Analysis of RSV2 data:

Code: SR\_230529\_treespace.R

The data for this analysis was obtained from BEAST tutorial on “Time stamped data”:  
<https://taming-the-beast.org/tutorials/MEP-tutorial/>

In this Beast run there were 2001 samples with interval of 500 samples between each. I took the first 200 which corresponds to the 1e+5:

A graph of a line with numbers and a red line

Description automatically generated with medium confidence

I cut-off the first 50 samples which corresponds to the 25000th sample (red line). The analysis was done on the rest 150 samples (each with interval of 500).

I used “phangorn” package to calculate the difference of tree metrics between the sampled trees. I calculated the correlation between tree metrics and difference of likelihood using cor() function in R. Table 1 shows the tree metrics used in this analysis and their computed correlations.

|  |  |
| --- | --- |
| Metric | Correlation |
| Kendal-Colijn | 0.20586854 |
| Subtree Prune Regraft | 0.39551556 |
| Robinson Foulds | 0.77220202 |
| Robinson Foulds Weighted | 0.754315 |
| Info Robinson Foulds | 0.79078397 |
| Jaccard Robinson Foulds | 0.62871472 |
| Kuhner Felsenstien | 0.761936 |
| Path Difference | 0.53508628 |
| Matching Split Distance | 0.55084975 |
| Nye Similarity | -0.6287147 |
| Tree Distance | 0.67894617 |

In table 1, Info Robinson Foulds has the highest correlation (0.79). Scatter plot below shows this metric and likelihood:

A graph with numbers and circles

Description automatically generated

I also did the an optimization to find a linear combination of metrics that minimize the difference to likelihoods, the value is the coefficient (or weight) of each metric given:

where shows the index of metric and is the likelihood difference. The optimization was done using optim() function with “Nelder-Mead” method in R.

The optimized coefficients ():

|  |  |
| --- | --- |
| Metric | Coefficients |
| Kendal-Colijn | 0.46678194 |
| Subtree Prune Regraft | -0.017475 |
| Robinson Foulds | 0.10353513 |
| Robinson Foulds Weighted | 0.32000334 |
| Info Robinson Foulds | 0.02977533 |
| Jaccard Robinson Foulds | -0.2194599 |
| Kuhner Felsenstien | -0.1437046 |
| Path Difference | 0.20545441 |
| Matching Split Distance | 0.35045625 |
| Nye Similarity | 0.30054578 |
| Tree Distance | -0.1372816 |

The calculated correlation for optimized case is: 0.7869199 (so, it’s not better than IRF metric). Scatter plot of optimized metric:

A graph of a tree-growing graph

Description automatically generated with medium confidence

In previous case, the optimization was forced to not have any intercept. I can add an intercept as a parameter to the objective function:

Where is the intercept. The coefficients after optimization are as follows:

|  |  |
| --- | --- |
| Metric | Coefficients |
| Kendal-Colijn | 0.34610905 |
| Subtree Prune Regraft | 0.49513603 |
| Robinson Foulds | 0.03413773 |
| Robinson Foulds Weighted | 0.36688447 |
| Info Robinson Foulds | 0.03188507 |
| Jaccard Robinson Foulds | 0.21683434 |
| Kuhner Felsenstien | 0.27033998 |
| Path Difference | 0.12228505 |
| Matching Split Distance | 0.37723783 |
| Nye Similarity | 0.36076627 |
| Tree Distance | -1.0238428 |
| intercept | 0.01470947 |

And the correlation with intercept is: 0.7877834

In addition, I did a regularization technic (also known as LASSO regression). This method is a modified version of the optimization method:

Using In this case, the coefficients are as follows:

|  |  |
| --- | --- |
| Metric | Coefficients |
| Kendal-Colijn | 0.46678194 |
| Subtree Prune Regraft | -0.017475 |
| Robinson Foulds | 0.10353513 |
| Robinson Foulds Weighted | 0.32000334 |
| Info Robinson Foulds | 0.02977533 |
| Jaccard Robinson Foulds | -0.2194599 |
| Kuhner Felsenstien | -0.1437046 |
| Path Difference | 0.20545441 |
| Matching Split Distance | 0.35045625 |
| Nye Similarity | 0.30054578 |
| Tree Distance | -0.1372816 |

The correlation in this case is: 0.7869199 which is same as the result of optimization.

Finally, I used linear regression model using lm() function in R. Coefficients of metrics in this case are as follows:

|  |  |
| --- | --- |
| Metric | coefficients |
| intercept | -154.91216 |
| Kendal-Colijn | -0.2544046 |
| Subtree Prune Regraft | -5.4794072 |
| Robinson Foulds | 7.63520371 |
| Robinson Foulds Weighted | -0.1105525 |
| Info Robinson Foulds | -0.0066915 |
| Jaccard Robinson Foulds | -8.3090608 |
| Kuhner Felsenstien | 7.60900224 |
| Path Difference | 0.1117178 |
| Matching Split Distance | 0.14330719 |

Calculated correlation in this case is: 0.791446, so it is slightly better than Info Robinson Foulds.

Scatter plot of optimized metric using linear regression method:

A graph of growth of trees

Description automatically generated