RARE: Machine Learning Approach for Binning Rare Variant Features to Detect Association with Disease

Satvik Dasariraju, The Lawrenceville School

Introduction: Rare Variant Features

In biomedical data, features with rare variant states (frequency below 0.05) pose challenges for analysis. The contribution of rare variant features in disease etiology remains largely unexplained, though they are widely believed to hold the key to explaining missing heritability in complex diseases [1-2].

Traditional univariate association analyses lack sufficient statistical power due to the low frequency of rare variants, thus binning (i.e., grouping) has been presented as a strategy [3].

Limitations of previous rare variant binning methods [4-7]:

- Depend on expert knowledge for initialization
- Not data-driven, rely on assumptions
- Fail to consider feature interactions (e.g., epistasis)
- Applications restricted to only genetics

Background: HLA Amino Acids in Transplantation

Kidney transplantation failure is a serious condition affecting about 20% of transplant recipients [8]. HLA amino acid mismatches are risk factors for kidney transplantation failure [9].

However, the low frequency and multicollinearity of amino acid mismatches poses difficulties in univariate analysis. Thus, a data-driven rare variant binning method would be very useful to elucidate patterns in kidney transplantation failure. Relevant questions include the roles of HLA-DRB1 vs. -DQB1 loci.

Purpose and Hypothesis

An evolutionary algorithm (a type of machine learning approach that optimizes by mimicking natural selection) is a suitable method for iteratively constructing bins of rare variant features to optimize bin association to outcome.

To overcome the limitations of previous rare variant feature analysis methods, this study aimed to:

1) Invent an evolutionary algorithm (RARE) that can reliably construct bins of rare variant features with optimal association to outcome (i.e., disease)
2) Validate RARE by developing rare variant data simulators
3) Apply RARE to HLA amino acids and kidney graft failure

The hypothesis is that RARE will display accurate performance on the simulated data trials and construct bins with higher association to transplantation failure compared to expert knowledge bins by taking a data-driven approach.

Materials

For the application of RARE to construct bins of HLA amino acid positions with association to kidney transplantation failure:

- Source of data: Scientific Registry of Transplant Recipients
- 49,459 deceased donor kidney transplantations from 2005-17
- 306 HLA amino acid positions (features)
- Data from 5 HLA loci (A, B, C, DQB1, and DRB1)
- Data was processed through imputation from serologic antigen specificities and haplotype frequencies

Methods: The RARE Algorithm

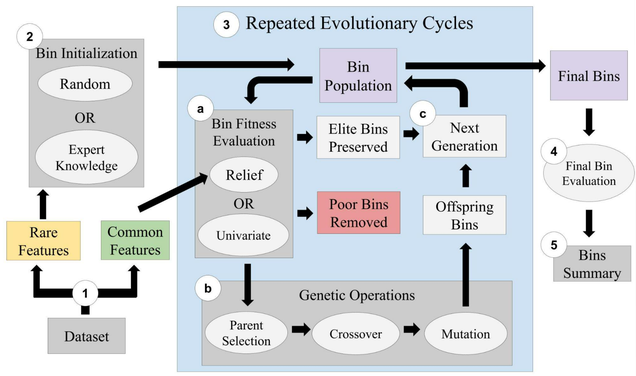

Figure 1. Schematic of the RARE algorithm's stems, including (1) data preprocessing, (2) bin initialization, (3) evolutionary cycles consisting of bin fitness evaluation, genetic operations, and creation of next generation of bins, (4) final bin scoring, and (5) summary of top bins. *All figures by student researcher.*

Bin Initialization

While previous methods depended on expert knowledge for initialization of bins, RARE offers flexibility:

1. Random initialization: rare variant features are randomly grouped into bins (option for flexible or constant bin size).
2. Expert knowledge initialization: groups of features believed to be informative fromdomain knowledge can be imported.

Bin Scoring

Bin fitness evaluation drives optimization of bin association to outcome. RARE presents two options for bin scoring based on what is relevant in the problem domain:

1. Univariate: RARE quantifies a bin's univariate association to outcome through chi-square test of independence.
2. Relief: RARE uses the MultiSURF algorithm to detect interactions between bins of rare variant features and/or features to construct bins whose interactions are informative. Previous methods lack the ability to detect interactions.

Genetic Operations

In each evolutionary cycle of RARE, elite bins are preserved for the next generation while poor bins are removed. High scoring bins are probabilistically chosen (through tournament selection) to be parent bins. Each pair of parent bins undergoes crossover to initialize a pair of offspring bins. The pair of offspring bins undergoes the mutation operation, which randomly adds, removes, or swaps a rare variant feature in the bin, and then the pair is added to the next generation of bins.

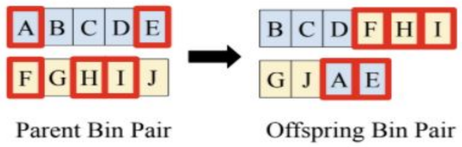

Figure 2. RARE's crossover operation. Letters represent rare variant features.

Methods: Data Simulators for Validation of RARE

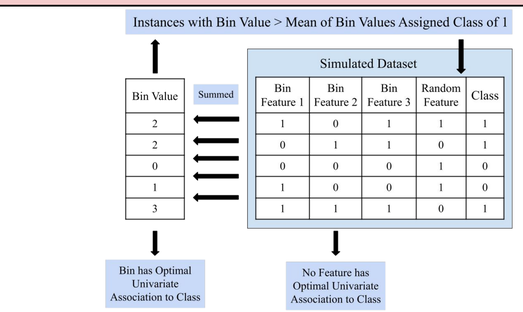

Figure 3. Data simulator for univariate association bin.

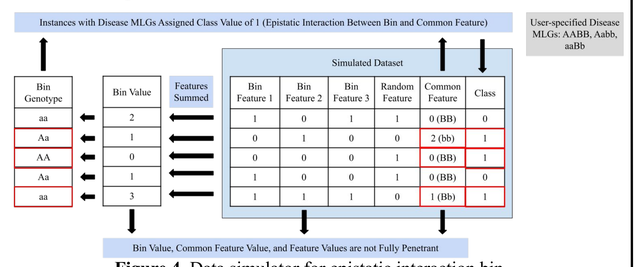

Figure 4. Data simulator for epistatic interaction bin.

Results: Validation with Simulated Data

The data simulators were used to set up experiments where the optimal bin was known (since the data was simulated) and RARE's accuracy in constructing optimal bins was measured.

Comparison 1: Univariate Data, Different Levels of Noise

The data simulator for univariate bin tested RARE's accuracy on data with varying noise. RARE with MultiSURF scoring achieved 94.22% accuracy in binning features that belonged in the known optimal bin with data without noise, 91.56% with 0.05 noise, 86.11 with 0.1 noise, and 9.00% with 0.5 noise (30 replicates for each noise level, 0.5 noise was negative control).

Comparison 2: Different Configurations of RARE

Tested on data from the simulator for univariate bin, RARE achieved 100.00% accuracy with univariate scoring, 94.22% accuracy with MultiSURF scoring, and 99.00% accuracy with constant bin size (30 replicates for each configuration).

Comparison 3: Expert Knowledge Vs. Random Initialization

RARE with partial expert knowledge input (99.00% accuracy, average of 21.57 cycles across 30 trials to reach 80% accuracy) was compared to RARE with random initialization (94.22% accuracy, average of 286.48% accuracy). **Results indicate that expert knowledge helps RARE's accuracy and efficiency.**

Comparison 4: Data with Epistatic Interaction

RARE with univariate scoring failed to construct bins when tested on the data simulator for epistatic bin (33% accuracy), but RARE with MultiSURF scoring achieved 96.00% accuracy (30 replicates for each scoring method).

Results validate RARE's ability to reliably construct optimal bins with both univariate and epistatic effects.

Results: Application to Kidney Transplantation

RARE with univariate scoring and constant bin size was used to analyze the relationship between HLA amino acid mismatches and kidney transplantation failure. Bins were compared to expert knowledge bins extracted from sequence feature variant type (SFVT) categories for protein domains and amino acid motifs with structural-functional annotation for peptide-binding pockets and T-cell receptor contact sites.

Table 1. Results from RARE's amino acid bins and transplantation data compared to the amino acid position and SFVT with highest associations to graft failure.

	Bin Size	χ^2 Value	p-value
DRB1-13	1	19.11	12.30e-6
Best SFVT	21	208.68	2.66e-47
All Loci	15	240.22	3.53e-54
A	15	76.43	2.28e-18
B	15	43.69	3.85e-11
C	15	69.99	5.96e-19
DQB1	15	208.68	4.58e-52
DRB1	15	154.47	1.82e-35

First, RARE was used to bin across all 5 available HLA loci with a constant bin size of 15 (informed by the size of typical sequence feature variant type categories). RARE's best bin shared 12 amino acid positions with the expert knowledge bin with highest association to outcome, demonstrating RARE's ability to output biologically relevant bins. RARE's best bin also had a higher association compared to the best SFVT bin, displaying the utility of RARE's data-driven approach. RARE's best bin was composed of primarily amino acid positions at peptide binding sites of DRB1.

RARE was also used to bin amino acid positions in each individual HLA locus separately. The DRB1 bin had the highest association, next DQB1, then A, C, and lastly B.

Conclusions

RARE, an evolutionary algorithm for constructing bins with optimal association to outcome, was presented.

- Through tests on simulated data, RARE's performance with both univariate and epistatic bins was validated.
- RARE can operate both with or without expert knowledge.

The algorithm was applied to elucidate the role of HLA amino acid mismatches in kidney transplantation; results suggest that mismatches at peptide-binding sites of DRB1 confer high risk of kidney transplantation failure.

Applications of this study are threefold:

1) The presented algorithm overcome previous limitations and can be applied in a wide range of domains such as genetics.
2) Results from RARE analysis on transplantation failure will guide collaborators' work on developing the new CRPA (a sensitization test conducted to guide donor-recipient pairings)
3) Results can guide the Organ Procurement and Transplantation Network's new continuous distribution initiative and generally aid in predicting graft failure [10].

Ongoing further work will analyze transplantation survival within and beyond one year separately.

References

1. Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? The American Journal of Human Genetics, 69(1), 124–137. https://doi.org/10.1086/321272
2. Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. Trends in Genetics, 17(9), 502–510. https://doi.org/10.1016/S0168-9525(01)02410-6
3. Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. The American Journal of Human Genetics, 83(3), 311–321. https://doi.org/10.1016/j.ajhg.2008.06.024
4. Morgenthaler, S., & Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 615(1–2), 28–56. https://doi.org/10.1016/j.mrfmmm.2006.09.003
5. Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genetics, 5(2), e1000384. https://doi.org/10.1371/journal.pgen.1000384
6. Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., & Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. The American Journal of Human Genetics, 91(2), 224–237. https://doi.org/10.1016/j.ajhg.2012.06.007
7. Moore, C. B., Wallace, J. R., Frase, A. T., Pendergrass, S. A., & Ritchie, M. D. (2013). BioBin: A bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. BMC Medical Genomics, 6(2), S6. https://doi.org/10.1186/1755-8794-6-S2-S6
8. Davis, S., & Mohan, S. (2021). Managing patients with failing kidney allograft: Many questions remain. Clinical Journal of the American Society of Nephrology, CJN.14620920. https://doi.org/10.2215/CJN.14620920
9. Kamoun, M., McCullough, K. P., Maiers, M., Fernandez Vina, M. A., Li, H., Teal, V., Leichtman, A. B., & Merion, R. M. (2017). HLA amino acid polymorphisms and kidney allograft survival. Transplantation, 101(5), e170–e177. https://doi.org/10.1097/TP.0000000000001670
10. Continuous distribution. Retrieved December 20, 2021, from https://optn.transplant.hrsa.gov/policies-bylaws/a-closer-look/continuous-distribution/

Acknowledgments

This project was conducted during my time as an employed research intern at Penn Medicine under Dr. Ryan J. Urbanowicz. Collaborators include Dr. Malek Kamoun (Penn Med), Dr. Loren Gragert (Tulane), and Grace Wager (Tulane).

This project has been published and presented as a paper at the Genetic and Evolutionary Computation Conference and an abstract at the American Society for Histocompatibility and Immunogenetics. I am first author of these publications.

Code for open source at https://github.com/UrbsLab/RARE

Supported by NIH NIAID HLA/KIR Region Genomics 1U01AI152960-01.

Penn Medicine