SUBRATA SAHA

Associate Research Scientist ⋄ Irving Medical Center
Columbia University, NY 10032
www.subrata-saha.com ⋄ (860) 208-0449 ⋄ subrata.saha@uconn.edu

EDUCATION

University of Connecticut, Storrs, CT, United States *May 2017*
(Ranked 8^{th} in the world for computational biology and bioinformatics research[1])
Ph.D. in Computer Science and Engineering
(2× Research Excellence Awards; GPA: 3.98/4.0)
Dissertation Title: *Novel Algorithms for Big Data Analytics*
Advisor: Professor Sanguthevar Rajasekaran

Bangladesh University of Engineering and Technology, Dhaka, Bangladesh *March 2009*
B.Sc. in Computer Science and Engineering
Thesis: *Scalable Address Autoconfiguration in Mobile Ad Hoc Networks*
Advisor: Professor A.K.M. Ashikur Rahman

RESEARCH AND TEACHING INTERESTS

Algorithms	Data Structures	Data Mining	Machine Learning
Bioinformatics	Biomedical Informatics	Computational Biology	Computational Genomics

EMPLOYMENT

Irving Medical Center, Columbia University, NY November 2020 - Present
Associate Research Scientist

- Developed novel graph theoretic and machine learning algorithms, study various data analytics methods in the fields of cancer biology, bioinformatics, and computational genomics
- Worked on machine learning and graph theoretic methods, statistics, rectal neuroendocrine tumors (RNETs), TF-target prediction, genome-wide association study, gene-based collapsing analysis, ethnicity prediction, modeling protein activity, among others

IBM T.J. Watson Research Center, Yorktown Heights, NY June 2017 - October 2020
Postdoctoral Research Scientist

- Developed novel algorithms and statistical inference techniques in the fields of computational genomics, big data mining, and bioinformatics in computer science
- Worked on machine learning algorithms, biostatistics, data reduction, edge sparsification, biclustering, genomics of Alzheimer's disease and Parkinson's disease, biological networks (e.g. pathway, protein-protein interaction, and gene co-expression), and haplotype phasing and reconstruction

University of Connecticut, Storrs, CT August 2011 - May 2017
Graduate Assistant

- Developed efficient algorithms and data structures in the fields of big data mining and bioinformatics
- Worked on a wide array of big data problems, e.g. genome-wide association studies, biological sequence classification, correction, compression and assembly, hierarchical and spectral clusterings, feature selection, frequent itemset mining, and similar pairs of points detection

[1] CSRankings: Computer Science Rankings [`http://csrankings.org/#/index?bio&world`]

- Taught graduate and undergraduate courses (e.g. algorithms and complexity, big data analytics, etc.), designed problems, prepared solutions, evaluated assignments, and corresponded to students

Bentley Systems, Inc., Watertown, CT May 2015 - August 2015
Research Intern

- Designed and implemented novel techniques to calibrate finite element models to monitor current health of civil structures (e.g. buildings and bridges) by employing machine learning algorithms
- Improved `C++` codebase and automated dataset generation techniques from finite element models

TigerIT Bangladesh Ltd., Dhaka, Bangladesh June 2009 - July 2011
Senior Software Engineer

- Developed software products currently being served for millions of people worth millions of dollars
- Worked on *Voters Registration System for Bangladesh Election Commission*, *Machine Readable Passport System for Government of Nepal*, and *Biometric Identification System for Banking and Security Services*

SELECTED PUBLICATIONS

Referred Journal Publications

1. **S. Saha**, J. Johnson, S. Pal, G. M. Weinstock and S. Rajasekaran: **MSC: a metagenomic sequence classification algorithm.** *Bioinformatics*, 35(17):2,932–2,940, 2019
   • Impact Factor = 6.937. **Ranked** 1^{st} in Bioinformatics & Computational Biology by Google Scholar
2. **S. Saha** and S. Rajasekaran: **NRGC: a novel referential genome compression algorithm.** *Bioinformatics*, 32(22):3,405-3,412, 2016
   • Impact Factor = 6.937. **Ranked** 1^{st} in Bioinformatics & Computational Biology by Google Scholar
3. **S. Saha** and S. Rajasekaran: **ERGC: an efficient referential genome compression algorithm.** *Bioinformatics*, 31(21), 3,468-3,475, 2015
   • Impact Factor = 6.937. **Ranked** 1^{st} in Bioinformatics & Computational Biology by Google Scholar
4. **S. Saha**, A. Soliman and S. Rajasekaran: **A robust and stable gene selection algorithm based on graph theory and machine learning.** *BMC Human Genomics*, 31(21), 3,468-3,475, 2021 [Impact Factor = 4.860]
5. D. He, **S. Saha**, R. Finkers and L. Parida: **Efficient algorithms for polyploid haplotype phasing.** *BMC Genomics*, 19 (Suppl 2):110, 2018
   • Impact Factor = 4.478. **Ranked** 9^{th} in Genetics & Genomics by Google Scholar
6. **S. Saha** and S. Rajasekaran: **Efficient and scalable scaffolding using optical restriction maps.** *BMC Genomics*, 15 (Suppl 5):S5, 2014
   • Impact Factor = 4.478. **Ranked** 9^{th} in Genetics & Genomics by Google Scholar
7. **S. Saha** and S. Rajasekaran: **EC: an efficient error correction algorithm for short reads.** *BMC Bioinformatics*, 16 (Suppl 17):S2, 2015
   • Impact Factor = 3.629. **Ranked** 4^{th} in Bioinformatics & Computational Biology by Google Scholar
8. **S. Saha**, S. Rajasekaran, J. Bi and S. Pathak: **Efficient techniques for genotype-phenotype correlational analysis.** *BMC Medical Informatics and Decision Making (BMC MIDM)*, 13:41, 2014
   • Impact Factor = 3.394. **Ranked** 8^{th} in Medical Informatics by Google Scholar
9. **S. Saha**, A. Soliman and S. Rajasekaran: **A novel pathway network analytics method based on graph theory.** *Journal of Computational Biology*, https://doi.org/10.1089/cmb.2021.0257, 2021 [Impact Factor = 1.479]
10. S.R. Hussain, **S. Saha** and A. Rahman: **SAAMAN: Scalable Address Autoconfiguration in Mobile Ad Hoc Networks.** *Journal of Network and Systems Management (JNSM)*, 19(3):394-426, 2011 [Impact Factor = 2.139]

11. **S. Saha**, S. Rajasekaran and R. Ramprasad: **Novel Randomized Feature Selection Algorithms.** *International Journal of Foundations of Computer Science (IJFCS)*, 26(3):321-341, 2015 [IF = 0.523]

Referred Conference Publications

12. **S. Saha**, A. Soliman and S. Rajasekaran: **MSPP: A Highly Efficient and Scalable Algorithm for Mining Similar Pairs of Points.** *International Conference on Advanced Data Mining and Applications (ADMA)*, pp. 30-37, 2020
13. **S. Saha**, Z. Wang and S. Rajasekaran: **HMSC: a Hybrid Metagenomic Sequence Classification Algorithm.** 11^{th} *ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB)*, 2020
14. S. Rajasekaran*, **S. Saha*** and X. Cai: **Novel Exact and Approximate Algorithms for the Closest Pair Problem.** *IEEE International Conference on Data Mining (IEEE ICDM)*, pp. 1,045-1,050, 2017 [*1^{st} **author**]
15. S. Dey, A. Bose, **S. Saha**, P. Chakraborty, M. Ghalwash, A.G. Saenz, F. Utro, K. Ng, J. Hu, L. Parida and D. Sow: **Impact of Clinical and Genomic Factors on SARS-CoV2 Disease Severity.** To appear in *American Medical Informatics Association (AMIA) 2021 Annual Symposium*, 2021
16. S. Rajasekaran and **S. Saha**: **Efficient Algorithms for the Three Locus Problem in Genome-wide Association Study.** *IEEE International Conference on Data Mining (IEEE ICDM)*, pp. 1,155-1,160, 2016
17. S. Rajasekaran and **S. Saha**: **Efficient Algorithms for the Two Locus Problem in Genome-wide Association Study.** 25^{th} *ACM International Conference on Information and Knowledge Management (ACM CIKM)*, pp. 2,305-2,310, 2016
18. **S. Saha** and S. Rajasekaran: **POMP: a powerful splice mapper for RNA-seq reads.** 7^{th} *ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB)*, pp. 414-421, 2016
19. **S. Saha** and S. Rajasekaran: **REFECT: a novel paradigm for correcting short reads.** 6^{th} *ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB)*, pp. 57-66, 2015
20. **S. Saha** and S. Rajasekaran: **NRRC: A Non-Referential Reads Compression Algorithm.** 11^{th} *International Symposium on Bioinformatics Research and Applications (ISBRA)*, pp. 297-308, 2015
21. S. Rajasekaran and **S. Saha**: **A Novel Deterministic Sampling Technique to Speedup Clustering Algorithms.** 9^{th} *International Conference on Advanced Data Mining and Applications (ADMA)*, pp. 34-46, 2014
22. **S. Saha**, S. Rajasekaran and R. Ramprasad: **A Novel Randomized Feature Selection Algorithm.** 9^{th} *International Conference on Data Mining (DMIN)*, 2013
23. **S. Saha**, S.R. Hussain and A. Rahman: **RBP: Reliable Broadcasting Protocol in Large Scale Mobile Ad Hoc Networks.** 24^{th} *IEEE International Conference on Advanced Information Networking and Applications (IEEE AINA)*, pp. 526-532, 2010
24. S.R. Hussain, **S. Saha** and A. Rahman: **An Efficient and Scalable Address Autoconfiguration in Mobile Ad Hoc Networks.** 8^{th} *International Conference on Ad-Hoc, Mobile and Wireless Networks (ADHOC-NOW)*, pp. 152-165, 2009

FORTHCOMING PUBLICATIONS

1. **S. Saha** et *al.*: **Genetic landscape of rectal neuroendocrine tumors.** *Completed manuscript*, 2021
2. **S. Saha**, H.N. Singh, A. Soliman and S. Rajasekaran: **A novel computational methodology for GWAS multi-locus analysis based on graph theory and machine learning.** *medRxiv*, doi: https://doi.org/10.1101/2021.10.22.21265388, 2021

3. **S. Saha**, A. Soliman and S. Rajasekaran: **Novel graph theoretic biological pathway network analytics methods for analyzing and discovering Alzheimer's disease related gene.** *bioRxiv*, doi: https://doi.org/10.1101/2021.10.19.465019, 2021
4. **S. Saha**, A. Guzman-Saenz, A. Bose, F. Utro, D.E. Platt, L. Parida: **RubricOE: a learning framework for genetic epidemiology.** *medRxiv*, doi: https://doi.org/10.1101/2021.03.09.21253105, 2021
5. **S. Saha**, Z. Wang and S. Rajasekaran: **A novel algorithm to accurately classify metagenomic sequences.** *bioRxiv*, doi: https://doi.org/10.1101/2020.10.01.321067, 2020

PATENT

Cognitive Identification of Pathogenic Pathways. A novel computational method to identify pathogenic pathways of a disease by employing a novel configuration of machine learning, redescription, and computational homology techniques (application filed; reference # P201803065US01)

HONORS AND AWARDS

Research and academic awards

- **IBM manager's choice award (2018).** Awarded for excellent performance, IBM Research
- **Topper, graduate fellowship prize (2015-2016).** Awarded twice the top prize for research excellence, Computer Science and Engineering Department, University of Connecticut, Storrs
- **Doctoral dissertation fellowship award (2016).** Awarded for doctoral dissertation, Graduate School, University of Connecticut, Storrs
- **Pre-doctoral fellowship award (2014).** Awarded for high quality research works, Computer Science and Engineering Department, University of Connecticut, Storrs
- **National merit scholarship (2004-2009).** Awarded for exceptional academic achievement in the public exam of 12^{th} grade, Government of the People's Republic of Bangladesh

Student travel awards

- **NSF travel awards (2014-2016).** Awarded for attending and presenting research articles in: • 7^{th} ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB) • 25^{th} ACM International Conference on Information and Knowledge Management (CIKM) • 6^{th} ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB) • 4^{th} International Conference on Computational Advances in Bio and medical Sciences (ICCABS)
- **Student travel award (2016).** Awarded for attending and presenting a research article, 25^{th} ACM International Conference on Information and Knowledge Management (CIKM)

RESEARCH EXPERIENCE

Post-PhD Research Experience

- **Genetic landscape of rectal neuroendocrine tumors (RNETs).** Proved that there is strong germline predispositions in RNETs with experimental confirmation. In addition, we demonstrated that the identified germline variants together have perfect prediction power. Consequently, those variants can be used as an early screening tool to assess the future risk of RNET-susceptibility of an individual
- **Machine learning based TF-target prediction.** Developed a machine learning algorithm to accurately predict the transcription factors (TFs). The TF-target networks were then reconstructed based on normalized mutual information based graph-centric approach
- **Machine learning based rare variant collapsing analysis.** Developed a machine learning algorithm that can meaningfully collapse a set of rare variants within their predefined genomic regions and can identify clusters of genomic regions highly correlated with the underlying genetic disease

- **Scalable biomarker identification techniques.** Developed novel statistical and machine learning algorithms that can meaningfully reduce very high dimensional feature space, and correctly identify genetic markers (e.g. SNPs and genes) responsible for various genetic diseases, such as Alzheimer's disease (AD), Parkinson's disease (PD), Astigmatism, Coronavirus disease 2019, among others
- **Haplotype phasing and reconstruction.** Developed an efficient method (first of this kind) to correctly construct contiguous sequence of haplotypes (a set of single nucleotide polymorphisms) from a set of disconnected haplotypes by cleverly utilizing shared information across individual subsequences
- **Pathway and PPI network analysis.** Developed efficient and novel graph theoretic methods that can accurately analyze and extract hidden information from complex network of biological pathways along with can discover novel disease-specific genes by exploiting protein-protein interactions (PPI) network

Doctoral Research Experience

- **Metagenomic sequence classification.** Devised highly accurate and fast metagenomic sequence classification algorithms to detect microbes and their abundances from metagenomic sequences with a very high accuracy, less memory consumption, and execution time; outperform the state-of-the-art algorithms, such as MetaPhlAn2, RTG, Kraken, and CLARK
- **Genome-wide splicing events detection.** Invented a multi-core self-learning algorithm to accurately detect genome-wide RNA splicing events (i.e. intron-exon or exon-intron boundaries); it is 2× as fast as and 2.9% more accurate compared to the existing state-of-the-art algorithms (e.g. TopHat2)
- **Time and space efficient biological data compression.** Employed greedy placement and novel hashing schemes to develop time and space efficient algorithms for compressing high volume of biological sequence data (e.g. genomes and reads); offers 85% improvement in compression ratios while being 2× as fast as existing algorithms namely GDC, iDoComp, and ERGC
- **Effective error correction algorithms.** k-mer spectrum-based algorithms have been devised to effectively correct substitution errors in short biological sequence reads; up to 1.5× faster and 5% more accurate compared to the best-known algorithms (e.g. Racer, Coral, Musket, and BLESS)
- **Genome-wide association study.** Developed highly efficient and scalable algorithms to solve 2-locus and 3-locus problems by employing novel random sampling and hashing schemes; 177× faster with higher accuracy compared to the currently best-known algorithms
- **Genotype-phenotype correlational analysis.** Employed random projection and support vector machines (SVMs) to develop classification algorithms; up to 12× faster and 6% more accurate compared to widely popular multifactor dimensionality reduction (MDR) and principal component analysis (PCA)
- **Efficient scaffolding algorithms.** Introduced a series of novel algorithms for correct placement of contigs (overlapping biological sequence reads) that exploit optical restriction maps (ORMs); up to 30% higher in accuracy and about 3 order of magnitude faster than that of best performing algorithm
- **Closest pair detection algorithms.** Studied the problem of finding the exact and approximate closest pair from millions of points in a very high dimensional space using novel projection and transformation methodology; best performing algorithms in this domain
- **A scalable clustering algorithm.** Developed deterministic sampling techniques to reduce quadratic time and space complexity of exact hierarchical clustering algorithms; 200× faster and 2% more accurate than the exact hierarchical clustering algorithms
- **Novel feature selection algorithms.** Invented a set of randomized search methods that are generic in nature and can be applied for any learning algorithms; accuracy increased by 3% over kernel ridge regression (KRR) and support vector machine (SVM)

Undergraduate Research Experience

- **Reliable broadcasting protocol in mobile ad hoc networks.** This research investigated a novel broadcasting protocol to send a message in the entire network where every node is guaranteed to receive a copy of it without flooding the message
- **Scalable address autoconfiguration in mobile ad hoc networks.** This research focused on automatically configuring a mobile ad hoc network by assigning unique IP addresses to all the nodes with a very low overhead and a minimal cost using the quad-tree architecture

GRANT WRITING EXPERIENCE

- Led the preparation of "Novel Precision Medicine Approach for Rectal Neuroendocrine Tumors" and was submitted to Neuroendocrine Tumor Research Foundation (NETRF), 2021
- Led the preparation of "Novel Computational Techniques for Correlating Microbial Data and Genotypes with Phenotypes" and was submitted to National Institutes of Health (NIH), 2017
- Aided in preparation of "Novel Out-of-core and Parallel Algorithms for Processing Biological Big Data" and was awarded $1.2M by National Science Foundation (NSF) in 2014
- Aided in preparing multiple research grant proposals focused on scalable and robust machine learning methods for genetic epidemiology in IBM Research, 2017-2020

TEACHING AND ADVISING EXPERIENCES

- **Student teaching.** Served as a teaching assistant for a varied number of courses at both undergraduate and graduate course of studies, including Algorithms and Complexity (CSE 3500), Advanced Sequential and Parallel Algorithms (CSE 5500), and Big Data Analytics (CSE 4502/5717)
- **Student mentoring.** Mentored five graduate students; helped them designing and analyzing algorithms, finding relevant research articles to study, and writing manuscripts for publications

TECHNICAL EXPERIENCE

- **Multi-tier web applications (2010-2011).** Designed and implemented business logic and web services to efficiently store, update, extract, and analyze demographic and biometric data (e.g. fingerprints, signatures, and images) of `100+` million of people using Java Enterprise Edition (`Java EE`)
- **Distributed printing and reporting Systems (2010-2011).** Designed and developed efficient and scalable distributed printing and reporting systems using Internet Printing Protocol (IPP) and `Java EE`
- **Relational database management systems (2009-2011).** Architected and implemented relational database schema for Oracle 9i and MySQL making database operations simple and fast
- **Biometric identification systems (2009-2011).** Designed and developed user interfaces to capture and authenticate biometric data (e.g. fingerprints and signatures) over the Internet using `C++` and `Java`

PROFESSIONAL SERVICE

Conference Organizing and Technical Committee

- **Program committee (PC) member (2017-2019).** Served three times as a PC member in the IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS)
- **Technical program committee (TPC) member (2014-2015).** Served two times as a TPC member in the IEEE Symposium on Computers and Communications (ISCC)
- **Local arrangements chair (2013-2014).** Served two times as a local chair in the IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS)

Conference and Journal Review Committee

- BMC Bioinformatics • BMC Supplements • ACM Computing Surveys • IEEE Transactions on Computers • IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) • IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS) • European Conference on Computational Biology (ECCB) • International Conference on Language and Automata Theory and Application (LATA) • IEEE Symposium on Computers and Communications (ISCC) • Engineering in Medicine and Biology Conference (EMBC) • International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB) • Life Science Alliance (LSA)

RESEARCH TALKS

1. **Novel algorithms for big data analysis (2017).** IBM Research, Yorktown Heights, NY
2. **Novel algorithms for biological big data analysis (2017).** Broad Institute, Cambridge, MA
3. **2-locus problem in genome-wide association study (2016).** 25^{th} ACM International Conference on Information and Knowledge Management (CIKM), Indianapolis, IN
4. **A self-learning algorithm to discover genome-wide splicing events (2016).** 7^{th} ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB), Seattle, WA
5. **2-locus and 3-locus problems in genome-wide association study (2016).** 7^{th} ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB), Seattle, WA
6. **A hybrid error correction algorithm for short reads (2015).** 6^{th} ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB), Atlanta, GA
7. **Referential genome compression algorithms (2015).** 6^{th} ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB), Atlanta, GA
8. **A *de novo* biological sequence reads compression algorithm (2015).** 11^{th} International Symposium on Bioinformatics Research and Applications (ISBRA), Norfolk, VA
9. **Error correction and sequence compression algorithms (2014).** IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS), Miami, FA
10. **A novel randomized feature selection algorithm (2013).** 9^{th} International Conference on Data Mining (DMIN), Las Vegas, NV

TECHNICAL STRENGTHS

• **Programming languages**	`C`, `C++`, Java, `C#`
• **Modeling and analysis**	Python, R, MatLab
• **Machine learning frameworks**	Java-ML, Weka, Smile, scikit-learn
• **Application frameworks**	OpenMP, MPI, Amazon EC2, Spring, Hibernate
• **Miscellaneous**	OOP, MVC, TCP/IP, SOAP, SOA, Agile, Linux