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

Satvik Dasariraju

Purpose

To develop an evolutionary algorithm (RARE) to construct bins of rare variant features with association to disease outcome

- Overcome the limitations of traditional, univariate association analyses (lack of statistical power)
- Add novel, critical capabilities to rare variant feature methods: detect interactions, remove dependence on expert knowledge, and versatility for applications beyond solely genetics
- Apply algorithm to HLA amino acid mismatches and transplantation

Validation of RARE Algorithm

- RARE's accuracy in binning rare variant features measured using simulated datasets with known optimal bins
- **Validation of Univariate Binning**: 100% accuracy using chi-square scoring as fitness metric, 94.22% accuracy using MultiSURF scoring
- **Validation of Epistatic Interaction Binning:** 96% accuracy using MultiSURF scoring, first-ever rare variant binning method to detect interactions between features (e.g., epistasis)
- Expert knowledge makes RARE more efficient, but algorithm can operate without expert knowledge unlike previous methods

Methods

- Development of RARE algorithm in python (made open source)
- Steps of RARE: bin initialization (option for expert knowledge input), evolutionary cycles of optimization, and final bin output
- Development of rare variant data simulators to validate RARE
- For application of RARE on kidney transplantation:
  - HLA mismatch data on 49,459 donor-recipient pairs
  - 306 HLA features (from A, B, C, DQB1, and DRB1 loci)
  - Data from Scientific Registry of Transplant Recipients
  - Imputation from serologic antigen specificities and haplotype frequencies

Application of RARE in Transplantation

- RARE applied to construct bins of HLA amino acid positions to detect association with kidney graft failure
- RARE's best bin, $X^2(29) = 240.22$, $p < 10e\text{-}53$, outperformed the best expert knowledge (sequence feature variant type) bin, $X^2(33) = 208.68$, $p < 10e\text{-}46$
- RARE's best bin with 15 amino acid positions shared 12 amino acid positions with the best expert knowledge bin
- RARE will be applicable for evaluating donor-recipient pairs and inform the U.S. gov's OPTN new continuous distribution initiative for transplant allocation