Discerning Systematic Bias in *S. Cervisiae* Pathways Using Novel Bayesian Statistics Problem Structuring Methods

Jacob M. Luber[1,2], Albert X. Jiang[1], Matthew A. Hibbs[1,2]
[1]Department of Computer Science, Trinity University, San Antonio, Texas 78212 USA
[2]The Jackson Laboratory, Bar Harbor, Maine 04609 USA

BACKGROUND

One of the major problems in computational biology is developing algorithms that cope with multi-functionality of proteins due to the reuse of biological pathways and components in different contexts. Because of this inherent complexity, new work coping with genome scale problems is potentially pressured to conform to existing biases or limitations present in the literature.

DESCRIPTION OF ALGORITHM

We propose an algorithm that utilizes a novel form of Bayesian problem structuring to correlate overfitting with these aforementioned problems in an ensemble of classifiers attempting to predict gene function based on incomplete or inherently biased expert knowledge. We tested this novel form of Bayesian problem structuring through iteratively removing random genes from the well-studied MAPK pheromone response pathway in *S. cerevisiae*. We then subsequently used this modified pathway as a training set in our classifier ensemble that is applied to features from freely available heterogeneous data sources. We then performed cross validation, creating features from the normalized results and the amount of overfitting observed in the first round of classification. Classification outputs from this initial ensemble are converted into features and fed into a second "Black Box" classifier that outputs the probability of expert knowledge being inherently biased.

CONCLUSION

Early results indicate that this problem structuring method has potential, based on a small subset of very well studied *S. cerevisiae* pathways, to predict pathways that may significantly suffer from study biases or lack of complete knowledge.

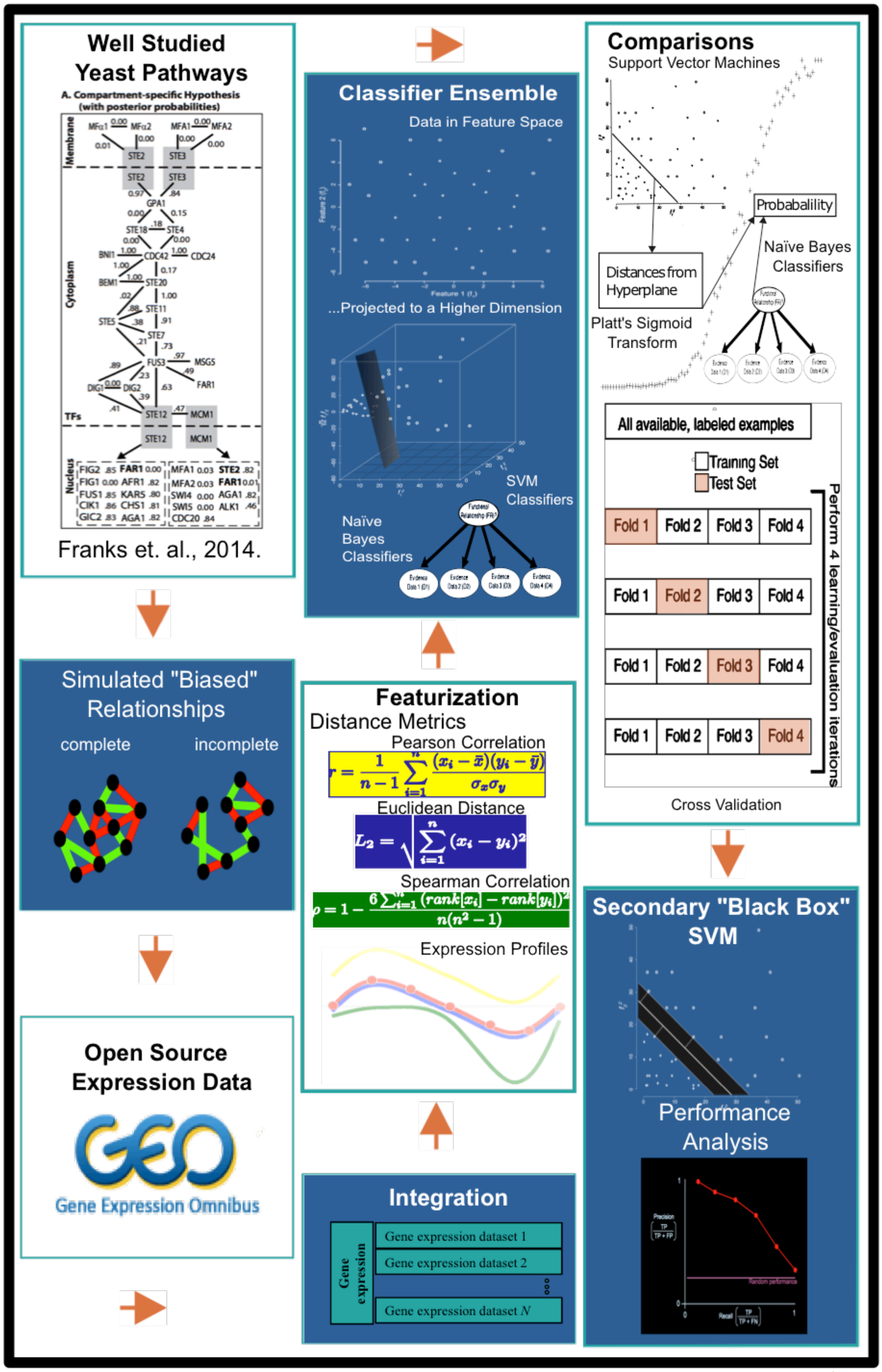
REFERENCES

Franks et. al. Estimating cellular pathways from an ensemble of heterogeneous data sources. biorxiv. 2014

Jordan, Michael I. What Are The Open Problems In Bayesian Statistics. ISBA Bulletin Vol 18, Number 1. 2011

Mavrovouniotis ML. Describing Multiple Levels of Abstraction in the Metabolism.ISMB. 1994.

Platt, John C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Microsoft Research. 1999.



This work was supported by NSF Division of Biological Infrastructure Grant DBI-1262049, an ISMB/ECCB Travel Fellowship, and a Trinity University Mach Undergraduate Research Fellowship.

cs.trinity.edu/~jluber

www.jax.org