**Result explanation**

**Dataset**

Correlation:

Arabidopsis\_thaliana.txt.probes.intersection.mean.wona.pearson.allConditions

Glycine\_max.txt.probes.intersection.mean.wona.pearson.allConditions

MedicagoNormExp.txt.probes.intersection.mean.wona.pearson.allConditions

Orthology:

Athaliana\_167\_vs\_Gmax\_109.blp

Athaliana\_167\_vs\_Mtruncatula\_135.blp

Gmax\_109\_vs\_Athaliana\_167.blp

Gmax\_109\_vs\_Mtruncatula\_135.blp

Mtruncatula\_135\_vs\_Athaliana\_167.blp

Mtruncatula\_135\_vs\_Gmax\_109.blp

**Preprocess**

1. Partition Arabidopsis, Glycine Max, Medicago into 3 different dataset: top & bottom 5%, 10%, 20%; *I don’t understand what you mean by partition. The top 10% should include the top 5% for example. Partitions are disjoint.*
2. Find the blast scores in orthology files. Eg. Athaliana\_167\_vs\_Gmax\_109.blp and Gmax\_109\_vs\_Athaliana\_167.blp. If both have the same gene pair, save the gene pair and their mean identity.
3. Find the candidates. Eg. If gene pair (g1, g2) in Arabidopsis has orthology scores with (g1’, g2’) in Glycine Max, (g1, g1’) and (g2, g2’), (g1, g2, g1’, g2’) can be a candidate. *This completely confuses me. It should be that if g11, g12, g13 each has a good orthology score with g1’ and g21, g22, g23 each has a good orthology score with g2’ and {g11, g21} has a top correlation value as do {g12, g22} and {g13, g23} then all three influence the prediction on {g1’, g2’}.*
4. Generate the statistics.

**Features**

Instead of using:

Where

Now I used 13 features:

|  |  |  |
| --- | --- | --- |
|  | Mean correlation |  |
|  | Median correlation |  |
|  | Mean orthology |  |
|  | Median orthology |  |
|  | Mean p-value |  |
|  | Median p-value |  |
|  | Min correlation |  |
|  | Max correlation |  |
|  | Min orthology |  |
|  | Max orthology |  |
|  | Min p-value |  |
|  | Max p-value |  |
|  | Weighted mean = Sigma\_(correlation \* identity) |  |

**Calculation method**

True positive= prediction is positive and reality is positive

False positive= prediction is positive and reality is negative

True negative= prediction is negative and reality is negative

False negative = prediction is negative and reality is positive

## 5per

1. **Training using Arabidopsis and Glycine Max 5per, testing Glycine Max inbetween**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Positive | | | | Negative | | | | Harmonic Mean | | Avg deviation | |
|  | Recall | | Precision | | Recall | | Precision | |
|  | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML |
| All edges | 44315/45448 (%97.51) | 30967/45448 (%68.14) | 44315/76719 (%57.76) | 30967/38516 (%80.40) | 1673/34077 (%4.91) | 26528/34077 (%77.85) | 1673/2806 (%59.62) | 26528/41009 (%64.69) | (%54.70) | (%72.77) | 0.25 | 0.28 |

Predicted weights: -0.0250 0.1638 0.1389 0.1656 -0.3338 -0.0084 0.0124 -0.0794 0.1950 -0.3589 -0.2123 0.1343 0.0600

1. **Training using Arabidopsis and Glycine Max, testing Medicago**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Positive | | | | Negative | | | | Harmonic Mean | | Avg deviation | |
|  | Recall | | Precision | | Recall | | Precision | |
|  | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML |
| All edges | 97/188 (%51.60) | 97/188 (%51.60) | 97/207 (%46.86) | 97/208 (%46.63) | 87/197 (%44.16) | 86/197 (%43.65) | 87/178 (%48.88) | 86/177 (%48.59) | (%47.87) | (%47.62) | 0.45 | 0.47 |

Predicted weights: -0.0174 0.4425 0.2756 -0.0146 -0.2870 -0.0978 0.2466 0.2072 -0.0024 -0.0666 -0.1661 -0.1728 0.0033

1. **Training using 80% Arabidopsis and Glycine Max, testing 20% Arabidopsis and Glycine Max**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Positive | | | | Negative | | | | Harmonic Mean | | Avg deviation | |
| Recall | | Precision | | Recall | | Precision | |
| SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML |
| 6862/10117 (%67.83) | 6866/10117 (%67.87) | 6862/8468 (%81.03) | 6866/8462 (%81.14) | 5951/7557 (%78.75) | 5961/7557 (%78.88) | 5951/9206 (%64.64) | 5961/9212 (%64.71) | (%73.06) | (%73.15) | 0.21 | 0.274 |

Predicted weights: -0.0757 0.2404 -0.1005 0.0562 -0.0791 0.2546 0.3623 -0.0660 -0.1070 0.2580 -0.1735 0.0554 0.0016

## 10per

1. **Training using Arabidopsis and Medicago, testing Glycine Max**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Positive | | | | Negative | | | | Harmonic Mean | | Avg deviation | |
|  | Recall | | Precision | | Recall | | Precision | |
|  | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML |
| All edges | 47749/83681 (%57.06) | 53744/83681 (%64.22) | 47749/105157 (%45.41) | 53744/72012 (%74.63) | 10605/68013 (%15.59) | 49745/68013 (%73.14) | 10605/46537 (%22.79) | 49745/79682 (%62.43) | (%35.18) | (%68.61) | 0.51 | 0.26 |

Predicted weights: -0.0072 -0.0662 0.3457 0.2497 0.1666 0.1089 0.0548 -0.1558 -0.2413 -0.1470 -0.2584 0.0744 -0.0858

1. **Training using Arabidopsis and Glycine Max, testing Medicago**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Positive | | | | Negative | | | | Harmonic Mean | | Avg deviation | |
|  | Recall | | Precision | | Recall | | Precision | |
|  | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML |
| All edges | 179/342 (%52.34) | 179/342 (%52.34) | 179/374 (%47.86) | 179/375 (%47.73) | 187/382 (%48.95) | 186/382 (%48.69) | 187/350 (%53.43) | 186/349 (%53.30) | (%50.65) | (%50.51) | 0.79 | 0.40 |

Predicted weights: 0.2255 0.3801 0.1493 0.0073 0.2625 -0.2236 0.0498 0.3736 -0.0763 -0.0825 0.1352 -0.3328 0.0014

1. **Training using 80% Arabidopsis and Glycine Max, testing 20% Arabidopsis and Glycine Max**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Positive | | | | Negative | | | | Harmonic Mean | | Avg deviation | |
|  | Recall | | Precision | | Recall | | Precision | |
|  | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML |
| All edges | 12289/18569 (%66.18) | 12048/18569 (%64.88) | 12289/16672 (%73.71) | 12048/16100 (%74.83) | 10890/15273 (%71.30) | 11221/15273 (%73.47) | 10890/17170 (%63.42) | 11221/17742 (%63.25) | (%68.65) | (%69.11) | 0.46 | 0.25 |

Predicted weights: 0.1608 -0.1190 0.0765 -0.0367 -0.0535 0.0186 0.2107 0.3190 -0.1958 0.1358 0.1610 0.2205 0.0032

## 20per

1. **Training using Arabidopsis and Medicago, testing Glycine Max**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Positive | | | | Negative | | | | Harmonic Mean | | Avg deviation | |
|  | Recall | | Precision | | Recall | | Precision | |
|  | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML |
| All edges | 27671/152033 (%18.20) | 92711/152033 (%60.98) | 27671/34107 (%81.13) | 92711/133844 (%69.27) | 123000/129436 (%95.03) | 88303/129436 (%68.22) | 123000/247362 (%49.72) | 88303/147625 (%59.82) | (%60.70) | (%64.57) | 0.49 | 0.23 |

Predicted weights: 0.1027 0.1202 0.2315 -0.0038 -0.4201 0.0638 0.0209 0.2365 -0.0895 -0.3847 0.1568 0.0324 -0.0040

1. **Training using Arabidopsis and Glycine Max, testing Medicago**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Positive | | | | Negative | | | | Harmonic Mean | | Avg deviation | |
|  | Recall | | Precision | | Recall | | Precision | |
|  | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML |
| All edges | 372/700 (%53.14) | 368/700 (%52.57) | 372/758 (%49.08) | 368/750 (%49.07) | 363/749 (%48.46) | 367/749 (%49.00) | 363/691 (%52.53) | 367/699 (%52.50) | (%50.80) | (%50.79) | 0.68 | 0.35 |

Predicted weights: 0.4967 -0.0594 0.3149 -0.2000 -0.1071 0.0629 0.1398 0.2781 -0.0810 -0.0229 -0.0860 0.1399 -0.0021

1. **Training using 80% Arabidopsis and Glycine Max, testing 20% Arabidopsis and Glycine Max**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Positive | | | | Negative | | | | Harmonic Mean | | Avg deviation | |
|  | Recall | | Precision | | Recall | | Precision | |
|  | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML | SGD | NONML |
| All edges | 21523/32999 (%65.22) | 20668/32999 (%62.63) | 21523/31903 (%67.46) | 20668/29843 (%69.26) | 18665/29045 (%64.26) | 19870/29045 (%68.41) | 18665/30141 (%61.93) | 19870/32201 (%61.71) | (%64.72) | (%65.50) | 0.44 | 0.23 |

Predicted weights: 0.3245 -0.0288 0.0517 0.0221 0.0711 -0.1606 0.2081 -0.0985 0.0206 0.0016 0.1316 0.1745 -0.0015