Machine Learning Approach for Binning Donor-Recipient HLA Amino Acid Position Mismatches to Detect Association with Kidney Graft Failure

Satvik Dasariraju[4], Grace L. Wager[1], Loren Gragert[2], Malek Kamoun[3], Ryan Urbanowicz[4]

1. Tulane University, New Orleans, LA, United States
2. Tulane Cancer Center, Tulane University, New Orleans, LA, United States
3. Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, United States
4. Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA, United States

Aim: Analyze associations of combinations of HLA amino acid (AA) mismatches (MMs) with kidney graft failure (GF) using the Relevant Association Rare-variant-bin Evolver (RARE), a novel machine learning approach that discovers novel AA-MM bins as constructed predictive features.

Methods: Data on 49459 deceased donor kidney transplants from 2005-2017 (both donors and recipients were White) were obtained from the Scientific Registry of Transplant Recipients. We imputed high resolution HLA-A, -B, -C, -DRB1, and -DQB1 alleles (306 in total) from serologic antigen specificities using haplotype frequencies developed by National Marrow Donor Program, then assigned AA polymorphisms for each position. HLA AA-MM studies are challenged by rare variants and multicollinearity, so we analyzed additive effects of multiple MMs. We applied RARE (Fig. 1), a data-driven algorithm that evolves AA-MM bins to optimize bin association to GF. RARE binned AAs across all five loci, as well as AAs within each locus separately. For comparison, we grouped AA positions using expert knowledge Sequence Feature Variant Type (SFVT) categories for protein domains as well as AA motifs with structural/functional annotation including AAs composing peptide binding pockets, T-cell receptor contact sites, etc.

Results: When binning across all five loci, RARE's best bin (containing 15 AAs) shared 12 AAs with the top SFVT bin (21 AAs), supporting the utility of SFVT expert knowledge and RARE's ability to discover informative bins. RARE's top bin had stronger association to GF, X^2 (29) = 240.22, $p < 10^{-53}$, than the best SFVT bin, X^2 (33) = 208.68, $p < 10^{-46}$. When RARE was applied to each individual locus, the HLA-DRB1 bin had the highest association, X^2 (28) = 230.53, $p < 10^{-51}$, followed by -DQB1 bin, X^2 (29) = 154.47, $p < 10^{-34}$. Excluding -DQB1 AA-MMs only slightly decreased bin association with GF.

Conclusions: RARE effectively automates the discovery of AA-MM bins with optimal association to GF. Results suggest that MMs at peptide-binding sites of HLA-DRB1 are most strongly associated with GF, while MMs in -DQB1 have a weaker association. We expect this work to be applicable in a clinical setting for evaluating donor-recipient pairs and predicting GF.

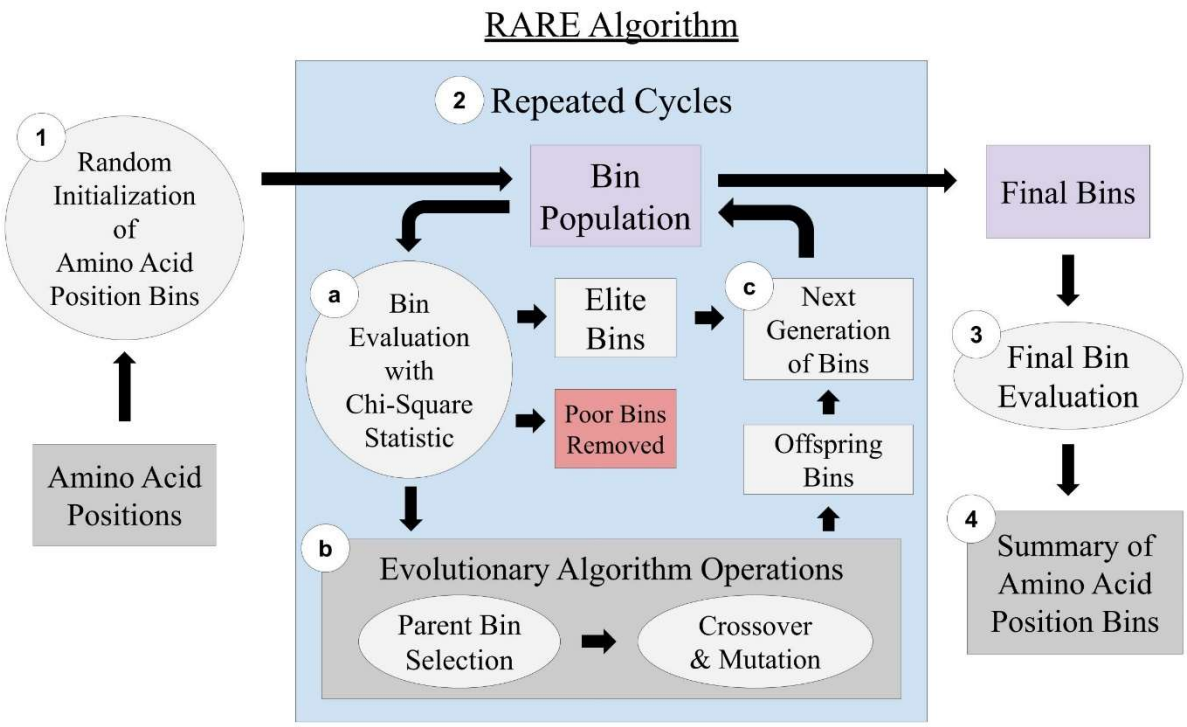

Figure 1. Schematic illustration of the RARE algorithm: (1) random initialization of bins of amino acid positions, (2) evolutionary algorithm cycles consisting of bin scoring with chi-square statistic, operations to select 'parent' bins and produce 'offspring' bins, and creation of the next generation of bins, (3) final bin evaluation, and (4) summary of discovered bins.