

Saket Choudhary

CONTACT INFORMATION	2677 Ellendale Place Apartment #215 Los Angeles California 90007, USA	Email: saketkc@gmail.com Homepage: https://saketkc.github.io [Google Scholar] [CrossValidated] [Github]
EDUCATION	University of Southern California , Los Angeles, USA <i>Ph.D. Candidate</i> , Computational Biology and Bioinformatics	[2014 - Ongoing]
	University of Southern California , Los Angeles, USA <i>Masters in Computer Science (Data Science)</i> Department of Computer Science	[2018 - 2019]
	University of Southern California [USC] , Los Angeles, USA <i>Masters in Statistics</i> Department of Mathematics	[2016 - 2018]
	Royal Statistical Society , London, England <i>4/5 Modules in Graduate Diploma in Statistics</i>	[2016 - 2017]
	Indian Institute of Technology Bombay [IITB] , Mumbai, India <i>Bachelor of Technology, Master of Technology</i> Department of Chemical Engineering Masters Thesis: Pattern Recognition in Clinical Data	[2009 - 2014]
HONORS AND AWARDS	<ul style="list-style-type: none">- Open Bioinformatics Foundation Travel Fellowship- International Society for Computational Biology Travel Fellowship- Outstanding Teaching Assistant, USC- Provost Fellowship, awarded to outstanding incoming PhD students at USC- Gandhian Young Technological Innovation Award by Indian Institute of Management Ahmedabad, for designing a low-cost water impurity detection device- Institute Technical Special Mention for three consecutive years at IITB- Undergraduate Research Award, IITB- Kishor Vaignyanik Protsahan Yojana (KVPY) Fellowship by Department of Science and Technology (DST), Government of India- Homi Bhabha Young Scientists' Gold Medal	[2019] [2019] [2019] [2014] [2013] [2010-12] [2012] [2007] [2005]
OLYMPIADS	<ul style="list-style-type: none">- Top 6 to be selected for Indian National Mathematics Olympiad (INMO), selection level exam for International Mathematical Olympiad(IMO)- Top 30 in Regional Mathematics Olympiad (RMO)- Top 250 in Indian National Physics Olympiad (INPhO)- Top 300 in Indian National Astronomy Olympiad (INAO)	[2008] [2009] [2009] [2009]

1. **Saket Choudhary**[†], Wenzheng Li[†], and Andrew D. Smith *Accurate detection of short and long active ORFs using Ribo-seq data*. Bioinformatics (2019). [\[Online\]](#) | [\[Preprint\]](#)
2. **Saket Choudhary**[†], Suzanne C. Burns[†], Hoda Mirsafian, Wenzheng Li, Dat T. Vo, Mei Qiao, Andrew D. Smith, and Luiz O. Penalva *Integrated analyses of early responses to radiation in glioblastoma identify new alterations in RNA processing and candidate target genes to improve treatment outcomes*. [\[Preprint\]](#)
3. **Saket Choudhary**. *pysradb: A Python Package to Query next-Generation Sequencing Meta-data and Data from NCBI Sequence Read Archive*. F1000Research, vol. 8, F1000 (Faculty of 1000 Ltd), Apr. 2019, p. 532. [\[Online\]](#)
4. 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*. eLife (2019). [\[Online\]](#)
5. **Saket Choudhary**. *Golden-Thompson via pinching inequality*. arXiv preprint arXiv:1811.00544 (2018).[\[Preprint\]](#)
6. 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. [\[Online\]](#)
7. **Saket Choudhary**, Leyla Garcia, Andrew Nightingale, and Maria-Jesus Martin. *BioJS-HGV Viewer: Genetic Variation Visualizer*. bioRxiv (2015): 032573. [\[Preprint\]](#)

As an Undergraduate Student:

8. 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), 13895. [\[Online\]](#)
9. 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. [\[Preprint\]](#)
10. **Saket Choudhary**, and Santosh B. Noronha. *GalDrive: Pipeline for comparative identification of driver mutations using the Galaxy framework*. bioRxiv (2014). [\[Preprint\]](#)
11. **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. [\[Online\]](#)
12. 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. [\[Online\]](#)
13. 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. *Accurate detection of active short and long ORFs*, USC-QCB Retreat 2019 [Talk]
2. *BioJS-HGV: Human Genetic Variation Viewer*, USC-UCLA Joint Symposium 2015 [Talk]
3. *MoCA: A tool for Motif Conservation Analysis*, USC-MCB Retreat 2015 [Poster]
4. *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**
*Guide: Dr. Radhakrishna Bettadapura*May 2018 - July 2018
Strand Life Sciences, India

I developed a core library for applied machine learning on histopathology images. Our method used a combination of unsupervised segmentation and random forests. Though its performance was not at par with other deep learning approaches that have been applied to this problem in the literature, it serves as a good proof of concept.

[Report](#) | [Code](#)**Computational methods for deciphering translational regulation, Ph.D. Project** Oct 2016 - Ongoing*Guide: Dr. Andrew Smith**Computational Biology and Bioinformatics, USC*

The most significant investment of energy by cells is in the process of translation of messenger RNAs into protein. Our understanding of the mechanisms of the 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 the development of computational methods that can handle the diversity of public Ribo-seq studies. Through my work I am trying to achieve the following key goals: 1) develop a computational method to identify actively-translating ORFs using Ribo-seq data (ribotracer) 2) develop a standardized database of uniformly processed public Ribo-seq studies 3) characterize the prevalence of upstream ORF 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.

Code:

- ribotracer: <https://github.com/smithlabcode/ribotracer>
- riboraptor: <https://github.com/saketkc/riboraptor>

[Paper](#) | [Poster](#) | [Database](#)**Higher-order generalized SVD based alignment-free method for inferring orthology, Course Project**

Oct 2018 - Dec 2018

*Instructor: Dr. Mahdi Soltanolkotabi**Dept. of Electrical Engineering, USC*

Comparative analysis 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 motivates development 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 a varying number of rows corresponding to the genes and a 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](#) | [Preprint](#) | [Code](#)

Tools for Motif Conservation Analysis, Ph.D. Project

May 2015 - Feb 2016

Guide: Dr. 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.

[Poster](#) | [Code](#)

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 humans, 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 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](#) | [Code](#)

Pattern Recognition in Clinical Data, Masters Thesis

Apr 2013 - Jul 2013

Guide: Dr. Santosh Noronha

Dept. of Chemical Engineering, IIT Bombay

Awarded Graduate Research Award

Multiple methods exist for determining oncogenic 'driver' mutations. These tools often have non-overlapping predictions, and the 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 that are predicted to be drivers by a majority of the tools.

[Preprint](#)

In a separate project, we analyzed proteomics data from Glioblastoma patients and predicted a smaller set of marker genes.

[Paper](#)

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 a significant reduction in job runtime.

Next Generation Sequencing, Supervised Learning Project

Jul 2012 - Dec 2012

Guide: Dr. 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: Dr. 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 a web-browser, much like the modern-day IPython notebooks, thus enabling accessing Scilab remotely, even on low configuration devices.

Paper**PROFESSIONAL
EXPERIENCE****Google Summer of Code 2015 | Mixed Effect Models for statsmodels**
Student Contract Developer

May 2015 - Jul 2015

- 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 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
- Progress Report: <http://galaxy-gsoc2013.blogspot.com>

OTHER PROJECTS	Solutions to various examinations in Statistics <i>Personal Project</i> June 2015 - Ongoing
	<ul style="list-style-type: none"> - Royal Statistical Society Examinations Solutions: https://saketkc.github.io/rss-graduate-diploma-solutions/ - <i>Piddling Pertinent</i> - Solutions to several trivial problems in statistics: https://saketkc.github.io - <i>Screening Exam Solutions</i> - 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 <i>Personal Project</i> Jan 2019 - Ongoing
	<ul style="list-style-type: none"> - Scikit-learn compatible python implementation of higher order generalized singular value decomposition. https://github.com/saketkc/sklearn-hogsvd
	pysradb <i>Personal Project</i> Nov 2018 - Ongoing
	<ul style="list-style-type: none"> - Python package for interacting with SRADB and downloading datasets from NCBI Sequence Read Archive (SRA). https://github.com/saketkc/pysradb
	pyseqlogo <i>Personal Project</i> Nov 2017 - Ongoing
	<ul style="list-style-type: none"> - Python package to plot sequence logos https://github.com/saketkc/pyseqlogo
	Image Analysis of Tuberculosis Samples <i>Supervised Learning Project, Collaborator: Hinduja Hospital, Mumbai</i> Jan 2013 - Apr 2013
	<ul style="list-style-type: none"> - 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 <i>India's First Students' Satellite Team, IIT Bombay</i> May 2010 - Oct 2010
	<ul style="list-style-type: none"> - Executed hardware testing of the On-board Computer system - Implemented signal processing pipeline for communications subsystem
TEACHING EXPERIENCE	<ul style="list-style-type: none"> - Teaching Assistant, Computer Programming and Utilization Fall 2011 - Teaching Assistant, Artificial Intelligence in Process Engineering Fall 2013 - Teaching Assistant, How the Body Works Fall 2016 - Spring 2019 - Teaching Assistant, Statistics for the Biological Sciences Fall 2019
STANDARDIZED TEST SCORES	<ul style="list-style-type: none"> - 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