Surag Nair

5th Year PhD Student | Stanford University
www.stanford.edu/~surag | surag@stanford.edu

Research Interests

Deep Learning for Regulatory Genomics, Single-cell Genomics, Computational Biology, iPSC Reprogramming

Education

Stanford University — Sep 2019 - Dec 2023 (Expected)
PhD in Computer Science
Cumulative Grade Point Average: 4.21

Stanford University — Sep 2017 - Jun 2019
MS in Computer Science

Indian Institute of Technology, Delhi — Jul 2013 - Jun 2017
B.Tech in Electrical Engineering
Cumulative Grade Point Average: 9.35/10

Relevant Coursework

Computer Science	:	Artificial Intelligence, Machine Learning, Probabilistic Graphical Models, Natural Language Processing, Computer Systems and Organization, Databases, Analysis of Networks, Cryptography, Operating Systems
Biology/Biocomputation	:	Deep Learning in Genomics and Biomedicine, Structure of Biomolecules, Chromatin Regulation of the Genome

Research and Technical Projects

Deep Learning Methods for Genomics and Single-cell Epigenomics — Apr 2018 - Present
Supervisor: Dr. Anshul Kundaje, Stanford University

Mitigating spurious feature learning in conventional models of regulatory genomics

- Demonstrated that conventional machine learning models of regulatory genomics learn spurious features that result in inconsistent predictions, misleading feature interpretation, and erroneous sequence design.
- Exploring training design decisions that mitigate the learning of spurious sequence features.

Single-cell dissection of human skin cell reprogramming with DNA sequence models

- Analyzed single-cell RNA-seq and ATAC-seq of a time course of human skin cell reprogramming to iPSCs.
- Trained DNA sequence models to predict ATAC-seq at high resolution and performed model interpretation.
- Linked concentration of transcription factors and their DNA motif sequences to reprogramming progression.
- Proposed detailed mechanisms for how skin cells lose their identity within the first two days of reprogramming.

Speeding up In-silico Saturation Mutagenesis (ISM) for convolutional sequence models

- ISM is an interpretability method for deep learning sequence models in which each position in input sequence is perturbed and propagated through a trained model to measure the effect of the mutation on the output.
- Developed fastISM, an algorithm that speeds up ISM by over 10x for convolutional neural networks.

Cis-trans deep learning models for chromatin accessibility

- Developed models that incorporate DNA sequence with RNA expression data to predict chromatin accessibility.
- Demonstrated the ability to make predictions in unseen cell types and impute missing chromatin accessibility.
- Improved state-of-the-art by introducing new features, a ResNet model architecture, and training procedures.

Timely detection of extreme failure cases for Siri — Jun 2018 - Aug 2018
Summer Internship, Apple Inc. (Siri International team), Cupertino

- Studied causes for failure in the multi-component machine learning pipeline for Siri, Apple's voice assistant.
- Devised and implemented an NLP pipeline for real-time detection of failure cases based on usage logs data.
- Workflow consisted of periodic PySpark and Python scripts running on terabytes of real-time data.

Selected Publications and Preprints

- **Transcription factor stoichiometry, motif affinity and syntax regulate single-cell chromatin dynamics during fibroblast reprogramming to pluripotency**: Surag Nair*, Mo Ameen*, Laksshman Sundaram, Anusri Pampari, Jacob Schreiber, Akshay Balsubramani, Will Wang, David Burns, Helen Blau, Ioannis Karakikes, Kevin Wang, Anshul Kundaje. bioRxiv 2023.
- **The dynseq genome browser track displays context-specific sequence features at single-nucleotide resolution**: Surag Nair*, Arjun Barrett*, Daofeng Li*, Brian Raney, Brian Lee, Peter Kerpedjiev, Vivek Ramalingam, Anusri Pampari, Fritz Lekschas, Ting Wang, Maximilian Haeussler, Anshul Kundaje. Nature Genetics 2022.
- **fastISM: Performant in-silico saturation mutagenesis for convolutional neural networks**: Surag Nair, Avanti Shrikumar, Jacob Schreiber, Anshul Kundaje. Bioinformatics 2022.
- **Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts**: Surag Nair*, Daniel Kim*, Jacob Perricone, Anshul Kundaje. Bioinformatics 2019.
- **Single-cell multiome of the human retina and deep learning nominate causal variants in complex eye diseases**: Sean Wang, Surag Nair, Rui Li, Katerina Kraft, Anusri Pampari, Aman Patel, Joyce Kang, Christy Luoung, Anshul Kundaje, Howard Chang. Cell Genomics 2022.
- **Accelerating in silico saturation mutagenesis using compressed sensing**: Jacob Schreiber, Surag Nair, Akshay Balsubramani, Anshul Kundaje. Bioinformatics 2022.
- **AP-1 is a temporally regulated dual gatekeeper of reprogramming to pluripotency**: Glenn Markov, Thach Mai, Surag Nair, Anna Shcherbina, Yu Xin Wang, David Burns, Anshul Kundaje, Helen Blau. PNAS 2021.
- **Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays**: Rajiv Movva, Peyton Greenside, Georgi K. Marinov, Surag Nair, Avanti Shrikumar, Anshul Kundaje. PLOS ONE 2019.
- **Inferring Temporal Knowledge for Near-Periodic Recurrent Events**: Dinesh Raghu*, Surag Nair*, Mausam. International Joint Conference on Artificial Intelligence (IJCAI) 2018.

** equal contribution*

Open Source Contributions

Multi-framework Alpha Zero [3000+ stars] WINTER 2017-18

- Developed a package for self-play based learning following the Alpha Zero paper by DeepMind. Allows easy addition of new games and works with all major deep learning frameworks (PyTorch, TensorFlow, Keras).

Stanford CS230 Deep Learning Starter Code [2000+ stars] WINTER 2017-18

- Developed starter code for deep learning projects in PyTorch with accompanying tutorials.

PyTorch Implementation of seqGAN Algorithm [600+ stars] AUTUMN 2017-18

- Implemented a LSTM based deep learning language model and trained it using a Generative Adversarial Network framework using Policy Gradients, in PyTorch, based on the paper by Lantao Yu et al. 2016.

Technical Skills

Languages	:	Python, R, C++, SQL
Softwares/Tools	:	PyTorch, TensorFlow, NumPy, PySpark

Invited/Selected Talks

- **Invited**: IMSc Chennai (Aug 2023): Motif syntax of fibroblast reprogramming
- **Selected (Travel Fellowship)**: ISMB (France, Jul 2023): Feature leakage in models of regulatory DNA
- **Invited**: IGVF Seminar Series (Apr 2023): Motif syntax of fibroblast reprogramming
- **Invited**: Bay Area Chromatin Club (Jul 2022): Motif syntax of fibroblast reprogramming
- **Selected**: Cold Spring Harbor Systems Biology (Mar 2022): Motif syntax of fibroblast reprogramming
- **Invited**: Genentech (Mar 2022): Motif syntax of fibroblast reprogramming
- **Selected**: ISMB 2021 (Jul 2021): Motif syntax of fibroblast reprogramming ▶
- **Selected**: MLCB 2020 (Nov 2020): Speeding up in-silico saturation mutagenesis (fastISM) ▶
- **Selected**: ISMB 2019 (Switzerland, Jul 2019): Cis-trans deep learning models for chromatin accessibility ▶

Professional Activities

Journal Reviewer

- OUP Bioinformatics: 2019, 2023
- BMC Bioinformatics: 2022, 2023
- PLOS Computational Biology: 2022, 2023
- IEEE Transactions on Computational Biology and Bioinformatics: 2023
- Journal of Open Source Software: 2022, 2023
- Review Commons: 2021

Conference Reviewer

- ICML Workshop on Computational Biology: 2023
- International Conference on Intelligent Systems for Molecular Biology (ISMB): 2023
- Machine Learning in Computational Biology (MLCB): 2021, 2022

External Teaching

- Cold Spring Harbor Statistical Analysis of Genome Scale Data, Cold Spring Harbor, USA: 2022
- Machine Learning in Health and Disease, International Centre for Theoretical Sciences, India: 2023

Teaching Experience

CS230: Deep Learning WINTER 2017-18

Course Instructors: Dr. Andrew Ng & Kian Katanforoosh, Stanford University

- Project mentor for 14 teams who applied deep learning to domains including space imagery, translation, genomics, and photography. Developed questions for midterm examination. Graded exams and projects for 400+ students.

Scholastic Achievements

2017 Department Rank 5 (top 6%) among students of Electrical Engineering Department, Batch of 2017
2013 All India Rank 280 in IIT Joint Entrance Exam-Advanced out of over 1 million students
2013 National Top 40 in Indian National Chemistry Olympiad

Mentoring

- Areeb Gani (Montgomery Blair High School): 2022
- Arjun Barrett (The Harker School): 2021