Rajan Kapoor

PhD, Electrical Eng
rkapr.github.io

r.kapoor@tamu.edu
979-398-2987

linkedin.com/in/rkapr
github.com/rkapr

Objective: Looking for internship/co-op in data science/bioinformatics.

Summary: 3+ years of experience in data analysis of bioinformatics datasets in crop science with focus on wheat and sorghum. Worked with plant breeders in Texas A&M AgriLife to identify research problems of interest, creating bioinformatics pipelines to collect, clean and process multi-TB publicly available gene expression datasets, and identify and evaluate predictive statistical models from published research for genome wide gene network analysis.

Education:

PhD, Electrical and Computer Engineering (GPA: 3.82) 2017 – 2021
Texas A&M University, College Station

MS, Electrical and Computer Engineering (GPA: 3.75) 2015 – 2017
Texas A&M University, College Station

BTech, Electrical Engineering (GPA: 8.57/10) 2010 – 2014
Indian Institute of Technology, Patna

Projects:

Understanding nitrogen transport pathways related to grain protein content in wheat [Spring 2019]
• Identifying gene regulatory pathways that transport nitrogen from senescing leaves to grain with goal of understanding mechanisms affecting grain protein content trait.
• Performed WGCNA based gene clustering for seed gene expression data on computing cluster and GO analysis.
• Collected and cleaned publicly available alignment files (CRAM) from multiple SRA projects. using globus command line, extracted reads mapping to UMAMI genes using SAMtools for verifying gene models (shell).
• Performed exploratory analysis of UMAMI gene expression in Arabidopsis to verify orthologs and identify phylogenetically conserved expression patterns (R).
Key achievement: Identified two modules involved in activating storage molecules including starch, lipids and protein, discovered multiple TFs as possible key candidate genes among modules involving amino acid transporters.
Gaussian Graphical model with fused lasso penalty for learning causal transcriptional regulations of Sorgoleone biosynthesis genes in Sorghum [Fall 2020]
• Inferring transcriptional regulation of sorgoleone biosynthesis to understand herbivory resistance in sorghum roots with goal of developing biological herbicides.
• Used HISAT2/featureCounts/DEseq2 to map, count and extract differentially expressed genes from RNAseq data.
• Verified the algorithm was able to extract causal gene interactions for sorghum circadian genes.
• Used root and leaf development data from stay green sorghum to identify TF regulators of sorgoleone genes.
• Verified TF-gene regulations using TF motifs from plant transcription factor database and comparative genomics.
Key achievement: Manuscript under preparation.
***Zeroinfl*: Zero Inflated Poisson regression (ZIP) in Python** (with Eric Chuu)[Spring 2018]
• ZIP regression was use to model a process switching between perfect state with no errors and imperfect one with Poisson distributed errors.
• The algorithm provided better fit to model systemic departure from Poisson regression than generalized linear regression model (GLM) (quasi-likelihood Poisson, negative binomial) and generalized linear mixed models.
• Implemented logit, probit, complementary log log, Cauchy and log link functions for GLM fit using object-oriented programming.
• Wrote functions for likelihood, log-likelihood, gradient and max likelihood estimation using BFGS optimization, with expectation maximization-based initialization.
• Wrote functions *summary* for pretty-printing results including p-value, standard error, z-statistics and *predict* for using fitted model for prediction.
• The functionality was extended to include zero inflated negative binomial and zero inflated geometric regressions.

Key achievement: The code was the first open source contribution to provide zero inflated regression functionality in python based on *pscl* package in R.

Adaboost based face detection using Voila Jones framework [Fall 2019]

• Implemented five stage Adaboost classifier in python with decision stump as weak learner for detecting face images using vertical, horizontal and diagonal Haar features. Code vectorization was extensively used for speedup.

• Observed the effect of changes in false negative and false positive penalties on empirical accuracy.

Key achievement: Achieved **3.5x improvement** in training time at the cost of 2.25% reduction in empirical accuracy by filtering robust features.

Minimum description length based Boolean network learning [Spring 2020]

• Used cost function consisting of sum of the error and model-description entropies as analogues of resubstitution-error and VC dimension for model-complexity regularization.

• Used model and noise codelengths as universal sufficient statistics decomposition of total codelength under normalized log-likelihood model by partitioning the parameter space such that adjacent partitions are separated by fixed Kullback-Leibler distance.

• The code was tested and verified on Boolean time series data generated using BoolNet package in R.

Key achievement: The MATLAB code based on Dougherty et al. paper was released for open source use.

Gaussian Mixture Modeling for Cancer Heterogeneity [Fall 2016]

• Used gaussian mixture modelling to model cancer heterogeneity as mixture of Boolean networks.

• Expectation maximization was then used to estimate the proportion of each subpopulation.

Key achievement: Results were published in IEEE/ACM TCBB journal.

Skills:

• R, Python, Shell scripting, SLURM and LSF batch processing, MATLAB, SQL, Jupyter notebooks, R Notebooks, C/C++, visualization: plotly, matplotlib, ggplot2, flexdashboard, python: numpy, pandas,scikit-learn, scipy.

• ML models: SVM with kernels, K-means, decision trees, Adaboost, linear/logistic/generalized linear regression and parsimonious model selection (AIC/BIC), lasso, ridge regression, mixture models, stochastic gradient/coordinate descent, random forests, time series analysis (ARMA/ARIMA/ARCH/GARCH), ANOVA.

• BLASTing on cluster, bioinformatics tools (HISAT2, featureCounts, SAMtools, BAMtools, DESeq2), phylogenetic analysis, experience using EMBL-EBI and Uniprot REST APIs, globus-cli, GitHub.

• Some experience with mixed models, image processing (imutils, cv2), graphical models.

Relevant Courses:

Engineering/Statistics: Regression Analysis (A), Statistical computing in R & Python (A), Applied Statistics & Data Analysis (B), Pattern Recognition (A), Distribution Theory (A), Information Theory (A).

Biology/Bioinformatics: Bioinformatics (S), Bioinformatics Command Line (A), Metagenomics (A), Differential Gene Expression (A), Genome Assembly (A).

Online Certifications: Introduction to SQL, Learn the Command Line, Algorithms & Data Structures.

Publications:

A Gaussian Mixture-Model Exploiting Pathway Knowledge for Dissecting Cancer Heterogeneity, in IEEE/ACM Transactions on Computational Biology and Bioinformatics, doi: 10.1109/TCBB.2018.2869813.

Differential Graphical Modeling Identifies Possible Transcriptional Regulators of Sorgoleone Biosynthesis Genes, under preparation.

Blogs/Tutorials:

Unix scripting tutorial for processing public RNAseq datasets for gene model verification in wheat.

Honors/Awards:

• Texas Engineering Experiment Station (TEES) Research Assistantship, Sept 2015-current.

• Texas International Student Scholarship, Fall 2015, Spring 2016.

• Runner up Best BTech Project Award in Electrical Engineering, 2014.

• State of Rajasthan (India) High School Academic Excellence Award, 2009.