Josh Joseph

Mike Ferguson

EC500 D1 - Homework 2

Objective

In this assignment, we were tasked with maximizing the yield of a given biosynthetic pathway. This involved a set of target genes (factors), which could be assigned a variety of promoter/terminator combinations (modules). To mirror real-world constraints, we had a limited budget with which to simulate experiments, and needed to employ an iterative process of designing experiments, abstracting a yield model from the experiments, and using the model's predictions in order to guide further experiments. While our initial objective was to find a set of modules which could produce a target yield of greater than 500 mg/L, we sought to continue the experiment until we had a more robust model that could reliably predict good yields with a variety of modules.

Methods

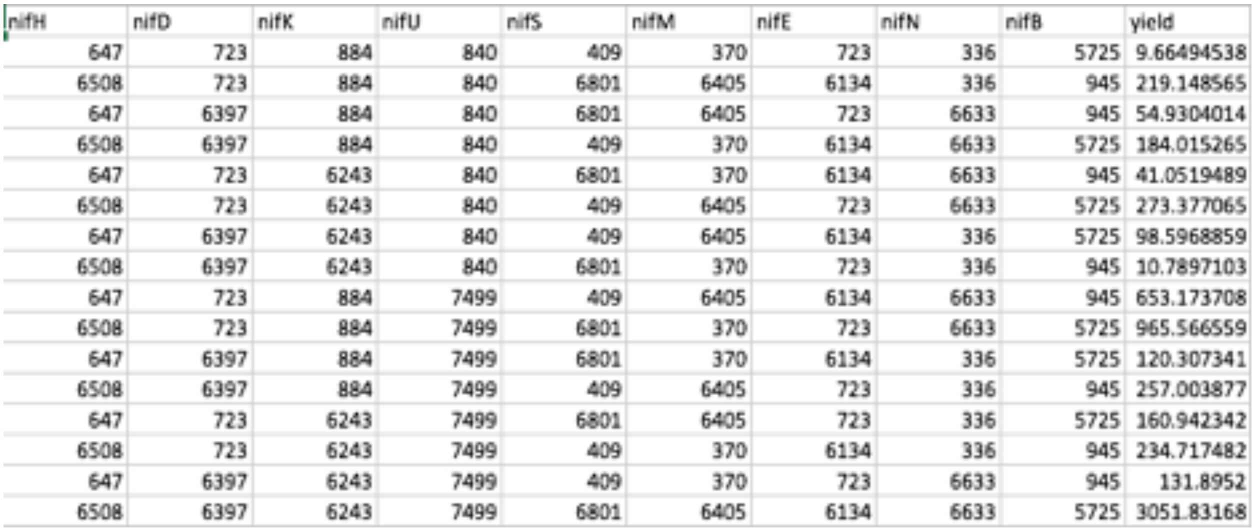
To simulate our experimental model and yields, we used the Double Dutch Pathway Designer (www.doubledutchcad.org). Our approach was to start with a 2k-p fractional factorial resolution IV model in order to screen for the main effects of individual genes, and after an initial set of regression analyses, to move to a resolution V model to assess for the effects of factor interactions. In resolution IV model fractional factorial designs, main effects are not confounded with other main effects or with two-factor interactions, but two-factor interactions are confounded with each other. Comparatively, in resolution V models, both main effects and two-factor interactions are not confounded with one another, but two-factor interactions can be confounded with three-factor interactions [Box 1961]. The comparative advantage of the resolution IV model as an initial model is that it is economical, and in the absence of yield data, we wished to spend as little of our available credit as possible before being able to narrow the design space.

Once our initial yield data was available, we used the R statistical package (www.r-project.org) to perform a multiple linear regression on it to model the relationship between the expression strength of the factors and the overall pathway yield. The regression was also performed on a log transformation of the data, and the significance of the correlation of individual terms was used to identify factors that potentially were not contributors to the pathway. These terms were then run in a resolution V model to evaluate for interaction effects, and the results were then re-evaluated using another round of multiple linear regression, which was itself evaluated using a factorial assignment.

Our next step hinged on the outcome of the regression model. If the model did not have a strong correlation beyond the original results of the resolution IV model, we would submit a factorial design, which would incur a significant cost, but with a smaller number of factors, might be within our budget. Alternatively, if the model proved parsimonious at this stage, we planned to use an automated, factorial evaluation of our model using a custom Python script in order to test a set of candidate module assignments, aiming to ensure that there were no overlapping module components (promoters or terminators) between factors.

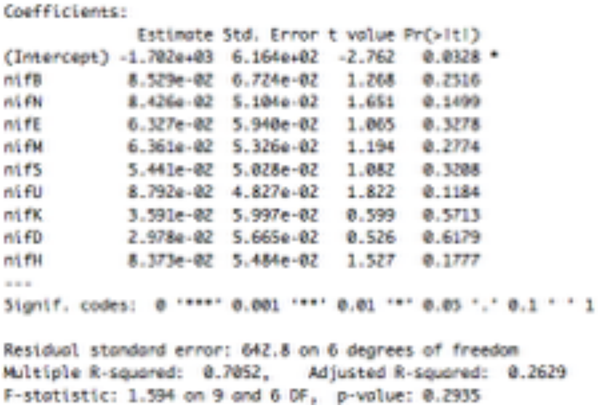
Results

Our initial, blind resolution IV model had a wide range of yield data.



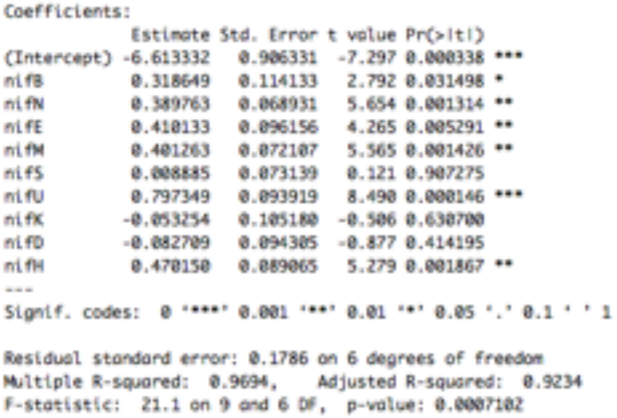
nifH	nifD	nifK	nifU	nifS	nifM	nifE	nifN	nifB	yield
647	723	884	840	409	370	723	336	5725	9.66494538
6508	723	884	840	6801	6405	6134	336	945	219.148565
647	6397	884	840	6801	6405	723	6633	945	54.9304014
6508	6397	884	840	409	370	6134	6633	5725	184.015265
647	723	6243	840	6801	370	6134	6633	945	41.0519489
6508	723	6243	840	409	6405	723	6633	5725	273.377065
647	6397	6243	840	409	6405	6134	336	5725	98.5968859
6508	6397	6243	840	6801	370	723	336	945	10.7897103
647	723	884	7499	409	6405	6134	6633	945	653.173708
6508	723	884	7499	6801	370	723	6633	5725	965.566559
647	6397	884	7499	6801	370	6134	336	5725	120.307341
6508	6397	884	7499	409	6405	723	336	945	257.003877
647	723	6243	7499	6801	6405	723	336	5725	160.942342
6508	723	6243	7499	409	370	6134	336	945	234.717482
647	6397	6243	7499	409	370	723	6633	945	131.8952
6508	6397	6243	7499	6801	6405	6134	6633	5725	3051.83168

While it was tempting to immediately trial permutations of our outlier pathway, it unfortunately involved multiple homologous components.



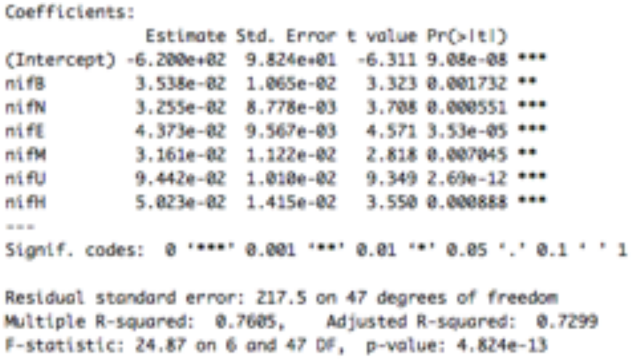
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.702e+03  6.164e+02  -2.762   0.0328 *
nifB         8.529e-02  6.724e-02   1.268   0.2516
nifN         8.426e-02  5.104e-02   1.651   0.1499
nifE         6.327e-02  5.940e-02   1.065   0.3278
nifM         6.361e-02  5.326e-02   1.194   0.2774
nifS         5.441e-02  5.028e-02   1.082   0.3208
nifU         8.792e-02  4.827e-02   1.822   0.1184
nifK         3.591e-02  5.997e-02   0.599   0.5713
nifD         2.978e-02  5.665e-02   0.526   0.6179
nifH         8.373e-02  5.484e-02   1.527   0.1777
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 642.8 on 6 degrees of freedom
Multiple R-squared:  0.7052,	Adjusted R-squared:  0.2629
F-statistic: 1.594 on 9 and 6 DF,  p-value: 0.2935
```

Our initial regression model for the data did not have a particularly high correlation coefficient; however, log transformation of the data yielded a much more robust model.



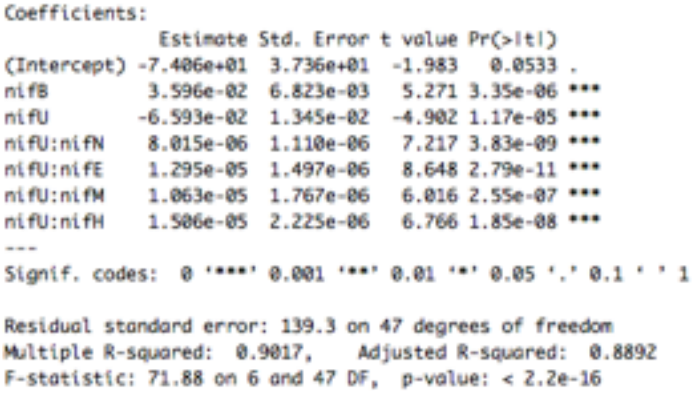
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.613332   0.906331  -7.297 0.000338 ***
nifB         0.318649   0.114133   2.792 0.031498 *
nifN         0.389763   0.068931   5.654 0.001314 **
nifE         0.410133   0.096156   4.265 0.005291 **
nifM         0.401263   0.072107   5.565 0.001426 **
nifS         0.008885   0.073139   0.121 0.907275
nifU         0.797349   0.093919   8.490 0.000146 ***
nifK        -0.053254   0.105180  -0.506 0.630700
nifD        -0.082709   0.094305  -0.877 0.414195
nifH         0.470150   0.089065   5.279 0.001867 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1786 on 6 degrees of freedom
Multiple R-squared:  0.9694,	Adjusted R-squared:  0.9234
F-statistic:  21.1 on 9 and 6 DF,  p-value: 0.0007102
```

We then submitted the resolution V design, dropping the factors which were not significant on the log model. While this raised the possibility of a costly set of no-yield set of experiments if we were wrong, the wide range of results led to a much better regression model with a relatively compatible adjusted R-squared.



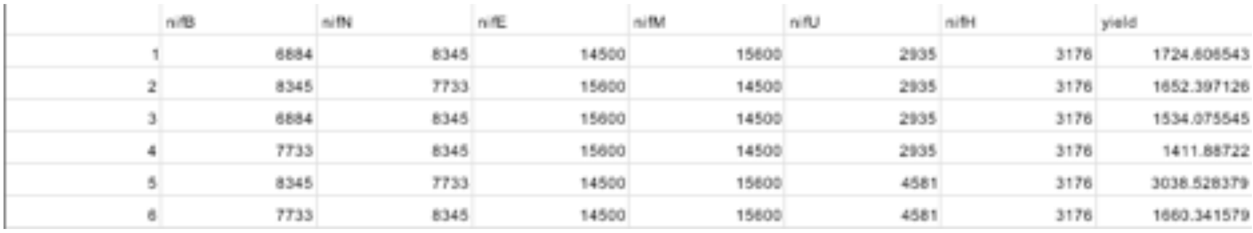
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.200e+02  9.824e+01  -6.311 9.08e-08 ***
nifB         3.538e-02  1.065e-02   3.323 0.001732 ** 
nifN         3.255e-02  8.778e-03   3.708 0.000551 ***
nifE         4.373e-02  9.567e-03   4.571 3.53e-05 ***
nifM         3.161e-02  1.122e-02   2.818 0.007045 ** 
nifU         9.442e-02  1.010e-02   9.349 2.69e-12 ***
nifH         5.023e-02  1.415e-02   3.550 0.000888 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 217.5 on 47 degrees of freedom
Multiple R-squared:  0.7605,	Adjusted R-squared:  0.7299 
F-statistic: 24.87 on 6 and 47 DF,  p-value: 4.824e-13
```

Using the leaps package (cran.r-project.org/web/packages/leaps/index.html), which evaluates coefficients from sets of multiple linear regressions by performing an exhaustive search for optimal subsets of potential regressors, we were able to create a regression model that could adequately incorporate relevant second-level effects.



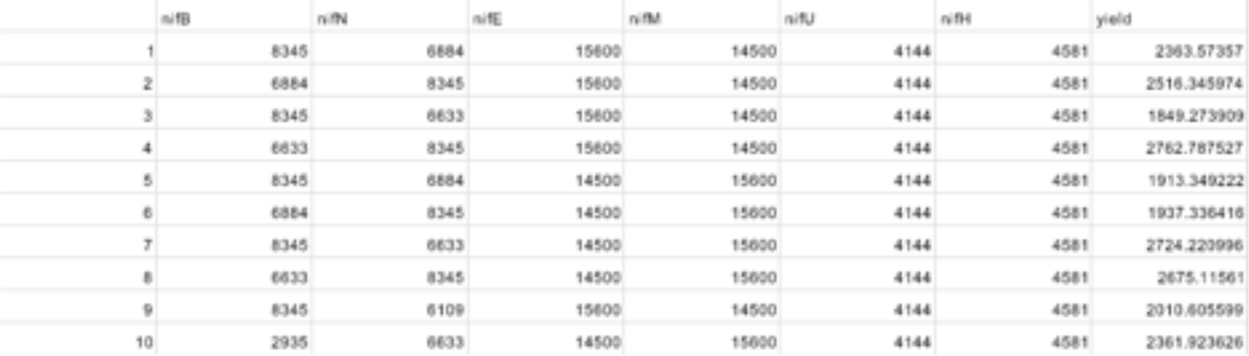
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.406e+01  3.736e+01  -1.983   0.0533 .  
nifB         3.596e-02  6.823e-03   5.271 3.35e-06 ***
nifU        -6.593e-02  1.345e-02  -4.902 1.17e-05 ***
nifU:nifN    8.015e-06  1.110e-06   7.217 3.83e-09 ***
nifU:nifE    1.295e-05  1.497e-06   8.648 2.79e-11 ***
nifU:nifM    1.063e-05  1.767e-06   6.016 2.55e-07 ***
nifU:nifH    1.506e-05  2.225e-06   6.766 1.85e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 139.3 on 47 degrees of freedom
Multiple R-squared:  0.9017,	Adjusted R-squared:  0.8892 
F-statistic: 71.88 on 6 and 47 DF,  p-value: < 2.2e-16
```

The next step of the process was to evaluate the model prospectively and ideally, to find pathway variants that had superior yields. There were several potential strategies available to us at this stage. The most direct of which was to calculate the gradient of our response equation and find a set of global maxima, then to find the combination of factor modules that was at closest distance to this point. However, while this strategy could find a theoretical maximum for the model, we could not find a combination of modules without a significant amount of overlap that could approach this point. In the interest of obtaining additional practical results, we then composed a custom Python script (https://github.com/jwjoseph/Assignment2/recombinator2.py) which used the response function to iterate through a set of factorial combinations of non-overlapping modules. The algorithm has a high degree of computational complexity (O(n!)), so we employed it on a relatively small portion of the sample space to keep runs within reasonable limits. The advantage of testing a significant number of designs *in silico* is that the amount of actual runs needed to validate the model were very small. Accordingly, the results were fairly close to the yields predicted by our model.



	nifB	nifN	nifE	nifM	nifU	nifH	yield
1	6884	8345	14500	15600	2935	3176	1724.606543
2	8345	7733	15600	14500	2935	3176	1652.397126
3	6884	8345	15600	14500	2935	3176	1534.075545
4	7733	8345	15600	14500	2935	3176	1411.88722
5	8345	7733	14500	15600	4581	3176	3038.528379
6	7733	8345	14500	15600	4581	3176	1660.341579

Only a slight expansion of our search space yielded consistently higher results.



	nifB	nifN	nifE	nifM	nifU	nifH	yield
1	8345	6884	15600	14500	4144	4581	2363.57357
2	6884	8345	15600	14500	4144	4581	2516.345974
3	8345	6633	15600	14500	4144	4581	1849.273909
4	6633	8345	15600	14500	4144	4581	2762.787527
5	8345	6884	14500	15600	4144	4581	1913.349222
6	6884	8345	14500	15600	4144	4581	1937.336416
7	8345	6633	14500	15600	4144	4581	2724.220996
8	6633	8345	14500	15600	4144	4581	2675.11561
9	8345	6109	15600	14500	4144	4581	2010.605599
10	2935	6633	14500	15600	4144	4581	2361.923626

Discussion

Using only a pair of fractional factorial designs, we were able to create a model that could consistently deliver high-yield pathways. This is in keeping with the progression from resolution IV to resolution V models as an economical means of designing experiments. However, our approach had several notable limitations. The most significant of which is that it required a relatively high degree of algorithmic complexity, and despite this, it is not guaranteed to obtain an optimal solution.

One of the reasons for this is that our algorithm examines only one module from the library per promoter and per terminator. While this prevents pathway homology and significantly limits the overall runtime, it also tends to emphasize designs that increase the difference between module levels.

A potential solution to this would be using a dynamic or memorized algorithm, which could dynamically expand the search space to modules which are close to locally optimal solutions. It is important to note that finding a global maximum using a gradient of the response function may yield a value that is only a theoretic, rather than actual maximum yield. Without the ability to tune the characteristics of factor modules, there are a limited, number of discrete combinations of modules available with which to optimize a pathway, and the closest fit among modules to a theoretic best solution of the module may not actually be the optimal combination. This is the relative advantage of our approach - it begins with the available factor modules, and builds a set of solutions quickly - which can be further refined simply by increasing the search depth of the algorithm.

References

Box GE, Hunter JS. The 2 k—p fractional factorial designs. Technometrics. 1961 Aug 1;3(3): 311-51.