

Dr. Ajay Arya

CONTACT INFORMATION	#202 (skbioinfo Lab) National Institute of Plant Genome Research NIPGR, New Delhi-110067	+91-9756507133 +91-8985409235 ajayarya05@gmail.com ajayarya@nipgr.ac.in
---------------------	--	--

RESEARCH INTERESTS	Structural Biology, Cell-Specific Protein-DNA Interactions, NGS (Next Generation Sequencing) Analysis, Machine Learning and Data Mining, Deep Learning, Data analysis, Bioinformatics.
--------------------	--

RESEARCH EXPERIENCE	Project Scientist-I National Institute of Plant Genome Research, New Delhi, India Bioinformatics Lab July, 2023-Present
---------------------	---

Project I: PFusionDB: Development of a comprehensive plant fusion database. July, 2023-Sep, 2023

- A pipeline has been developed for the identification of fusion transcripts in plants.
- Fusion identification was done in 2795 RNA-Seq samples of *Arabidopsis*, rice, and chickpea.
- Multiomics datasets were used for the in-silico characterization of fusion transcripts.
- Developed an extensive database encompassing four plant species..
- PFusionDB: Plant Fusion database

Project II: athisomiRDB: Development of a comprehensive plant isomiR database. August, 2023-Present

- A pipeline has been developed for the identification of isomiR in *Arabidopsis thaliana*.
- Development of a comprehensive database for isomiR of *Arabidopsis thaliana*. (In Progress...)

Project III: Comparative study of tRFs and miRNA in plants and other organism. July, 2023-Present

- Work in Progress...!

Jawaharlal Nehru University, New Delhi, India
Centre for Computational Biology and Bioinformatics, SC&IS
Projects During Ph.D., from 2017 to 2022

Project I: Development of computational predictive model to predict Protein-DNA binding residues. 2018-2021

- A novel dataset is prepared of Transient Residues Transient Binding (TRTB) from high resolution and non-redundant DNA-Binding Protein complexes.
- Sequence, Structure and gene-expression features were created for the protein sequences.
- Structural and biophysical labeling and loss/gain of DNA-binding in single residue positions were analyzed.
- Development of a machine learning model (LightGBM) that is a gradient boosting framework that uses tree based learning algorithms.
- First author of one peer-reviewed scientific publication in *Journal of Molecular Biology*.

Project II: Role of atomic solvent accessibility and other sequence-dependent features in longer-range DNase accessibility of genome-wide nucleosome occupancies. 2019-2022

- DNA structure dataset preparation from different sources.
- Solvent atomic accessibility of DNA calculation using NACCESS tool.
- Development of the regression model to predict solvent atomic accessibility using Dynaseq conformational parameters using R package.
- Analysis of chromatin accessibility data i.e. DNase accessibility in R language.
- Comparative analysis of Atomic and Chromatin accessibility.
- Tools used NACCESS, SamTool, BEDTools, bash scripting and R for Data processing, analysis, developing GLM model and plotting graphs.

Project III: Deterministic features and prediction of transcription factors specificities across cell-types. 2020-2022

- Design and implement experiment to test validity of new ideas.
- Retrieval of ChIP-seq data for transcription factors from ENCODE database and preprocessing of the data in R.
- Motif enrichment analysis on ChIP-seq data using galaxy tools i.e. meme-chip, meme, Tom-Tom etc.
- Data preprocessing and featurization of the binding sites.
- Development of machine learning model in python.

Project IV: Comparative 3D genome organization analysis of Mycobacterium tuberculosis strain (H37Ra and H37Rv) using HiC data. 2020-2022

- Collaborated with other scientists locally.
- Downstream analysis of next-generation sequencing was performed on Hi-C data.
- Hicexplorer is used to filter and process the aligned reads to identify the pairs that represent valid Hi-C interactions
- Identifying TADs and Loops, conducting a comparative analysis between H37Ra and H37Rv, and drawing biological inferences from the results.
- Tools used Hicexplorer, Bowtie, BWA, HiCUP and juicer.
- Co-authored scientific publication in bioRxiv

Note: During this period, I engaged in various side projects and had the privilege of mentoring four master's students throughout their research internships. My responsibilities included supporting project development, offering training and direction for experiment design and execution, and providing assistance with problem-solving and data analysis.

EDUCATION

Jawaharlal Nehru University, New Delhi.

Ph.D., Computational Biology & Bioinformatics 2017-2023

- Topic: *Computational modeling of genome-wide protein-DNA interactions using sequence, gene-expression and conformational dynamics.*
- Supervisor: Prof. Shandar Ahmad

University of Hyderabad, Hyderabad, Telangana.

M.Tech., Bioinformatics 2014-2016

- Topic: *Prevalence of HIV-1 Transcriptase Drug Resistance in India.*
- Supervisor: Prof. A. K. Kondapi
- Grade: 7.86 on the scale of 10.

Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Uttarakhand.

B.Tech., Biotechnology 2010-2014

- Topic: *Exploration of potential bioinoculants for improved crop productivity and diversity analysis*

- Supervisor: Prof. Reeta Goel
- Grade: 6.821 on the scale of 10.

TECHNICAL SKILLS

- **NGS:** BWA (Burrows-Wheeler Aligner), Bowtie, STAR, FastQC, Fastp, Cufflinks, DESeq2, MACS2 and GATK (Genome Analysis Toolkit).
- **Hi-C Technology:** HiCEXplorer, HiCUP and Juicer.
- **Programming Language:** Python, R, PHP, JavaScript, perl, C.
- **Deep learning tools:** PyTorch, Keras, Tensorflow, Theano, Torch.
- **Machine learning and data mining tools:** Scikit-learn, R-Packages (e1071, h2o, caret, glmnet).
- **Web scraping tools:** Scrapy, urllib/urllib2.
- **Bioconductor tools:** Biostrings, GRanges, bio3d, ChIPseeker etc.
- **Data analysis tools:** R, dplyr, tidyr, Pandas, SciPy, Numpy, Microsoft Excel.
- **Data visualization tools:** R: ggplot2, Plotly; Python: Matplotlib, Seaborn.
- **Database management tools:** MySQL, MongoDB, Joomla.
- **Operating System:** Windows, Linux.

FELLOWSHIP AND EMPLOYMENT

UPoE-II 1st August 2016 to 31st December 2016

- Project Title: *Genome-wide DNA-conformational dynamics in host-pathogen interactions.*

ICMR-SRF 12th April 2018 to 11th April 2021

- Project Title: *Investigating DNA-binding specificity of transcription factors using DNA conformational dynamics signatures.*

SERB-JRF 13th April 2021 to 12th July 2021

- Project Title: *Identification of genome-wide DNA-conformation Mycobacterium Tuberculosis novel markers for prognostic and clinical outcomes of the disease.*

DBT-Project Associate-II 1st September 2021 to 30th November 2021

- Project Title: *Data driven modeling of pathogen-specific host responses mediated by nucleic acid sensing proteins.*

ICPS-JRF 1st April 2022 to 30th June 2022

- Project Title: *Data reduction techniques for genomic and transcriptomic data sets.*

ACHIEVEMENTS AND AWARDS

- **Assisted a workshop** on “Protein Sequence and Structure Analysis Using Google Cloud Engine“, at InCoB-2019, Jakarta, Indonesia
- **ICMR-SRF** fellowship 2018-2021
- **Joint CSIR-UGC(NET 2018)** in Life Sciences
- **GATE-2014** qualified in Biotechnology
- **GATE-2016** qualified in Biotechnology

REFEREED PUBLICATIONS/BOOK CHAPTER

1. **Ajay Arya**, Dana Mary Varghese, Ajay Kumar Verma, Shandar Ahmad. “Inadequacy of evolutionary profiles vis-a-vis single sequences in predicting transient DNA-binding sites in proteins.” Journal of Molecular Biology(2022): 167640.
2. Dana Mary Varghese, **Ajay Arya**, Shandar Ahmad. “Feature-Engineering from Protein Sequences to Predict Interaction Sites Using Machine Learning.” Machine Learning in Bioinformatics of Protein Sequences: Algorithms, Databases and Resources for Modern Protein Bioinformatics. 2023. 129-151.
3. Mohit Mishra, **Ajay Arya**, Md. Zubair Malik, Rakesh Bhatnagar, Seyed E. Hasnain, Shandar Ahmad, Rupesh Chaturvedi. “High Throughput

Chromosome Conformation Capture identifies differential genome organization in virulent and avirulent strains of *Mycobacterium tuberculosis*.” bioRxiv (2022): 2022.11.10.515895

4. Dilip D. Karad, Ravi Tandon, **Ajay Arya**, Kailash D. Sonawane, Anusaya S. Chavan, Arun S. Kharat. “Subtype diversity and emergence of drug resistance in HIV-1 in solapur district of Maharashtra, India.” Iranian Journal of Microbiology 14.5 (2022): 730-739.
5. Ahmad, Shandar, Philip Prathipati, Lokesh P. Tripathi, Yi-An Chen, **Ajay Arya**, Yoichi Murakami, and Kenji Mizuguchi. “Integrating sequence and gene expression information predicts genome-wide DNA-binding proteins and suggests a cooperative mechanism.” Nucleic Acids Research 46, no. 1 (2017): 54-70.

WORKSHOPS AND CONFERENCES

- **International Conference on Recent Trends in Biotechnology and Applied Bioinformatics**, Sharda University, Greater Noida, India (Oral Presentation)

27-29 April, 2023

- **ICTP-ICTS Winter School on Quantitative Systems Biology: Machine Learning in Biology**, International Centre for Theoretical Physics, Italy (Oral and Poster Presentation)

05-16 December, 2022

- **ICTP - ICTS Winter School on Quantitative Systems Biology: Morphogenesis**, International Centre for Theoretical Sciences, India

03-20 December, 2019

- **International Conference on Bioinformatics (InCoB-2019)**, Universitas YARSI, Jakarta, Indonesia

10-12 September, 2019

- **NATIONAL SCIENCE DAY 2019**, JNU, New Delhi

28 February, 2019

- **Annual BSBE Meet**, IIT-Kanpur, Kanpur

07-09 December, 2018

- **International Conference on Bioinformatics (InCoB-2018)**, JNU, New Delhi

26-28 September, 2018

- **Har Gobind Khorana Memorial Symposium on Genes, Genomes and Membrane Biology**, Mohali, Punjab

03-05 December, 2017

- **86th Conference of Society of Biological Chemists**, JNU, New Delhi

16-19 November, 2017