This document builds on “Aim 2 predictors.docx”. Unlike in those documents, for this analysis all “fig|83333.1.rna.x” genes were excluded, leaving only “fig|83333.1.peg.x” genes. Also, some of the definitions of predictors have subtly changed, most notably SStat – we are now using LogSStat instead for numerical reasons.

# Updated list of predictors with definitions

* Mean of the gene’s eijs (**MeanEij**)
* SD of the gene’s eijs (**SDEij**)
* Multi BIC flag – no bonus or offset, just simply whether the 1-component model fit better than the 2-component one (**MultiBIC**)
* Uni BIC flag (same as the Multi one except univariate) (**UniBIC**)
* Mean of the gene’s aijs (**MeanAij**)
* Percent of the gene’s aijs which were < 0.5 (**%off**)
* Number of scenarios the gene is in (**NumSc**)
* Rank of the gene’s scenarios’ names lexicographically (genes without scenarios will be first in the ranking, then alphabetically by the gene’s first scenario, ties broken by gene’s second scenario, etc) (**ScName**)
* Number of genes in the gene’s operon (**NumOp**)
* The gene’s peg ID (**PegID**)
* Rank of the gene's common name alphabetically (**ComName**)
* Pre-existing weighted score (the components and weighting are documented in an email) (**WScore**)
* Log of the S statistic, where the S statistic is the probability the gene is always-inactive, assuming the aijs are calibrated and experiments are independent. I.e., take each of the gene’s 907 aijs (one for each experiment), subtract them from 1 to find the probability the gene was inactive in each experiment, and multiply them all together to find the probability the gene is inactive in all experiments. (**LogSStat**)
* Maximum correlation between the gene and any other gene (**MaxCor**)
* Peg ID of the gene the gene has maximum correlation with (**PegMaxCor**)
* Minimum correlation between the gene and any other gene (strongest anti-correlation) (**MinCor**)
* Peg ID of the gene the gene has minimum correlation with. (**PegMinCor**)
* Maximum correlation of the gene with any of its operon-mates. Genes with no operon-mates are assigned 1 as their maximum correlation with any operon-mate; this strongly influences the distribution of these correlations, though, and perhaps should be reconsidered? (**MaxCorOp**)

As before, ties for any of the “rank” predictors are broken at random, and the random tiebreak is recomputed every time the code is run.

# Individual GLM testing

Unlike the previous GLM testing, this analysis considers each predictor one-by-one – that is, each one is run in its own GLM with no covariates, and a p-value is obtained. Because the testing was conducted this way, WScore was left in the analysis for comparison, unlike previous GLMs. As before, family=binomial(logit) was used.

Even in this individual GLM testing, MultiBIC and UniBIC were still performing extremely poorly (p=0.958 and p=0.963 respectively), similar to what was observed in the previous GLM testing. I found this very strange, so I made contingency tables for how well MultiBIC and UniBIC line up with the correct answers.

|  |  |  |
| --- | --- | --- |
|  | Always-Inactive | Always-Active or Changing State |
| MultiBIC Flagged | 13.0% | 23.1% |
| MultiBIC Not Flagged | 0% | 63.9% |

|  |  |  |
| --- | --- | --- |
|  | Always-Inactive | Always-Active or Changing State |
| UniBIC Flagged | 13.0% | 34.4% |
| UniBIC  Not Flagged | 0% | 52.6% |

This shows that both MultiBIC and UniBIC correctly flagged all 13.0% of the genes which were actually always-inactive, although with significant numbers of false positives (MultiBIC doing slightly better). Examining the detailed logistic-regression results for MultiBIC, we have a model of -20.57 + 19.99\*MultiBIC, with SE of 383.01 on both coefficients. This model implies that the probability of being always-inactive, given not flagged, is , and the probability of being always-inactive, given flagged, is . Both of these are reasonable values; in fact, they’re almost identical to the actual probabilities computed from the contingency table. Only the SEs are unreasonably large.

The problem here, based on further research, is that having a 0 in any cell of the contingency table sometimes causes the logistic regression not to converge. One fix suggested online was to add 0.5 to all cell counts. This not being easy in R, I instead decided to add 1 to all cell counts.

Here’s the revised results, adding 1 to all cell counts for MultiBIC and UniBIC:

|  |  |
| --- | --- |
| **Predictor** | **Logistic Regression p-value** |
| WScore | 1e-102 |
| %off | 1e-68 |
| LogSStat | 1e-62 |
| SDEij | 7e-44 |
| MeanEij | 5e-37 |
| MeanAij | 8e-36 |
| MultiBIC | 1e-12 |
| UniBIC | 8e-11 |
| MaxCorOp | 7e-6 |
| PegMaxCor | 0.0001 |
| PegID | 0.0015 |
| MaxCor | 0.0140 |
| MinCor | 0.0322 |
| All others (5 predictors) | > 10% |

Next, I tested each of these single-predictor models vs. the null model (intercept only) using a Likelihood Ratio Test, to see if the results are any different. (Again, MultiBIC and UniBIC were adjusted by adding 1 to each cell count.) Listed are Chi-squared p-values.

|  |  |
| --- | --- |
| **Predictor** | **LRT Chi-square p-value vs. null model** |
| MultiBIC | 2e-219 |
| WScore | 1e-212 |
| LogSStat | 2e-191 |
| UniBIC | 3e-157 |
| %off | 3e-94 |
| SDEij | 2e-77 |
| MeanAij | 1e-41 |
| MeanEij | 2e-39 |
| MaxCorOp | 2e-5 |
| PegMaxCor | 0.0001 |
| PegID | 0.0014 |
| MinCor | 0.0028 |
| MaxCor | 0.0140 |
| All others (5 predictors) | > 10% |

This list (by LRT) is very similar to the previous list, but with a few differences. We see MultiBIC and UniBIC get a lot more credit for their stellar performance, with MultiBIC even beating WScore – which is surprising to me. (MultiBIC is one component of WScore.) LogSStat is also much closer to WScore in this list than it was in the previous list. MeanAij beats MeanEij, which should be expected, rather than the other way around as it was in the previous list. Overall, this list’s results make more intuitive sense to me – that is, they conform better to my expectations.

Next, we do the same test, but for all 17 choose 2 = 136 pairs of predictors (excluding WScore from this analysis). The fitted models just have two main-effects terms, not any interaction terms, and are compared against the null model (intercept only) using the LRT. (Re-running the tests including the interaction terms, e.g. using MultiBIC + LogSStat + MultiBIC:LogSStat instead of merely MultiBIC + LogSStat, produced almost exactly the same table.) Notably, in this analysis MultiBIC and UniBIC are **not** corrected by adding to cell counts, because that would require making up fictitious data for whatever other predictor was with them.

|  |  |  |
| --- | --- | --- |
| **Predictor Combination** | **LRT Chi-square p‑value vs. null model** | **“Fruitfulness”** |
| MultiBIC + LogSStat | 2e-279 | 60.1 |
| MultiBIC + %off | 1e-277 | 58.2 |
| MultiBIC + MeanAij | 2e-275 | 56.1 |
| MultiBIC + MeanEij | 4e-267 | 47.7 |
| UniBIC + LogSStat | 2e-251 | 60.2 |
| MeanAij + LogSStat | 2e-232 | 41.2 |
| MultiBIC + SDEij | 3e-225 | 5.8 |
| MultiBIC + UniBIC | 8e-225 | 5.4 |
| MultiBIC with any other predictor (10 combinations) | 1e-224 to 2e-220 | 5.3 to 1.1 |
| UniBIC + %off | 8e-222 | 64.5 |
| SDEij + LogSStat | 8e-207 | 15.5 |
| UniBIC + MeanAij | 2e-203 | 46.1 |
| LogSStat + MinCor | 1e-198 | 7.4 |
| LogSStat + %off | 2e-193 | 2.1 |
| LogSStat with any other predictor (10 combinations) | 3e-191 to 8e-190 | -0.5 to -1.6 |
| UniBIC + MeanEij | 2e-189 | 32.2 |
| UniBIC + SDEij | 1e-164 | 7.4 |
| UniBIC with any other predictor (10 combinations) | 5e-162 to 2e-158 | 4.7 to 1.1 |
| SDEij + %off | 7e-146 | 51.6 |
| MeanAij + %off | 5e-125 | 30.7 |
| SDEij + MeanAij | 2e-112 | 35.0 |
| %off + MinCor | 8e-99 | 4.6 |
| %off + MeanEij | 3e-97 | 3.0 |
| %off with any other predictor (9 combinations) | 4e-95 to 5e-93 | 0.8 to -1.4 |
| MeanEij + SDEij | 5e-87 | 9.6 |
| SDEij with any other predictor (10 combinations) | 8e-81 to 3e-76 | 3.4 to -1.3 |
| MeanEij + MeanAij | 7e-56 | 14.3 |
| MeanAij + MinCor | 4e-45 | 3.4 |
| MeanAij with any other predictor (9 combinations) | 2e-43 to 2e-40 | 1.9 to -1.2 |
| MeanEij + MinCor | 2e-41 | 1.9 |
| MeanEij with any other predictor (9 combinations) | 2e-40 to 3e-38 | 1.0 to -1.2 |
| Any other predictor combination (45 combinations) | 2e-8 to 0.3521 | 3.6 to -0.7 |

In the table, I made a distinction (based on my own classification) between predictors focused on identifying “not-changing-state” genes (these were colored red in the table), and predictors focused on identifying “usually inactive” genes (these were colored blue in the table).

Especially noteworthy are combinations which performed much better (had substantially lower p-values) than either of their individual components performed alone. I call these “fruitful” combinations, and introduce a metric for “fruitfulness”:

where is the stronger of the two LRT p-values of the individual components in the individual LRT analysis, and is the LRT p-value of the combination. “Fruitfulness” measures in some ways the contribution of the interaction to the overall p-value; it measures how much smaller the combination’s p-value is than its stronger component’s. Taking the ratio the other way might be more intuitive, but I chose this way so that higher “Fruitfulnesss” indicates the combination was more fruitful. Negative “Fruitfulness” indicates the combination did worse than one of its components alone did. “Fruitfulness” of each combination is shown in the rightmost column of the table.

I make the case that the most fruitful combinations include one of each kind of predictor (red and blue).

* First consider combinations involving MultiBIC. The four best combinations in the table, in terms of model fit, were pairings of MultiBIC, the best red predictor, with the four best blue predictors. Furthermore, these four combinations were the only significantly “fruitful” combinations involving MultiBIC – all other combinations involving MultiBIC produced about the same results as MultiBIC alone. In particular, the red predictors UniBIC and SDEij, although they performed very well on their own, barely added anything to the power of MultiBIC, whereas the blue predictors MeanAij and MeanEij, each much worse than UniBIC and SDEij on their own, significantly enhanced the predictive power of MultiBIC.
* UniBIC behaves similarly. Combinations with LogSStat, %off, MeanAij, and MeanEij are fruitful, whereas combinations with MultiBIC and SDEij are not, despite the individual power of those red predictors.
* Rounding out the strong red predictors, we see the same behavior in SDEij. Examining the table, we see that combinations with LogSStat, %off, MeanAij, and MeanEij are fruitful, and combinations with MultiBIC and UniBIC are not.
* Even red predictor MinCor, which is much weaker on its own than most other predictors we talk about a lot, follows this pattern. When paired with stronger red predictors – e.g. MultiBIC, UniBIC, SDEij – the combination is unfruitful; it doesn’t do much better than the stronger red predictor did on its own. However, when paired with blue predictors – e.g. LogSStat, %off, MeanAij, MeanEij – the combination is fruitful, and does (slightly but noticeably) outperform its individual components. Quantitatively, it has good fruitfulness given its individually low predictive power; perhaps we should introduce some kind of fruitfulness-to-power ratio.
* Conversely, blue predictors LogSStat and %off have fruitful combinations with the major red predictors, but not other blue predictors such as each other, or MeanEij.
* However, MeanAij does have fruitful combinations with blue predictors LogSStat, %off, and MeanEij, in addition to the red predictors mentioned above. In fact, this means MeanAij has fruitful combinations with every major predictor. One possible explanation for this is that, thanks to the sophisticated MultiMM-with-filtering process described in the Bayesian paper, MeanAij actually contains both blue and red information, meaning it has something to add to both red and blue predictors.