# Hirak Sarkar

Stony Brook – NY 11790

☏ 631-520-8131 • ✉ hsarkar@cs.stonybrook.edu • www.hiraksarkar.com

## OBJECTIVE

I am interested in designing and deploying efficient workflows to process raw sequences (such as RNA-seq) and alongside applying machine learning techniques to analyze and extract information from heterogeneous, large-scale public datasets (such as SRA).

## EDUCATION

**Stony Brook University (SBU)** — **Stony Brook,NY**
*Ph.D in Computer Science (3.99/4)* — *2014-2019(exp)*

**Indian Statistical Institute** — **Calcutta, India**
*M.Tech in Computer Science (1st class Hons.)* — *2011-2013*

**West Bengal University of Technology** — **Calcutta, India**
**(Kalyani Govt. Engg. College)** — *2007-2011*
*B.Tech in Computer Science (8.88/10)*

## EXPERIENCE

COMBINE-Lab (Stony Brook University)
*Research Assistant (https://github.com/COMBINE-lab)* — *Jan 2015-Present*

- Application of machine learning methods for publicly available massive genomic databases. (Python, sklearn, C++)
  - Available public databases are full of mislabeled samples which makes the downstream analysis extremely difficult. To mitigate the difficulty, we aim to build a workflow that can automatically learn the metadata features from a set of well-annotated databases. The project involves writing the modules for processing, cleaning and designing suitable learning algorithms.
- Development of graph based k-mer mapper, Pufferfish (C++)
  - Genome sequences (string in the order of gigabytes) are difficult to index and search in limited memory. Building a fast query efficient and memory efficient genome index is a challenging task. We used a minimum perfect hash based, rank-select algorithm to store the de-Bruijn graph based genome index which enables fast query of nucleotide sequences with manageable memory overhead. [bioRxiv'17]
- Developed an intermediate solution for accurate mapping of read sequences. (C++)
  - Alignments involve rigorous dynamic programming and therefore are costly. Mapping of reads are fast yet not accurate, to carry best of the both worlds we developed a selective-alignment based algorithm, implemented in C++, which achieved quantification accuracy comparable with complete aligners (Bowtie2, STAR), yet get to do so with almost half the time requirement. [bioRxiv'17]
- Development of compression algorithm for raw RNA-seq reads, Quark (C++)
  - We developed a semi-reference based compression scheme, which achieves state-of-the-art compression ratio. In this scheme the reference is needed while compressing the reads although it is not required at the decompression end, therefore enabling the compressed format completely self-sufficient. [*Bioinformatics'17*]
- Developed alignment free methods for sequence reads. (C++)
  - We developed *RapMap,* an ultra fast mapper, which builds an index over the transcriptomic sequence by using a suffix array and hash table. While comparing with alignment-based quantification tools, it achieved similar results and do so in substantially less time. [*ISMB'16*]
- Graph based clustering for novel organisms. (C++, Python)
  - We proposed equivalence class graph, an intermediate representation of isoform level expression and able to cluster isoforms in a *de-novo* setting.

*Collaboration with Siepel-Lab (Cold Spring Harbor Lab)* — *June 2016-Aug 2016*

- Developed probabilistic graphical model for inferring transcription rate from multi-assay dataset.
  - With the rise of different assays for the same biological specimen, it is possible to look into the cellular processes

at multiple resolution. We looked into the GRO-seq (a protocol developed in Cornell) and RNA-seq read datasets from the same sample and designed a probabilistic graphical model to estimate regulation rate and degradation rate.

## PUBLICATION

*Conferences and Journals*

- A space and time-efficient index for the compacted colored de Bruijn graph, by Fatemeh Almodaresi*, **Hirak Sarkar***, Rob Patro. [bioRxiv'17]
- Towards selective-alignment: Bridging the accuracy gap between alignment-based and alignment-free transcript quantification, by **Hirak Sarkar***, Mohsen Zakeri*, Laraib Malik, Rob Patro. [*Submitted to Bioinformatics'17*, bioRxiv'17]
- Quark enables semi-reference-based compression of RNA-seq data by **Hirak Sarkar** and Rob Patro. [*Bioinformatics'17*, bioRxiv'16].
- Fast, Lightweight Clustering of de novo Transcriptomes using Fragment Equivalence Classes by Avi Srivastava*, **Hirak Sarkar***, Laraib Malik and Rob Patro. [*RECOMB-seq'16*, arXiv'16]
- RapMap: A Rapid, Sensitive and Accurate Tool for Mapping RNA-seq Reads to Transcriptomes by Avi Srivastava, **Hirak Sarkar**, Nitish Gupta and Rob Patro. [*ISMB'16* (acceptance rate: 17%), *Bioinformatics'16*, bioRxiv'16]
- Voronoi Game on Graphs (Extended version) by S. Bandyapadhyay, A. Banik, S. Das and **H. Sarkar**. [Journal of Theoretical Computer Science- *TCS'15*, *WALCOM'13*].

*Posters:*

- Pufferfish: A fast graph-based indexing and query strategy for large genomic sequences by Fatemeh Almodaresi*, **Hirak Sarkar***, Rob Patro, Poster presented in *WABI'17*.
- Joint probabilistic model for multiple steps of gene regulation by **Hirak Sarkar**, Yi-Fei Huang and Adam Siepel, Poster presented in *BioData'16*.

## AWARDS

- Awarded Research Assistantship, SBU. (2016-present)
- Awarded Special CS Chair Fellowship. ($10K) from *SBU*
- Awarded Post-Graduate Scholarship from *Govt. Of India.*
- Awarded NUS Research Scholarship, NUS. (Jan'14-June'14)
- Received First Prize for Software Competition (IEM), Calcutta. (2011)

## RELEVANT COURSES

- Artificial Intelligence, Computational Biology, Analysis of Algorithms, Fundamental of Networks. (at Stony Brook University)
- Machine Learning & Pattern Recognition, Image Processing, Stochastic Processes, Optimization Algorithms, Graph Theory. (at Indian Statistical Institute)

## SKILLS

- Programming: *C++, Python, R*
- Data analysis: *Jupyter, Pandas*
- Machine Learning Tools: *sklearn, tensorflow*

## REFERENCE

- Prof Robert Patro, (Assistant Professor, Department of Computer Science, Stony Brook University)
- Prof Adam Siepel, (Professor, Watson School of Biological Sciences, Chair, Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory)