Saket Choudhary

CONTACT INFORMATION

2677 Ellendale Place
Apartment #215
Los Angeles
California 90007

Email: saketkc@gmail.com
Homepage: https://saketkc.github.io
[Google Scholar] [CrossValidated]
[Github]

EDUCATION

University of Southern California, Los Angeles, USA
PhD Candidate, Computational Biology and Bioinformatics [2014 - Ongoing]

University of Southern California, Los Angeles, USA
Masters in Computer Science (Data Science Track)
Department of Computer Science [2018 - 2019]

University of Southern California [USC], Los Angeles, USA
Masters in Statistics, Department of Mathematics [2016 - 2018]

Graduate Diploma in Statistics, Royal Statistical Society, London, England
Cleared four out of five modules [2016-2017]

Indian Institute of Technology Bombay [IITB], Mumbai India
Bachelor of Technology, Master of Technology, Chemical Engineering [2009-2014]
Masters Thesis: Pattern Recognition in Clinical Data

HONORS AND AWARDS

- Open Bioinformatics Foundation Travel Fellowship [2019]
- ISMB Travel Fellowship [2019]
- Outstanding Teaching Assistant, USC [2019]
- Provost Fellowship, awarded to outstanding incoming PhD students at USC [2014]
- Gandhian Young Technological Innovation Award by Indian Institute of Management Ahmedabad, for designing a low cost water impurity detection device [2013]
- Institute Technical Special Mention for three consecutive years at IITB [2010-12]
- Undergraduate Research Award, IITB [2012]
- Kishor Vaignyanik Protsahan Yojana (KVPY) Fellowship by Department of Science and Technology (DST), Government of India [2007]
- Homi Bhabha Young Scientists' Gold Medal [2005]

OLYMPIADS

- **Top 6** to be selected for **Indian National Mathematics Olympiad (INMO)**, selection level exam for International Mathematical Olympiad(IMO) [2008]
- **Top 30** in **Regional Mathematics Olympiad (RMO)** [2009]
- **Top 250** in **Indian National Physics Olympiad (INPhO)** [2009]
- **Top 300** in **Indian National Astronomy Olympiad (INAO)** [2009]

PUBLICATIONS /PREPRINTS

† = equal contribution.

1. Wenzheng Li†, **Saket Choudhary**† and Andrew D. Smith *Accurate detection of short and long active ORFs using Ribo-seq data*. [Under review]
2. Juhye M. Lee, Rachel Eguia, Seth J. Zost, **Saket Choudhary**, Patrick C. Wilson, Trevor Bedford, Terry Stevens-Ayers, Michael Boeckh, Aeron Hurt, Seema S. Lakdawala, and Scott E. Hensley. *Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin* bioRxiv (2019). [Preprint]
3. **Saket Choudhary**. *pysradb: A Python Package to Query next-Generation Sequencing Metadata and Data from NCBI Sequence Read Archive.* F1000Research, vol. 8, F1000 (Faculty of 1000 Ltd), Apr. 2019, p. 532. [Online]
4. **Saket Choudhary**. *Golden-Thompson via pinching inequality.* arXiv preprint arXiv:1811.00544 (2018).[Preprint]
5. Syed Asad Rahman, Gilliean Torrance, Lorenzo Baldacci, Sergio Martínez Cuesta, Franz Fenninger, Nimish Gopal, **Saket Choudhary**, John May, Gemma L. Holliday, Christoph Steinbeck and Janet M Thornton. *Reaction Decoder Tool (RDT): extracting features from chemical reactions.* Bioinformatics 32, no. 13 (2016): 2065-2066. [Online]
6. **Saket Choudhary**, Leyla Garcia, Andrew Nightingale, and Maria-Jesus Martin. *BioJS-HGV Viewer: Genetic Variation Visualizer.* bioRxiv (2015): 032573. [Preprint]
7. Syed, Parvez, Shabarni Gupta, **Saket Choudhary**, Narendra Goud Pandala, Apurva Atak, Annie Richharia, Heng Zhu et al. *Autoantibody Profiling of Glioma Serum Samples to Identify Biomarkers Using Human Proteome Arrays* Scientific reports 5 (2015). [Online]
8. Yachdav, Guy, Tatyana Goldberg, Sebastian Wilzbach, David Dao, Iris Shih, **Saket Choudhary**, Steve Crouch et al. *Anatomy of BioJS, an open source community for the life sciences.* eLife 4 (2015): e07009. [Preprint]
9. **Saket Choudhary**, and Santosh B. Noronha. *GalDrive: Pipeline for comparative identification of driver mutations using the Galaxy framework.* bioRxiv (2014). [Preprint]
10. **Saket Choudhary**, Vishnu Raj, K. Sanmugasundaram, Gyan Singh Patel, and Kannan Moudgalya. *Scilab on Cloud and Textbook Companion Project: A Web 2.0 Service for Open Source Education.* In 2013 International Conference on Cloud Computing and Big Data. [Online]
11. Gatkine, Pradip, Swati Gatkine, Sushanth Poojary, **Saket Choudhary**, and Santosh Noronha. *Development of piezo-electric sensor based noninvasive low cost Arterial Pulse Analyzer.* In Biomedical Engineering International Conference (BMEiCON), 2013 6th, pp. 1-4. IEEE, 2013. [Online]
12. Dilip Save, Yogesh, R. Rakhi, N. D. Shambhulingayya, Amit Srivastava, Manas Ranjan Das, **Saket Choudhary**, and Kannan M. Moudgalya. *Oscad: An open source EDA tool for circuit design, simulation, analysis and PCB design.* In Electronics, Circuits, and Systems (ICECS), 2013 IEEE 20th International Conference on, pp. 851-854. IEEE, 2013. [Online]

TALKS/POSTERS

1. *BioJS-HGV: Human Genetic Variation Viewer*, USC-UCLA Joint Symposium 2015 [Talk]
2. *MoCA: A tool for Motif Conservation Analysis*, USC-MCB Retreat 2015 [Poster]
3. *A reference database and method for identifying actively translating ORFs from Ribo-seq data*, Biology of Genomes 2019 [Poster]

Research Experience

Detecting cancer in Histopathology Images, Research Intern May 2018 - July 2018
Guide: Dr. Radhakrishna Bettadapura *Strand Life Sciences, India*

I developed a core library for applied machine learning on histopathology images. Using unsupervised segmentation and random forests, our method performs at par with other deep learning approaches that have been applied to this problem in literature.

Source: http://github.com/saketkc/pyvirchow

Computational methods for deciphering translational regulation, PhD Project Oct 2016 - Ongoing
Guide: Prof. Andrew Smith *Computational Biology and Bioinformatics, USC*

The single largest investment of energy by cells is in the process of translation of messenger RNAs (mRNAs) into protein. Our understanding of the mechanisms of translational process remains limited. However, the need to decipher translational regulation has motivated the development of experimental approaches to profile the translational landscape. Ribo-seq is a deep-sequencing based technique that captures snapshots of ribosome protected fragments revealing the positions of the entire pool of ribosomes engaged in translation, hence, providing a global view of the translational landscape. My research involves systematically using public Ribo-seq data to answer some key biological questions involving translation regulation. This involves development of computational methods that can handle the diversity of public Ribo-seq studies. Through my work I am trying to to achieve the following key goals: 1) develop a computational method to identify actively-translating ORFs using Ribo-seq data (ribotricer) 2) develop a standardized database of uniformly processed public Ribo- seq studies 3) characterize the prevalence of upstream ORF (uORF) translation and its regulatory role across different species 5) develop a computational method to identify ribosomal stalling sites and characterize its role in regulating translation.

Source:

- ribotricer: https://github.com/smithlabcode/ribotricer
- riboraptor: https://github.com/saketkc/riboraptor

Poster: https://f1000research.com/posters/8-618

Higher order generalized SVD based alignment-free method for inferring orthology, Course Project Oct 2018 - Dec 2018
Instructor: Prof. Mahdi Soltanolkotabi *Dept. of Electrical Engineering, USC*

The advent of next generation sequencing has made a plethora of biological data available. Comparative analysis of gene of gene expression datasets from multiple species can be used to enhance our fundamental understanding of biological mechanisms. Current methods rely on sequence information to infer conservation of functionality or *orthology*, but this information is *incomplete*. Gene expression datasets provide a sequence independent paradigm that can help separate the *conserved* from the *non-conserved*. However, the high-dimensionality of these datasets raises a need of an appropriate framework to narrow down our search space to genes that behave similarly across multiple species. I explored higher order generalized singular value decomposition (HOGSVD) to analyze large scale gene expression datasets tabulated as matrices with varying number of rows corresponding to the genes and fixed number of columns corresponding to the tissues across different species to identify genes with similar function. This approach resulted in identifying genes showing significant overlap with the orthologous genes.

Poster: https://f1000research.com/posters/7-1853

Tools for Motif Conservation Analysis, PhD Project May 2015 - Feb 2016
Guide: Prof. Anton Valouev *Dept. of Preventive Medicine, Keck School of Medicine, USC*

Motifs predicted by motif discovery tools can often not be the 'true motifs' and can have significant p-value(or E-values) for even 'false motifs'. We hypothesized that a 'true motif' should exhibit high evolutionary conservation scores. MoCA makes use of PhyloP and Gerp scores to assess the conservation profile of motif bases.

We used MoCA to analyze ENCODE Chip-Seq datasets and found that the 'true motifs'(ones which have been validated experimentally) do exhibit high conservation scores and that these are statistically significant when compared to the scores of flanking regions or randomly sampled regions.

Source: https://github.com/saketkc/moca
Poster: https://doi.org/10.6084/m9.figshare.1565626.v5

Predicting protein coding boundaries using Deep Learning Nov 2017 - Ongoing
Course Project *USC*

I explored how recurrent neural networks (RNNs) can be used to predict protein coding domains in a gene. A word embedding approach along with bi-directional LSTMs gave promising results using the entire pool of protein coding genes in human achieving an overall accuracy of 0.67. This model when used to predict protein coding domains in a different species, mouse, achieved an overall accuracy of of 0.70 when tested on non-orthologous genes (where orthogonality implies a gene in mouse shares significant sequence from a human gene owing to descent from a common ancestor).

Preprint: https://doi.org/10.6084/m9.figshare.5902726.v1

Pattern Recognition in Clinical Data, Masters Thesis Apr 2013 - Jul 2013
Guide: Prof. Santosh Noronha *Dept. of Chemical Engineering, IIT Bombay*

Awarded Outstanding Thesis Award

Multiple methods exist for determining oncogenic 'driver' mutations. These tools often have non overlapping predictions and input format is tool specific.

We developed a Galaxy based toolbox to run such prediction tools in parallel with a standard input format. The end results are presented as an intuitive heatmap indicating mutations which are predicted to be drivers by a majority of the tools. Preprint: "GalDrive: Pipeline for comparative identification of driver mutations using the Galaxy framework"

In a separate project, we analyzed proteomics data from Glioblastoma patients and predicted a smaller set of marker genes. Paper: "Autoantibody Profiling of Glioma Serum Samples to Identify Biomarkers Using Human Proteome Arrays".

Automated Mining of Reaction Patterns May 2012 - Jul 2012, Jan 2014 - Mar 2014
Guide: Dr. Syed Asad Rahman *Dr. Dame Janet Thornton Lab, EMBL-EBI, Cambridge(UK)*

EC-BLAST is a novel tool to compare enzymes and map reactions. We used clustering based approaches to highlight misclassified enzymes in the established enzyme classification system(EC).

We developed a web-service that facilitated automated job submissions to back end clusters at EBI that led to significant reduction in job runtime.

Next Generation Sequencing, Supervised Learning Project Jul 2012 - Dec 2012
Guide: Prof. Santosh Noronha *Dept. of Chemical Engineering, IIT Bombay & ACTREC*

We developed automated pipelines using Python to analyze whole genome sequence data of cancer tumors. As part of the project, I also contributed BWA and samtools wrappers to Biopython, a Python based open source library for bioinformatics.

Scilab On Cloud May 2012 - Jul 2012
Guide: Prof. Kannan Moudagalya *Dept. of Chemical Engineering, IIT Bombay*

Scilab is an open source software for numerical computation and is primarily command line/GUI based. We developed a back-end that allowed running Scilab through browser much like the modern day IPython notebooks. This enabled accessing Scilab remotely, even on low configuration devices.

Presented at: IEEE Conference Cloud Computing and Big Data (CloudCom-Asia), 2013

Professional Experience

Google Summer of Code 2015 | Mixed Effect Models for `statsmodels` May 2015 - Jul 2015
Student Contract Developer

- `statsmodels` is a Python based library for statistical modeling
- Implemented IPython based notebooks illustrating varied applications of Mixed Effects Models
- Implemented likelihood ratio tests
- Progress Report: http://statsmodels-mlm-gsoc2015.blogspot.com

Google Summer of Code | BioJavascript Jul 2014 - Sep 2014
Student Contract Developer

- BioJavascript is an open source library to facilitate biological data visualization
- Developed 'Human Genetic Variation Viewer', a d3.js based component to visualize genetic variations from SNP databases
- Preprint: BioJS-HGV Viewer: Genetic Variation Visualizer

Google Summer of Code | Galaxy Project Jul 2013 - Sep 2013
Student Contract Developer

- Galaxy Project is an open source web-based platform used for reproducible bioinformatics analysis
- Implemented 'nested workflows' that allows users to run a workflow inside a workflow, obviating the need to replicate steps
- Added 'edit on the go' functionality to edit default parameters before runtime
- Progress Report: http://galaxy-gsoc2013.blogspot.com

Google Summer of Code | Connexions Project Jul 2012 - Sep 2012
Student Contract Developer

- Developed a Python module to allow embedding slide-shows in online notebooks
- Created functionality to add user defined quiz as an additional achievement
- Progress Report: http://oerpub.github.io/oerpub.rhaptoslabs.slideimporter/

Other Projects

Solutions to various examinations in Statistics June 2015 - Ongoing
Personal Project

- Royal Statistical Society Examinations Solutions:
  https://saketkc.github.io/rss-graduate-diploma-solutions/
- *Piddling Pertinent* - Solutions to several trivial problems in statistics:
  https://saketkc.github.io/pertinent-blog/
- *Screening Exam Solutions* - Solutions to screening examinations held at USC:
  https://saketkc.github.io/usc-math-505A-screening-solutions/;
  https://saketkc.github.io/usc-math-541A-screening-solutions/;
  https://saketkc.github.io/usc-math-541B-screening-solutions/

sklearn-hogsvd Jan 2019 - Ongoing
Personal Project

- Scikit-learn compatible python implementation of higher order generalized singular value decomposition. https://github.com/saketkc/sklearn-hogsvd

pysradb Nov 2018 - Ongoing
Personal Project

- Python package for interacting with SRAdb and downloading datasets from NCBI Sequence Read Archive (SRA). https://github.com/saketkc/pysradb

pyseqlogo Nov 2017 - Ongoing
Personal Project

- Python package to plot sequence logos https://github.com/saketkc/pyseqlogo

Image Analysis of Tuberculosis Samples Jan 2013 - Apr 2013
Supervised Learning Project, Collaborator: Hinduja Hospital, Mumbai

- Used image processing algorithms to detect probable cases of TB from sputum images
- Developed a user friendly GUI to aid histologists thus reducing the overall delay in analysis

Pratham, Student Satellite Program May 2010 - Oct 2010
India's First Students' Satellite Team, IIT Bombay

- Executed hardware testing of the On-board Computer system
- Implemented signal processing pipeline for communications subsystem

Teaching Experience

- Teaching Assistant, Computer Programming and Utilization — Fall 2011
- Teaching Assistant, Artificial Intelligence in Process Engineering — Fall 2013
- Teaching Assistant, How the Body Works — Fall 2017
- Teaching Assistant, How the Body Works — Spring 2018

Positions of Responsibility

Web Manager, UG Academic Council Jul 2012 - Apr 2013

- Initiated a number of web portals, thus improving online accessibility of academic resources
- Awarded **Institute Organizational Award**

TechniC, Core Group Member Jul 2010 - Apr 2011

- Organized institute wide technical events; mentored students

Standardized Test Scores

- GRE: Quantitative: 170/170 Verbal: 153/170 Analytical Writing: 3.5/6
- TOEFL: Reading: 29/30 Listening: 28/30 Speaking: 24/30 Writing:28/30 Total: 109/120

COURSEWORK AT USC

- Machine Learning
- Deep Learning
- Wavelets
- Time Series Analysis
- Mathematical Statistics
- Applied Probability
- Numerical Analysis
- Analysis of Algorithms
- Introduction to Computational Biology
- Biostatistics
- Molecular Biology
- Seminar in Statistical Consulting
- Methods of Statistical Inference

RELEVANT COURSEWORK THOUGH COURSERA (VERIFIED)

- Machine Learning
- Mathematical Biostatistics Boot Camp 1
- Applied Logistic Regression
- Reproducible Research
- Case-Based Introduction to Biostatistics
- Linear Algebra
- Exploratory Data Analysis