells have a "1", whereas entries for test wells have a "0".

Similarly, the vector of plate effects and PPI effects is multiplied by the $\boldsymbol{Z}$ matrix. This is a vector of length 125 for all the plate effects and PPI effects. Among these, the first 6 correspond to plate effects, and the other 119 correspond to PPI effects. We only include 6 plate effects, instead of all 8, to avoid multicollinearity. There is only one control PPI for all the control wells. In addition, each plate simulates 60 distinct test PPIs. Among the 8 plates, there are 2 distinct plate designs, so there exist 121 distinct PPI effects. Again, to

avoid multicollinearity, we remove 2 PPI effect entries (τ_1 for the control well PPI, and τ_2 for the first test PPI) from the total of 121 τ values. Therefore, the $\boldsymbol{Z}$ matrix's dimensions are $3,072$ rows by 125 columns.

We use `R`, an open source statistical computing software, to create these design matrices and to generate the simulated data scores. Once we have our $\boldsymbol{Y}$ vector, $\boldsymbol{X}$ matrix, and $\boldsymbol{Z}$ matrix, we use them in (11) to obtain the fixed effect estimation $\hat{\boldsymbol{b}} = \begin{bmatrix} \hat{\mu} \\ \hat{\lambda} \end{bmatrix}$ and the random effect estimation $\hat{\boldsymbol{u}} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\tau}} \end{bmatrix}$. Also, we assume the variance components are the true values for the data in this simulation study. We then use $\hat{\boldsymbol{\tau}}$ to determine the significant PPIs.

6 Simulation Results

The PPI effect, which is the primary focus of this project, is used to determine whether a given interaction should be considered significant.

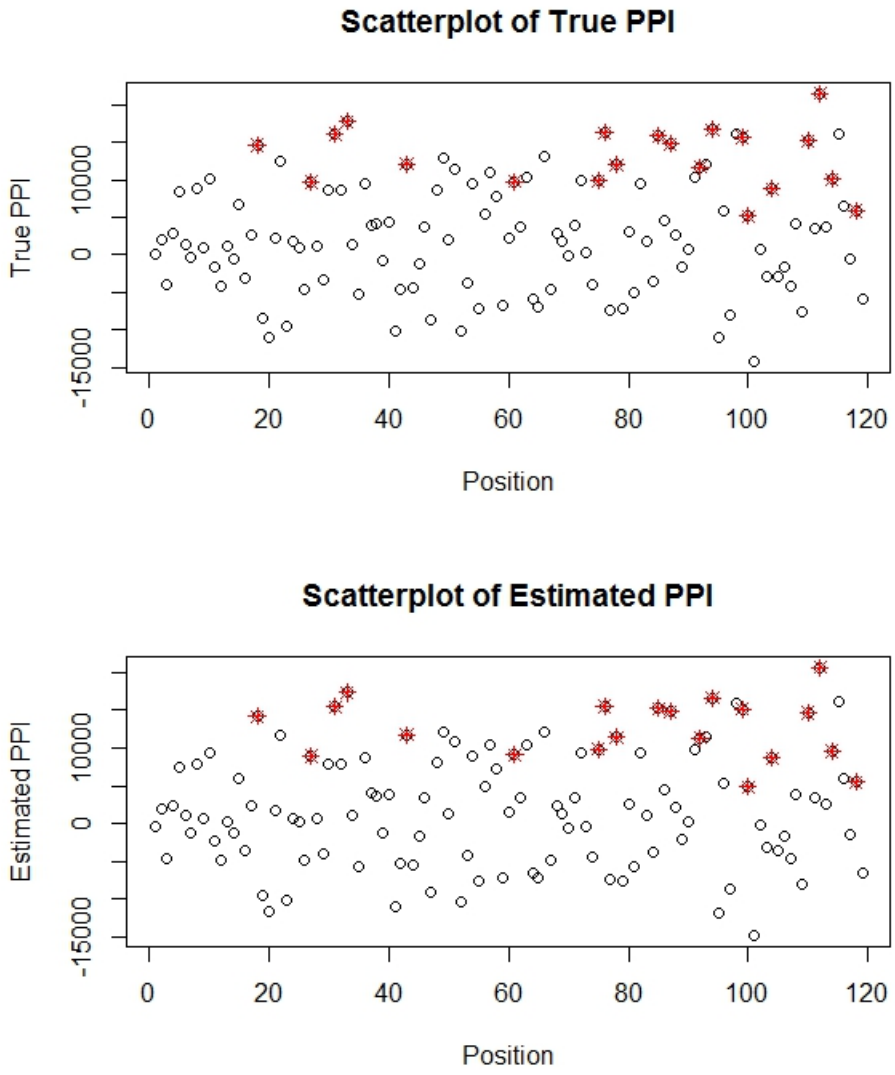

Figure 1: Plots of true PPI effects and estimated PPI effects

We assigned a position identifier to each of the distinct wells across the 8 plates; there are 119 position identifiers, corresponding to the 119 PPI effects. The above panel shows the true PPI effect versus position identifiers (Figure 1). The red stars indicate true significant

protein-protein interactions. The bottom panel shows the estimated PPI from the mixed model versus position identifiers. The red stars are the same protein pairs as the above panel. The majority of these points fall into the top half of the plot, where stronger PPIs would be expected.

The task, then, is establishing criteria to determine which of these PPIs are actually significant. Creating a threshold, for example, would create a cutoff for PPI values. Points above the line are estimated as significant PPIs, and those below are not.

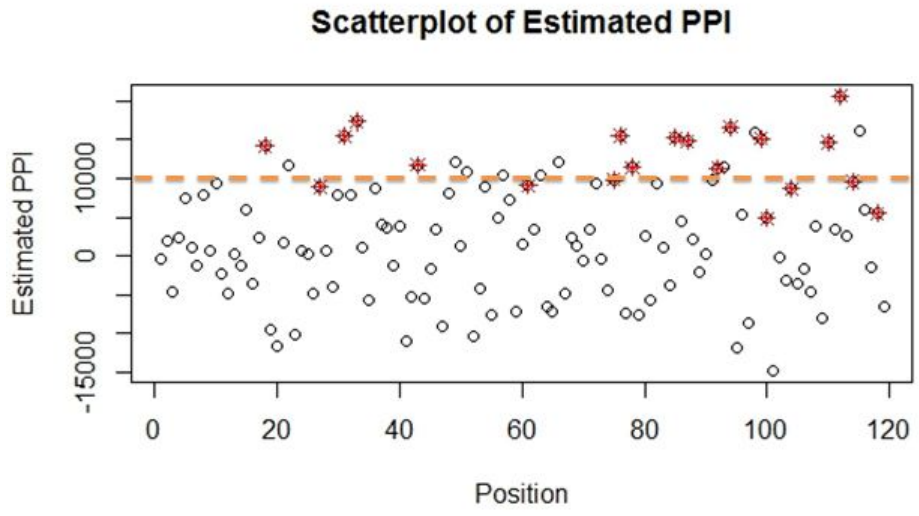

Figure 2: An example of a threshold for significant PPI effects

We can see from Figure 2 that some of the points above the threshold are not truly significant PPIs. These points are known as false positives, or type I error. Similarly, the significant PPIs that fall below the threshold are known as type II error. A contingency table is shown below to identify false positives and false negatives (Figure 3).

		Estimation	
Truth		Positive	Negative
	Positive	14	6
	Negative	9	90

Figure 3: Contingency table for PPI effects

Among all the protein pairs that are detected with significant PPI, 39.1% (9/(9+14)) are false positives; among all that are not detected, 6% (6/(6 + 90)) of the significant PPIs are false negatives. In creating the criteria for significant PPIs, minimization of type I and type II errors must be considered. Possible methods may include running numerous simulations to find trends in PPI effect or optimization of type I and type II error using thresholds.

7 Data Examination

As mentioned in the previous section, the variance components for the simulated data are known. Using prediction theory, the matrices $\mathbf{G}$ and $\mathbf{R}$ are constructed using (9) and (10) where $\sigma^2_\beta = 2.5 \times 10^7$, $\sigma^2_\tau = 3.6 \times 10^7$, and $\sigma^2 = 1 \times 10^6$. These matrices are used to solve the matrix equation (11) for $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{u}}$.

These two estimations are then used to determine the fit of the data to the proposed model. We use

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{b}} + \mathbf{Z}\hat{\boldsymbol{u}} \tag{20}$$

as an approximation of the actual observations, $\boldsymbol{y}$. Comparing the distributions of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$, the two follow similar patterns (Figure 4).

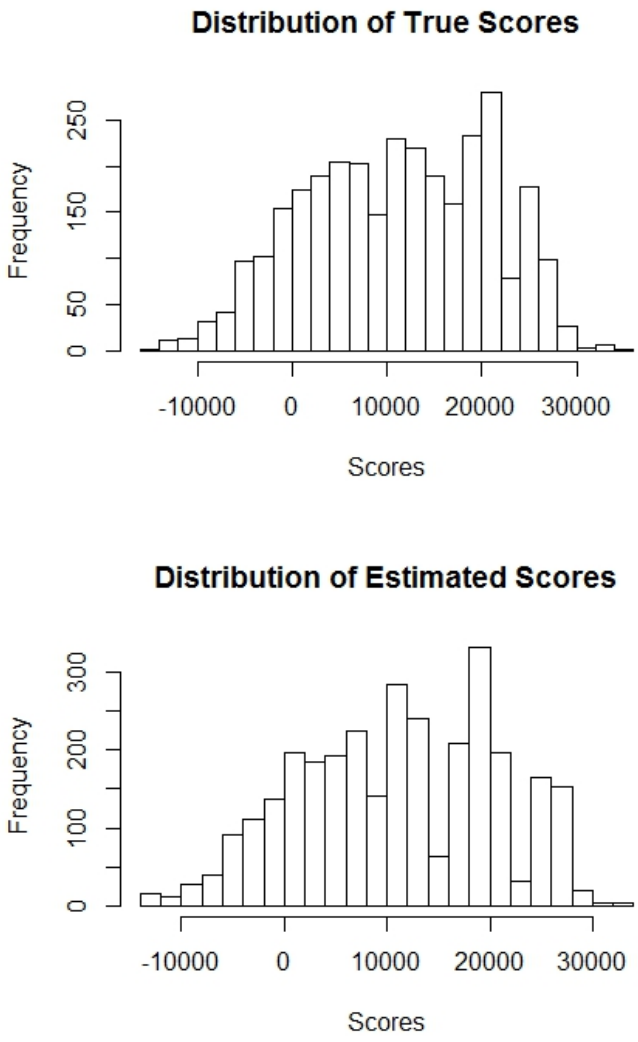

Figure 4: The distribution of $\boldsymbol{y}$, the true scores, and the distribution of the estimated scores $\hat{\boldsymbol{y}}$

Around the score value of $20,000$, we notice higher frequencies. These higher frequencies are due to the control group, which produces higher fluorescence scores. Observing the distribution with only the test group scores (Figure 5), the distribution shows characteristics of normal distributions.

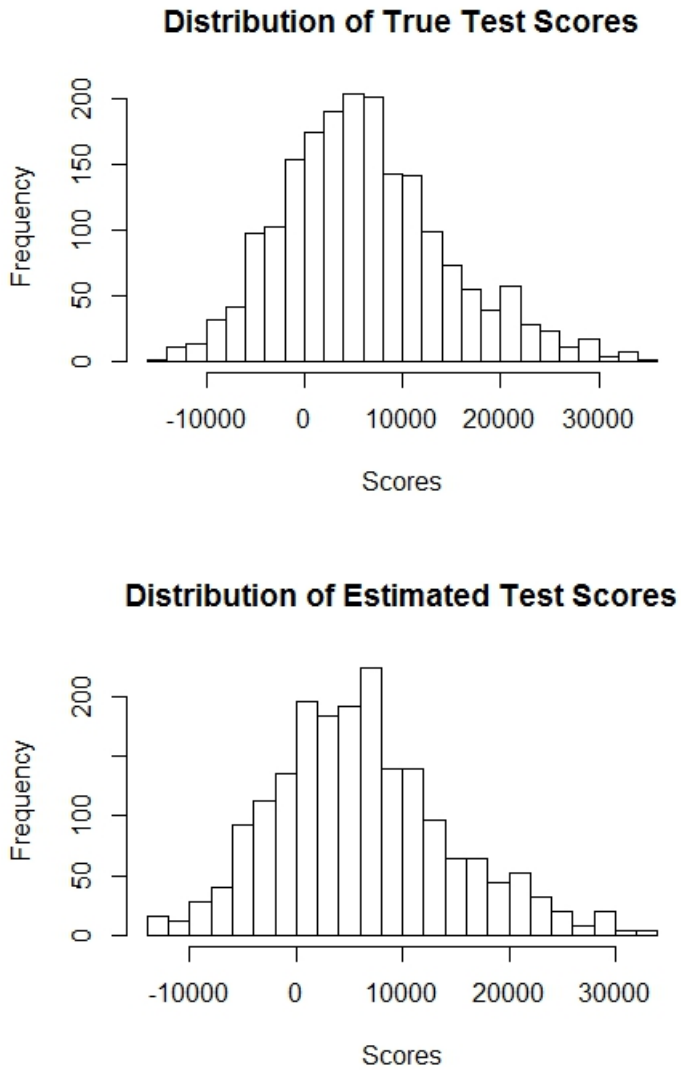

Figure 5: The distribution of only the test scores of $\boldsymbol{y}$, and the distribution of test scores of $\hat{\boldsymbol{y}}$

To analyze the accuracy of the model, we calculate the error, $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$, which is normally distributed (Figure 6), with mean 1.685×10^{-10} and variance 9.8×10^5. Recall that the model assumes $\sigma^2 = 1 \times 10^6$ and $E(\boldsymbol{e}) = 0$. Our estimated values are very close to the true values;

therefore, the model is a good fit.

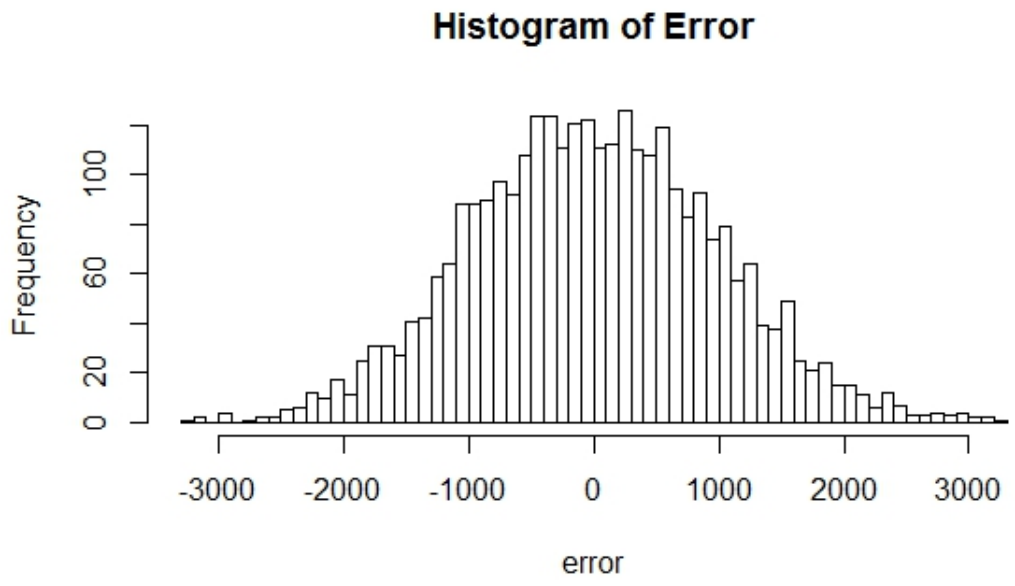

Figure 6: Distribution of error e

Additionally, when $\hat{\boldsymbol{y}}$ is plotted against $\boldsymbol{y}$, the resultant graph indicates a strong linear correlation between the true and estimated scores (Figure 7).

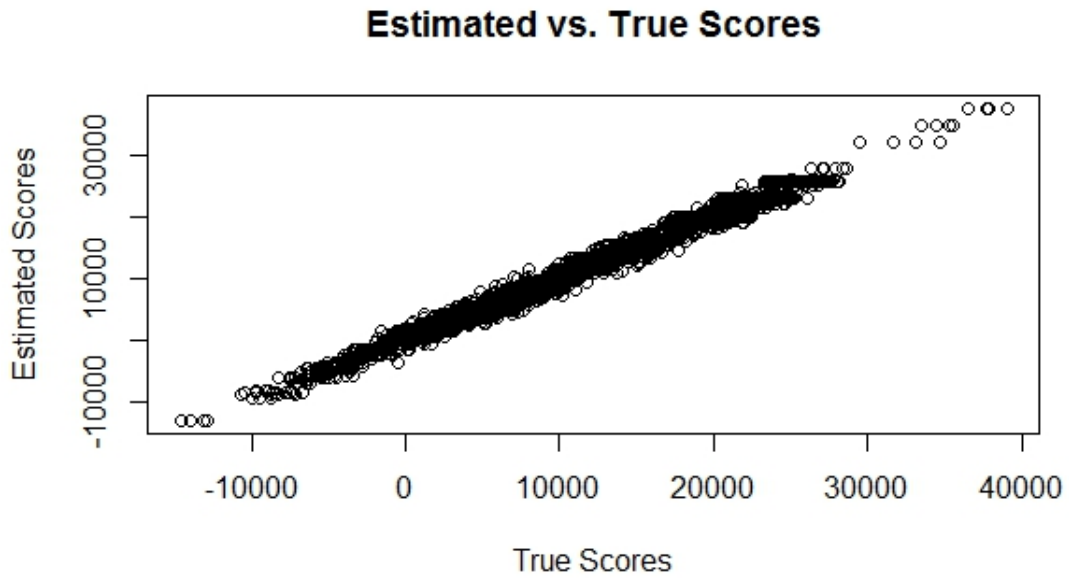

Figure 7: Plot of estimated versus true scores, correlation coefficient $R = 0.9947034$

The PPI effect $\boldsymbol{\tau}$ can then be examined (Figure 8). Note that the true and estimated PPI distributions have similar characteristics, with most PPI effects in the $-10,000$ to $20,000$ range. These results serve as evidence in favor of the proposed model.

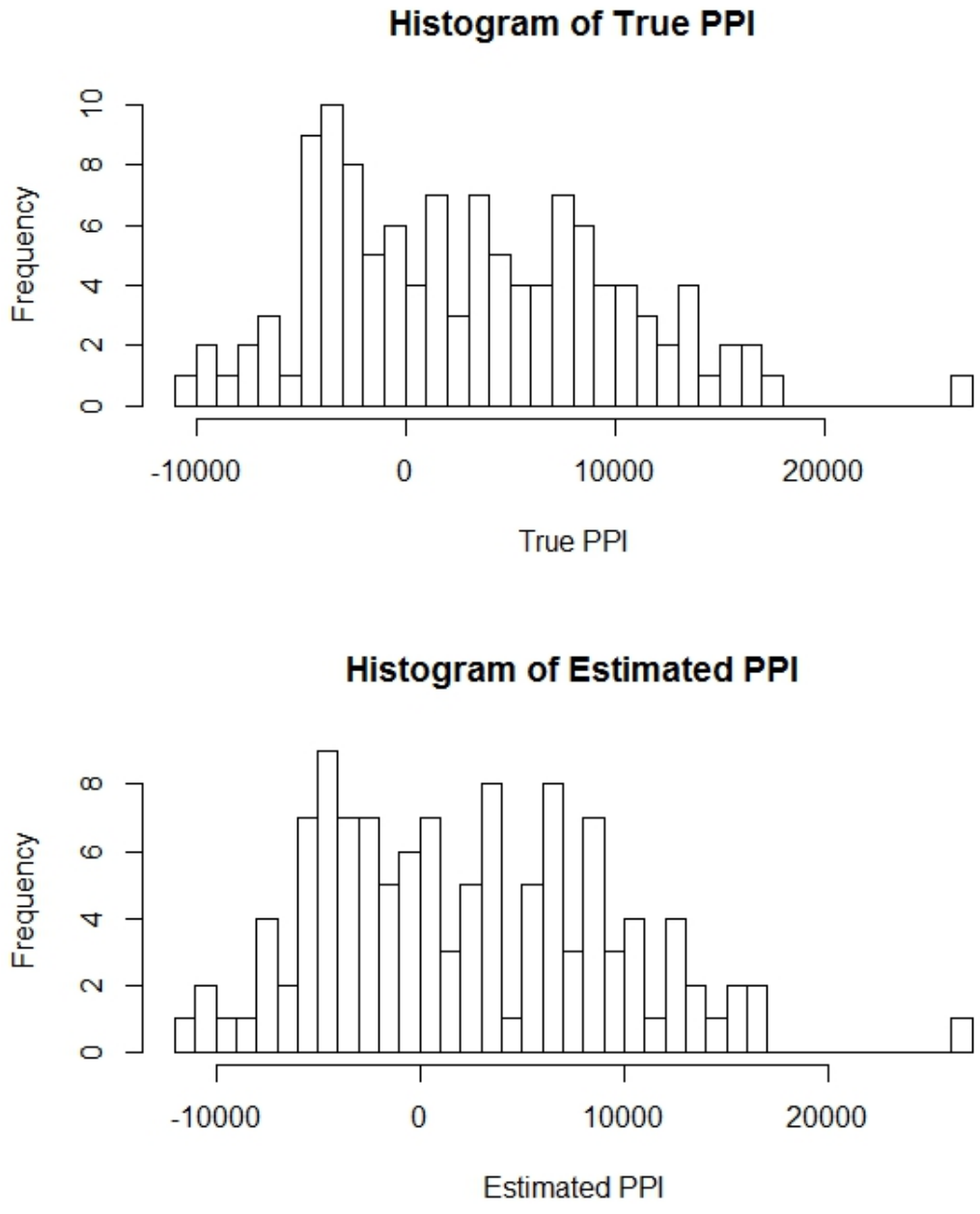

Figure 8: Distributions of true PPI effect and estimated PPI effect

8 Conclusion

We propose a statistical mixed model to detect PPIs among $1,417,701$ protein pairs for *A. thaliana* plants. In this model, we consider several sources of variations such as: positive control versus tests, variations from plates and PPIs. In claiming that there is no PPI effect, we expect that the PPI factor has no effect on the fluorescence score. On the other hand, if there exists a PPI effect, then it produces a higher fluorescence score. By applying our proposed model to a simulated data set, it has been demonstrated that our model works well in ideal situations. We also provide some illustration of type I error and type II error analyses. The success of the simulation study implies that our mixed model may fit the real PPI data, assuming a normal distribution in the real data set.

Future work includes applying the REML estimation on both the simulated data and the real data. The real data may not exhibit linearity and normality. Consequently, a link function is considerably useful. The purpose of a link function is to create a generalized linear model (GLM), so we can apply our mixed model to real data sets even if the normality assumption does not hold.

9 Acknowledgements

We would like to thank Dr. Heng Wang, Department of Statistics and Probability, Lyman Briggs College, Michigan State University, for sharing this project and allowing us to work on it with her. Also, we would like to thank Mengtian Shen, Department of Statistics and Probability, Michigan State University, for her contributions as a mentor. We would like to acknowledge Dr. Jin Chen, MSU-DOE Plant Research Laboratory, Computer Science and

Engineering Department, Michigan State University, for his collaboration in this project in providing us with the real data set. This work is supported by grants from the National Security Agency (under Grant Number H98230-13-1-0259) and the National Science Foundation (under Grant Number DMS 1062817). Finally, we are grateful to Lyman Briggs College and Michigan State University for their support of this research.

10 References

[1] Chen, J., Lalonde, S., Obrdlik, P., et al. (2012). Uncovering Arabidopsis Membrane Protein Interactome Enriched in Transporters Using Mating-based Split Ubiquitin Assays and Classification Models. *Frontiers in Plant Science, 3*, 124.

[2] Corbeil, R. R., & Searle, S. R. (1976). Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. *American Society for Quality, 18*, 1.

[3] [Image of gamma distributions]. Retrieved April 12, 2006, from www.clayford.net/statistics/wp-content/uploads/2011/08/gamma2.png.

[4] Lalonde, S., Ehrhardt, D., Loqué, D., et al. (2008). Molecular and Cellular Approaches for the Detection of Protein-protein Interactions: Latest Techniques and Current Limitations. *The Plant Journal, 53*, 610-635.

[5] University of Guelph (n.d.). Prediction Theory. Retrieved July 24, 2013, from www.aps.uoguelph.ca/∼lrs/ABModels/NOTES/predict.pdf.