Implementation of the Graphlet Screening Method in Genomic Analysis Using Hail

By: Rishit Avadhuta
Advisors: Zheyang Wu, Ph.D, Kevin Crowthers, Ph.D

BACKGROUND

- **Gene expression** controls every biological function
- **Genome-Wide Association Studies** (GWAS) helps treat cancer & complex diseases (Creighton C. J., 2023)

But it's also difficult because:

- Lots of different **SNPs** (variations of DNA, single nucleotide polymorphisms) connect to a trait (sparse) (Jin et al., 2014)
- Each variation in the DNA (SNPs) **contributes little** to the trait (weak) (Jin et al., 2014)

ENGINEERING NEED

Current methods (L0, LASSO) used in Genomic Analysis are **not accurate enough** and do not **penalize error** enough to accurately determine the associated gene expression for SNPs (Horowitz, 2015).

ENGINEERING GOAL

Optimize the implementation of a new algorithm (**Graphlet Screening**) in an efficient genomic parser (**Hail package**) that is better at working with genomic data than the current gold standard algorithms in the industry to detect genomic signals more accurately.

TAKEAWAYS

Graphlet Screening receives... ***And...***

-25%
Hamming Distance Error
When compared to L0-Regularization

-72%
Hamming Distance Error
When compared to the LASSO

P-value (ANOVA): 1.7e-148

Current Applications of Hail:

By the **Broad Institute**
+ 100s of other genetic studies that **save lives**
& More! ^

METHODOLOGY

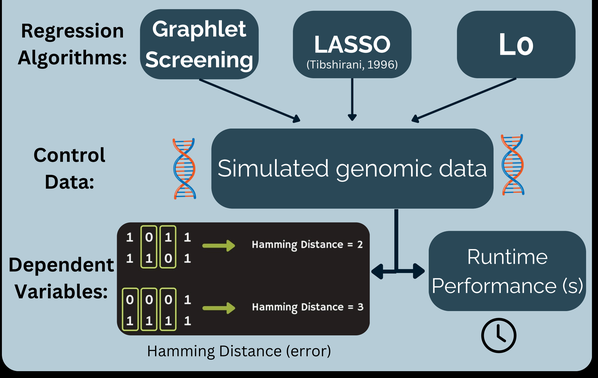
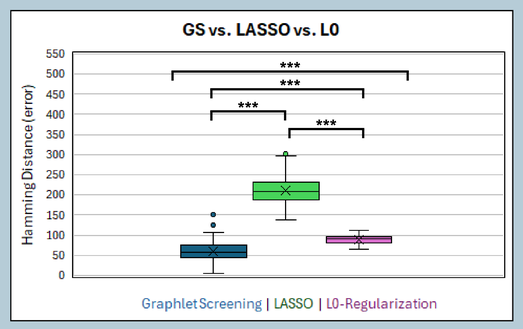


FIGURE 1

Box & Whisker Plot of Regression Model Performance Using Simulated Data



Variables of note in the graph:

- **Variances** of error (Hamming Distance)
- **Averages** of error (Hamming Distance)

Hamming Distance refers to how far off the genomic signals were for a certain algorithm. Therefore, **lower values are better**.

Graphlet Screening performs the best, but **L0** has the least **variance**. This is likely because **Graphlet Screening** (on about 2% of the trials) received **outliers** that **benefited** its performance. **Excluding these outliers** yields an average **still better** than L0 (61.48 vs. 89.28, diff = 27.8).

FUTURE STEPS

- Validate the success of the model in **new** genomic **studies**
- Make the model more **accessible** to non-programmers
- Expand **previous studies** using Graphlet Screening based pipelines (using *All of Us*)

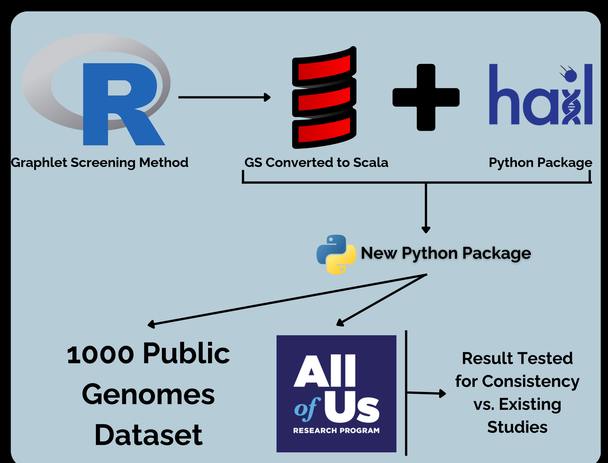


FIGURE 2

Table of Regression Model Performance Using Simulated Data

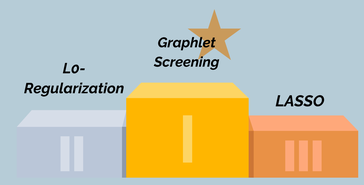
Regression Model	Simulation Count	Sum of Hamming Distance (error)	Average Hamming Distance (error)	Variance (Standard Deviation)
Graphlet Screening	**100**	**6041**	**60.41**	**572.81**
LASSO	100	21389	213.89	846.32
L0 Regularization	100	8928	89.28	93.41

Controls:

- High Sparsity (v = 0.35)
- Low Minimum Signal Strength (r = 3.5)

Dependent Variables:

- Average of Hamming Distance (error)
- P-value = **1.7E-148** (**ANOVA**)



Graphlet Screening received **25% less error** than L0 and **72% less error** than LASSO.

REFERENCES

Creighton C. J. (2023). Gene Expression Profiles in Cancers and Their Therapeutic Implications. *Cancer journal (Sudbury, Mass.)*, *29*(1), 9–14. https://doi.org/10.1097/PPO.0000000000000638

Horowitz, J.L. (2015), Variable selection and estimation in high-dimensional models. Canadian Journal of Economics/Revue canadienne d'économique, 48: 389-407. https://doi.org/10.1111/caje.12130

Jin, J., Zhang, C.-H., & Zhang, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *Journal of Machine Learning Research: JMLR, 15*(79), 2723–2772.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.* Series B (Methodological), 58(1), 267–288. http://www.jstor.org/stable/2346178

Special thanks to my advisors, Dr. Zheyang Wu and Dr. Kevin Crowthers