

# The Limit of Genomic Data Integration for Predicting Protein-Protein Interactions: An Assessment using Naïve Bayes Classifier

L. Jason Lu<sup>1</sup>, Mark Gerstein<sup>1</sup>

**Keywords:** protein-protein interaction, genomic features, integration, prediction, Naïve Bayes, Boosting

## 1 Introduction.

Genomic data integration – the process of statistically combining diverse sources of information from functional genomics experiments to make large-scale predictions – is becoming increasingly prevalent. One might expect that this process should become increasingly powerful with the integration of more evidence. Here, we explore the limits of genomic data integration, assessing the degree to which predictive power increases with the addition of more features. We focus on a predictive context that has been extensively investigated and benchmarked in the past – the prediction of protein-protein interactions in yeast. We start by using a simple Naïve Bayes classifier for integrating diverse sources of genomic evidence, ranging from co-expression relationships to similar phylogenetic profiles. We expand the number of features considered for prediction to 16, significantly more than previous studies. Overall, we observe a small but measurable improvement in prediction performance over previous benchmarks based on four strong features. This allows us to identify new yeast interactions with high confidence, which we make available from [networks.gersteinlab.org/intint/](http://networks.gersteinlab.org/intint/). It also allows us to quantitatively assess the inter-relations amongst different genomic features. It is known that subtle correlations and dependencies between features can potentially confound the strength of interaction predictions. We, thus, investigate this issue in detail through calculating mutual information. To our surprise, we find no appreciable statistical dependence between the many possible pairs of features. We further explore feature dependencies by comparing the performance of our simple Naïve Bayes classifier with a boosted version of the same classifier, which is fairly resistant to feature dependence. We find that boosting does not improve the performance, indicating that, at least for prediction purposes, our genomic

---

<sup>1</sup> Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Ave, New Haven, CT 06520, E-mail: [Jason.Lu@yale.edu](mailto:Jason.Lu@yale.edu)

features are essentially independent. We conclude that by integrating a few (i.e., four) good features, we approach the maximal predictive power of current genomic data integration; moreover, this limitation does not reflect (potentially removable) inter-relationships between the features.

## 2 Software and files.

All data are available at:  
<http://networks.gersteinlab.org/intint/index.html>

## 3 Figures and tables.

<b>Four Features used in [1, 2]</b>	<b>Functional Genomic Features</b>		<b>Sequence/Structure Features</b>
MRNA co-expression	Absolute mRNA Expression Level	Gene Neighborhood	Threading Scores
MIPS Functional Similarity	Marginal Essentiality	Rosetta Stones	Co-evolution Scores
GO Functional Similarity	Absolute Protein Abundance	Synthetic Lethality	
Co-essentiality	Co-regulation	Gene Clusters	<b>Other Features</b>
	Phylogenetic Profiles		Interologs

Table 1: Genomic Features Considered in this Study.

## 4 References and bibliography.

### References

- [1] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., et al. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, **302**, 449-53.
- [2] Lu, LJ, Xia, Y, Yu, H, Rives, A, Lu, H, Schubert, F, Gerstein, M, 2005, Protein interaction prediction by integrating genomic features and protein interaction network analysis. *Data Analysis and Visualization in Genomics and Proteomics*. in press. John Wiley & Sons, Ltd, New York.