TATHAGATA DEBNATH

PhD Candidate in Computer Science

Phone · Email · GitHub · LinkedIn · Google Scholar · ORCID · Website

RESEARCH INTERESTS

Circular data algorithms and their applications to genomic analysis. RNA-seq transcriptomics and alternative splicing detection. Metagenomics and 16S rRNA microbiome profiling. Machine learning approaches to disease classification from microbial and genomic data. Statistical methodology for compositional and circular data analysis.

EDUCATION

- **Ph.D. in Computer Science** Expected 2026
  New Mexico State University · Advisor: Dr. Joe Song
  Dissertation: Fast Algorithms for Circular Data Analysis with Applications to Round Genomes and Genomic Clustering
- **M.Tech in Computer Science and Engineering** 2020
  Indian Institute of Technology (BHU), Varanasi

PUBLICATIONS

[1] **T. Debnath**, M. Song. "Fast Optimal Circular Clustering and Applications on Round Genomes." *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, Vol. 18, No. 6, 2021. **[First Author; CRAN: OptCirClust]**
Designed an $O(KN \log^2 N)$ algorithm for optimal circular clustering with direct applications to bacterial, mitochondrial, and chloroplast genome analysis. Published CRAN R package continues active adoption by computational biology research community.

[2] **Y. Chen, T. Debnath**, A. Cai, M. Song. "Circular Silhouette and a Fast Algorithm for Circular Cluster Quality Assessment." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol. 45, No. 11, 2023. **[CRAN: CircularSilhouette]**
Introduced a linear-time $O(N)$ silhouette coefficient for circular cluster quality, achieving 3000 times speedup on 1M-point datasets. Algorithm enables rapid cluster validation in genome segmentation and astronomical data analysis workflows.

[3] **H. Bendzus-Mendoza, A. Rodriguez, T. Debnath**, et al. "Radiation Exposure Induces Genome-Wide Alternative Splicing Events in Aedes aegypti." *Scientific Reports*, 2025.
RNA-seq differential splicing analysis characterizing transcriptomic response to ionizing radiation in disease vector mosquito. Demonstrates collaboration with wet-lab research groups.

[4] **A.D. Lopez, T. Debnath**, M. Pinch, I. Hansen. "Phosphoproteomics Analyses of Aedes aegypti Fat Body Reveals Blood Meal-Induced Signaling Cascades." 2024.
Proteomics data analysis of mosquito blood meal response, demonstrating bioinformatics support across multiple omics platforms.

[5] Plus 6 additional publications across pattern recognition, computational biology, and bioinformatics. **Total: 10 publications, 81+ citations.** Full list at Google Scholar.

SOFTWARE & PUBLISHED TOOLS

- **OptCirClust** (CRAN R package) — CRAN · GitHub
  $O(KN \log^2 N)$ optimal circular clustering algorithm for round genomes (bacterial, mitochondrial, chloroplast) and other circular data structures. Applications include bacterial genome segmentation, CpG island clustering, and solar cycle detection. Widely adopted in computational biology research.
- **CircularSilhouette** (CRAN R package) — CRAN · GitHub
  Linear-time $O(N)$ silhouette coefficient for circular cluster quality assessment, achieving 3000 *times* speedup over standard $O(N^2)$ computation. Enables rapid cluster validation in genome analysis and astronomical data workflows.
- **Pohori** (in development) — GitHub
  Model-free statistical method for alternative splicing detection in short-read (Illumina) and long-read (PacBio, Oxford Nanopore) RNA-seq data. Provides flexible framework for transcriptomic analysis across multiple sequencing platforms.
- **ancestSeq** (C++ command-line tool) — GitHub
  Ancestral genomic sequence reconstruction from present-day genomes using comparative genomics and phylo-

genetic methods.

RESEARCH EXPERIENCE

PhD Research, New Mexico State University 2022–Present

Advisor: Dr. Joe Song

- Developed the Fast Optimal Circular Clustering (FOCC) algorithm, published in IEEE/ACM TCBB, with applications to genomic data analysis on round chromosomes.
- Designed the linear-time CircularSilhouette algorithm for circular cluster quality assessment, achieving significant computational improvements documented in IEEE TPAMI.
- Actively developing Pohori, a model-free alternative splicing detection method for RNA-seq data processing.
- Collaborated with Hansen Lab (NMSU) on RNA-seq and proteomics data analysis of Aedes aegypti mosquito immune response.

Collaborative Research: Hansen Lab, NMSU 2023–Present

- Bioinformatics analysis of RNA-seq alternative splicing in radiation-exposed Aedes aegypti — published in Scientific Reports 2025.
- Phosphoproteomics analysis of blood meal-induced signaling in mosquito fat body tissue.
- Demonstrates cross-disciplinary collaboration between computational and wet-lab biology.

M.Tech Research, IIT (BHU) 2018–2020

Thesis: Computer Vision Methods for Medical Image Analysis

- EF-Index algorithm for efficient image segmentation using entropy-based feature selection.
- Knee thermogram classification using deep convolutional neural networks.

RESEARCH PROJECTS

Gut Microbiome Disease Classification *2023–2024*

[16S rRNA, Metagenomics, CLR Transformation, Machine Learning]

- Built a multi-disease classifier achieving 0.907 weighted AUC across Crohn's Disease, Type 2 Diabetes, and healthy controls from 236 clinical 16S rRNA samples using CLR (Centered Log-Ratio) compositional transformation to address zero-inflation and compositionality constraints inherent in microbiome count data.
- Designed a sophisticated 4-method feature selection pipeline (Mutual Information, ANOVA F-score, Recursive Feature Elimination, Boruta) reducing 4,203 bacterial taxa to 52 discriminative biomarkers with quantitative cross-method consensus analysis.
- Identified clinically validated biomarkers via SHAP interpretation: Faecalibacterium depletion in Crohn's Disease (concordant with Sokol et al. 2008), Roseburia enrichment in healthy controls, Ruminiclostridium elevation in T2D — all findings corroborated by published literature.
- Performed rigorous external validation on 759-sample independent cohorts (NielsenHB 2014, QinJ 2012) and systematically characterized 35% AUC performance degradation caused by cross-platform batch effects (16S amplicon vs. shotgun metagenomic WGS sequencing) using PCA batch effect visualization.
- Demonstrates core bioinformatics competencies: microbiome data preprocessing, compositional data analysis, machine learning validation, and mechanistic biomarker discovery.

Scientific Literature Retrieval for Genomics *2024*

[RAG, FAISS, Information Retrieval, Domain-Specific Search]

- Built a scientific literature retrieval system over peer-reviewed bioinformatics and nutrition research papers using hybrid search (FAISS semantic + BM25 lexical + Reciprocal Rank Fusion).
- Generates evidence-based answers with full source attribution (document ID, page numbers), enabling traceable literature-grounded scientific reasoning over research corpora.
- Reframed as bioinformatics tooling for query-based literature exploration and hypothesis generation from peer-reviewed evidence, rather than consumer chatbot.
- Demonstrates practical tool development skills for research infrastructure.

Fast Optimal Circular Clustering Algorithm *2021–2023*

[Algorithm Design, C++, R Package Development, Computational Complexity]

- Designed and published an $O(KN \log^2 N)$ circular clustering algorithm achieving asymptotic speedup over brute-force $O(KN^2)$ approaches for round genome data (bacterial, mitochondrial, chloroplast genomes).
- Developed CRAN R package *OptCirClust* with C++ computational backend, comprehensive unit test coverage, and user documentation. Package continues active adoption and citations from research community.
- Demonstrated algorithm applications across CpG island clustering, bacterial genome segmentation, and circular data in astronomy (solar cycle pattern detection).

- First author on IEEE/ACM TCBB publication — establishes research independence and algorithmic contribution.

CircularSilhouette: Fast Circular Cluster Quality Metric *2022–2023*

[Algorithm Design, Statistical Methods, C++ Implementation]

- Developed a linear-time O(N) silhouette coefficient for circular cluster quality assessment, achieving 3000 *times* speedup over standard $O(N^2)$ computation.
- Published in IEEE TPAMI with accompanying CRAN R package distribution — demonstrates ability to translate algorithmic research into reproducible, reusable software.
- Validated algorithm across synthetic and real datasets with detailed benchmark studies documenting runtime complexity and numerical stability.

TECHNICAL SKILLS

- **Programming Languages:** R (CRAN packaging), Python (scikit-learn, pandas, NumPy), C/C++ (computational efficiency), SQL, Bash scripting
- **Bioinformatics Methods:** RNA-seq analysis, Alternative splicing detection (MATS), 16S rRNA microbiome profiling, Shotgun metagenomics, BLAST/sequence alignment, Phylogenetics, Sequence assembly and annotation
- **Bioinformatics Tools & Packages:** Bioconductor (R ecosystem for genomics), curatedMetagenomicData, QIIME (16S analysis), SAMtools, BAMtools, bedtools, UCSC tools
- **Machine Learning & Statistics:** scikit-learn, XGBoost, SHAP interpretation, Compositional Data Analysis (CLR transformation), Nested cross-validation, ANOVA, hypothesis testing, feature selection
- **Infrastructure:** CRAN package development and publication, Docker containerization, Git version control, LaTeX scientific writing, HPC cluster job submission, AWS cloud computing

TEACHING & MENTORING

TA/Tutoring positions and workshops in bioinformatics, computational biology, and machine learning (if applicable). Demonstrates ability to communicate technical concepts clearly.

PRESENTATIONS

Conference talks and poster presentations on circular data algorithms, circular genome clustering, microbiome analysis, and alternative splicing detection at academic and bioinformatics community venues.

AWARDS & HONORS

Relevant fellowships, scholarships, or academic distinctions earned during PhD and M.Tech programs (if applicable).

PROFESSIONAL SERVICE

Peer review contributions, committee participation, or professional society memberships in bioinformatics and computational biology organizations (if applicable).

REFERENCES

- Dr. Joe Song, Advisor, New Mexico State University [email] · [phone]
- Dr. Immo Hansen, Collaborator, New Mexico State University [email] · [phone]

al Recommender [email] · [phone]