PANKAJ PRADEEP

65 West 106th Street, New York, NY 10025 | 332-259-4539 | pp2831@columbia.edu | LinkedIn | GitHub Pages

SUMMARY

- Skilled in core biology and biotechnology with a focus on applying bioinformatics to drive advancements in precision medicine.
- Thrives in inclusive, inter-disciplinary environments with strong problem-solving abilities and excellent communication skills.
- Motivated to contribute to computational methods to push forward personalized therapy projects toward clinical applications, aiming to impact patient outcomes through innovative bioinformatics solutions.

EDUCATION

Columbia University — **New York, NY**
Master of Science in Biomedical Engineering (Bioinformatics Focus) — *Dec 2023*

Vellore Institite of Technology (VIT University) — **Vellore, India**
Bachelor of Technology in Biotechnology — *Jul 2022*

EXPERIENCE

Columbia Irving Medical Centre - Gertrude H. Sergivesky Centre — **New York, NY**
Graduate Student Researcher - Bioinformatics — *Sep 2023 to Dec 2023*

- Performed quality control analysis of 2500 FASTQ files, leveraging fq2vcf for Bioinformatics Analysis to convert samples to VCF for further investigation of Alzheimer's Disease in underrepresented populations
- Designed a visualization package in RStudio to assess 4 AD risk factors. Package was used by lab researchers for visualization of local ancestry distinguished by 3 major sub-groups.
- Collaborated with post-docs to troubleshoot errors in pipeline set-up and multi-tasking jobs (qsub, Unix) on HPC cluster
- Conducted deconvolution of RNAseq data to single cell compositions to enhance biomarkers discovery by 30% using CIBERSORTx

GlaxoSmithKline Pharmaceuticals — **Collegeville, PA**
Summer Intern – Computational Biology Oncology — *Jun 2023 to Aug 2023*

- Investigated clinical signatures for gene and protein expression from public databases to determine drivers of tumor state
- Implemented a correlation analysis pipeline in Python to visualize and distinguish patterns across transcriptomics (mRNA) and proteomics (mass spectrometry) in 947 cancer cell lines from the Cancer Cell Line Encylopedia, based on different subgroups
- Collaborated with remote researchers in GSK Germany to ingest 12 scRNAseq cancer datasets from GEO in Python into their in-house Bioinformatics pipelines
- Proposed the usage of a novel metric "Earth Mover Distance" during a quarterly company wide ideathon "GSK Science" participated by 200+ researchers. Idea was implemented in the Target Evaluation Framework by Computational Biology team

Columbia University - Herbert and Florence Irving Institute for Cancer Dynamics — **New York, NY**
Bioinformatics Research Intern — *Oct 2022 to Jun 2023*

- Leveraged AmpliconArchitect to identify regions in genome where fragmentation or breakpoints have occurred, to generate amplicons visualizations annotated with these regions to show presence of extrachromosomal DNA (ecDNA)
- Analyzed 80 whole genome samples in the HPC (bedtools, samtools, bwa) to locate and identify regions of 50 potential hits of extrachromosomal DNA using Integrated Genome Viewer, UCSC Browser for refining amplicon images
- Worked independently and as a team in setting up Python packages for further analyzing genome data for esophageal cancer
- Organized literature surveys (PubMed) with detailed reports to further understand mechanism of ecDNA occurrence

TECHNICAL SKILLS

- **Bioinformatics:** Omics Data Analysis, Samtools, GATK, Seurat, Nextflow, GEO, TCGA, CCLE, UCSC, NCBI
- **Programming:** R/RStudio, Python, Maching Learning, Bash, Unix, MATLAB, HPC, Intermediate C++/Java
- **Research/Biology:** scRNAseq, bulk RNAseq, RNA/DNA Microarray, Molecular Biology, Genomics, DNA/RNA Extraction, PCR
- **Packages:** Bioconductor, Biopython, Seaborn, Monocle3, Scanpy, ggplot, Scikit-learn, Pandas, NumPy, Keras, Tensorflow
- **Relevant Coursework:** Machine Learning for Functional Genomics, Statistical Machine Learning for Genomics, Applied Data Science, Computational Modeling of Physiological Systems, Topics in Immunology and Immunotechnology, Bioinformatics (Genetic Alignment and Protein Engineering), Molecular Biology, Cell Biology, Genetic Engineering

RESEARCH PROJECTS

Columbia University — **New York, NY**
Enhancing TIGER Model for Cas13 Off-Target Prediction for Indel gRNA Data — *Sep 2023 to Dec 2023*

- Initiated the evaluation of various RNN and Transformer architectures for Cas13d guides with indels off-target scoring
- Re-architected CNN model for compatibility with input data with up to 3 insertions/deletions by leveraging Levenshtein distance. Enhanced model robustness by increasing tolerance of input diversity
- Modelled utils package to assess performance of re-tuned model with indel's effect as labels to infer model compatibility

Columbia University **New York, NY**

BRCA-SAE Multi-Omics Data Integration using Stacked Autoencoders *Jan 2023 to May 2023*

- Pre-processed and normalized transcriptomic, proteomic and methylation data to ensure accuracy in subsequent analysis
- Invented a novel Machine Learning Stacked Autoencoder Algorithm (97% accuracy) to infer driving biomarkers for cancer progression by learning patterns across datasets for 1000 samples ingested from TCGA and GEO for breast invasive carcinoma
- Collaborated effectively with teammate to troubleshoot errors in model set-up and training for multi-omics analysis
- Performed Feature Attribution to discover top 20 biomarkers which are drivers of cancer progression across all datasets

Columbia University **New York, NY**

Predictive Model for Alzheimer's Disease Prediction using Machine Learning *Jan 2023 to May 2023*

- Acquired and pre-processed RNA microarray data for 800 Alzheimer's patients from public databases (GEO, TCGA), ensuring data quality and consistency by removing dataset artifacts
- Performed extensive comparative analysis of gene expression between normal and diseased states using ANOVA, providing insights into differential expression patterns to extract key drivers of disease state
- Leveraged Principal Component Analysis (PCA) extract key features and Stratified Shuffle Split to create robust training and test sets to feed into regression/ML models
- Explored and evaluated various predictive models, prioritizing those with the lowest false negative rates to enhance diagnostic accuracy (>96%)

PUBLICATIONS

Reviewing the Analeptic Activity of Calcium Citrate Malate

Bharat Kwatra, Chelsea Rumao, Hiya Abrol, Ishika Gulati, **Pankaj Pradeep**, Srashti Bajpai International Journal of Pharmaceutical Sciences Review and Research, 70(2), September - October 2021; Article No. 36, Pages: 294-304

OTHER SKILLS

Software	Trello, JIRA, Asana, Microsoft Office, LaTeX, Mendeley, Google Docs, Google Colab
Languages	English: Professional Proficiency. Tamil: Native. German: Conversational.