List of 10 best papers of Prof. Sanghamitra Bandyopadhyay
(Highlighting the important contributions in each)

1. S. Bandyopadhyay and R. Mitra,"TargetMiner: MicroRNA Target Prediction With Systematic Identification Of Tissue Specific Negative Examples", Bioinformatics, 25(20), pp.2625-2631, 2009 [IF: 6.937, citations: 215].

   The existing machine learning based approaches for microRNA target prediction relied on randomly selected or artificially generated negative samples for training, and hence suffered from high false positive rates. Development of TargetMiner, incorporating systematic identification of tissue specific, biologically relevant negative samples, is one of Sanghamitra's major contribution. It provided the best specificity-sensitivity trade-off as compared to almost all the existing methods. The importance of the contribution is evident from the 10000+ hits of the webserver and inclusion of the genome-wide predictions in miRBase, a heavily used microRNA database.

2. D. Sinha, A. Kumar, H. Kumar, S. Bandyopadhyay and D. Sengupta, "dropClust: Efficient Clustering of Ultra-large scRNA-seq Data", Nucleic Acids Research, vol. 46, iss. 6, pp. e36, 2018 [IF: 16.97, citations: 60].

   For the first time, a clustering method has been proposed that scales well to high-dimensional single cell transcriptomic data without compromising accuracy, while being able to detect rare cell types. Locality Sensitive Hashing, an approximate nearest neighbour search technique, is used for sampling while preserving the structure of the data. On a number of real datasets, dropClust outperformed the existing best practice methods in terms of execution time, clustering accuracy and detectability of minor cell sub-types.

3. D. Sinha, P. Sinha, R. Saha, S. Bandyopadhyay and D. Sengupta, "Improved dropClust R Package with Integrative Analysis Support for scRNA-seq Data", Bioinformatics, vol. 36, iss. 6, pp. 1946-1947, March 2020 [IF: 6.937, citations: 5 (recently published]].

   A novel approach for clustering large-scale single cell expression data is proposed that leverages Locality Sensitive Hashing (LSH) to speed up the processing. This improves the earlier method, dropClust, described in the previous work in this list. In this paper an improved dropClust is presented as a complete R package that is fast, interoperable and minimally resource intensive. The new dropClust features a novel batch effect removal algorithm that allows integrative analysis of single cell RNA-seq (scRNA-seq) datasets.

4. K. Gupta, M. Lalit, A. Biswas, C. D. Sanada, C. Greene, K. Hukari, U. Maulik, S. Bandyopadhyay, N. Ramalingam, G. Ahuja, A. Ghosh, D. Sengupta, "Modeling Expression Ranks for Noise-tolerant Differential Expression Analysis of scRNA-seq Data", Genome Research, doi: 10.1101/gr.267070.120, 2021 [IF: 9.043, citations: - (just published]].

   In this paper, a rank-based measure of gene expression is proposed that operates on single cell transcriptomic data. In the past few years, considerable efforts have been made to identify appropriate parametric models for single cell expression data. The zero-inflated version of Poisson/Negative Binomial and Log-Normal distributions have emerged as the

most popular alternatives due to their ability to accommodate high dropout rates, as commonly observed in single cell data. While the majority of the parametric approaches directly model expression estimates, in this paper the potential of modeling expression-ranks as robust surrogates for transcript abundance is explored. The performance of the Discrete Generalized Beta Distribution (DGBD) is examined on real data leading to the development of ROseq, a Wald-type test for comparing gene expression across two phenotypically divergent groups of single cells. Besides striking a reasonable balance between Type 1 and Type 2 errors, ROSeq is found to be exceptionally robust to expression noise and scales rapidly with increasing sample size. An R package called ROSeq is available on the Bioconductor platform.

5. D. Sengupta and S. Bandyopadhyay, "Topological Patterns in microRNA Gene Regulatory Network: Studies in Colorectal and Breast Cancer", Molecular Biosystems, 9(6), pp. 1360-1371, 2013 [IF: 3.336, citations: 36].

   This paper describes the construction and graph-theoretic analysis of the microRNA induced gene regulatory network, yielding novel markers for colon and breast cancer. The marker for breast cancer, miR-155, was thereafter been independently established through biological experiments in [Martin etal, Genes &Cancer, 5(9-10), 2014].

6. M. Aqil, Afsar R. Naqvi, S. Mallik, S. Bandyopadhyay, U. Maulik and S. Jameel, "The HIV Nef Protein Modulates Cellular and Exosomal miRNA Profiles in Human Monocytic Cells'", Journal of Extracellular Vesicles, vol. 3, 2014 [IF: 25.841, citations: 75].

   For the first time a comprehensive analysis of the miRNA cargo in the exosomes of human monocytic U937 cells expressing HIV Nef protein is carried out. Since Nef functions as a viral suppressor of RNA interference and disturbs the distribution of RNA-induced silencing complex proteins between cells and exosomes, it was hypothesized that it might also affect the export of miRNAs into exosomes. It was experimentally established that Nef expression affected a significant fraction of miRNAs in U937 cells. Our analysis showed 47 miRNAs to be selectively secreted into Nef exosomes and 2 miRNAs to be selectively retained in Nef-expressing cells. The exosomal miRNAs were predicted to target several cellular genes in inflammatory cytokine and other pathways important for HIV pathogenesis, and an overwhelming majority had targets within the HIV genome. We suggested that this is a novel viral strategy to affect pathogenesis and to limit the effects of RNA interference on viral replication and persistence.

7. A. Chakraborty, B. Morgenstern and S. Bandyopadhyay, ``S-conLSH: Alignment-free Gapped Mapping of Noisy Long Reads", BMC Bioinformatics, 22, 64, 2021 [IF: 3.242, citations: - (just published)].

   This paper provides a novel alignment-free algorithm for mapping of SMRT reads to the reference genome. While SMRT reads are longer with low GC bias, they suffer from high levels of noise. The proposed method called S-conLSH uses Spaced context based Locality Sensitive Hashing. With multiple spaced patterns, S-conLSH facilitates a gapped mapping of

noisy long reads to the corresponding target locations of a reference genome. The method achieves a sensitivity of 99% on human simulated sequence data, while being significantly faster than many state-of-the-art methods.

8. U. Maulik and S. Bandyopadhyay, "Genetic Algorithm Based Clustering Technique", *Pattern Recognition*, 33(9), pp.1455-1465, 2000 [IF: 7.196, citations: 1738].

   Prof. Bandyopadhyay was instrumental in developing this new clustering technique using genetic algorithms, with encoding of clustering centres, integrated with an iteration of local search. While earlier schemes often required the chromosomes of the genetic algorithm to be as long as the number of data points, the proposed strategy reduced the size of the chromosomes and resulted in significantly faster search. This approach is now the basis in many evolutionary clustering techniques.

9. U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12), pp.1650-1654, 2002 [IF: 16.389, citations: 1445].

   A new cluster validity index that enabled automatic identification of the model order was introduced in this article. Prof. Bandyopadhyay developed the theory behind the index, proving uniqueness and global optimality through its relationship with the well-known Dunn's index. The popularity of the index is evident from its 1445 citations in Google Scholar.

10. S. Bandyopadhyay, S. Saha, U. Maulik and K. Deb, "A Simulated Annealing Based Multiobjective Optimization Algorithm: AMOSA", *IEEE Transactions on Evolutionary Computation*, 12(3), pp.269-283, 2008 [IF: 17.13, citations: 875].

    This work represents the first effective technique for objective optimization (MOO) using simulated annealing. In MOO, up to four conflicting objectives, are simultaneously optimized. Such problems abound in real-life. Although simulated annealing had a strong theoretical background, its MOO version was rudimentary because of its point-based-search nature. By developing AMOSA, incorporating novel concepts of amount of domination and situation specific acceptance probabilities, Prof. Bandyopadhyay filled this gap in the literature. Because of non-greedy selection, AMOSA proved to be highly effective even for many-objective optimization, where the number of objectives to be optimized exceeds four. The importance of the contribution is evident from its applications in gene selection, microarray data clustering, finding protein modules and remote sensing image segmentation.

---------- XXX ----------