MaRS Centre
661 University Avenue, Suite 510
Toronto, Ontario Canada, M5G 0A3
✉ *falmodaresi@oicr.on.ca*
https://www.linkedin.com/in/fataltes
https://github.com/fataltes
https://fataltes.github.io

Fatemeh Almodaresi

Research Interests

Computational Biology
Probabilistic Modeling and Machine Learning
Algorithms and Data Structures

Skills

Programming Languages	**C++ (expert), Python (expert), R (familiar), Java Core(expert), MATLAB (familiar), NetLogo (familiar), C# (familiar), Rust (starter), MySQL (expert), Oracle (expert)**.
Tools, Libraries and Frameworks	**Bioinformatics Analysis Tools (e.g. SAMtools, Bedtools, Picard, BWA, BWA-MEM2, STAR, STARsolo, Bowtie2, Salmon, Minimap2, Seurat, IGV, StringTie2, DESeq2, etc.), LLM (e.g. LangChain framework), Popular Python Libraries (numpy, pandas, scipy.stats, sklearn, matplotlib, and seaborn), ShinyR**.
Databases	**TCGA, PCAWG, GO ontology and enrichment analysis tools, Human Cell Atlas, Reactome**.
Machine Learning	**Large-Language-Models (LLMs), feature selection, dimensionality reduction, (un)supervised learning, pattern recognition**.
Others	**Git, Bash script, Snakemake, Docker, Anaconda Platform, Jupyter Notebook, IntelliJ IDEA, CLion, Pycharm, Atlassian Jira**.

Education

Feb 2021-Now **PostDoctoral Fellow**, *Adaptive Oncology Department, Ontario Institute for Cancer Research (OICR)*, ON, CAN.
Advisor – Prof. Lincoln Stein

2019-Aug 2020 **Ph.D**, *Computer Science Department, University of Maryland (UMD)*, MD, US.
Advisor – Prof. Rob Patro

2015-2019 **MS**, *Computer Science Department, Stony Brook University (SBU)*, NY, US.
Advisor – Prof. Rob Patro

2004-2009 **BS**, *School of Electrical & Computer Engineering (ECE), University of Tehran*, Tehran, Iran.

Selected Research Projects

May. 2023 - Present **PathwayLLM**, *"Talk to your Pathway"*, A collaboration between Adaptive Oncology Dep., OICR and WangLab., Vector Institute.
Leading a team in developing a computational solution using **advanced language models**, **Reactome Knowledge Graph**, and **PubMed data** through text mining and embedding techniques. Our system employs retrieval-augmented generation (**RAG**), and utilizes **vector databases such as Pinecone** for text indexing, **LLM embeddings such as Huggingface embeddings** and **prompt engineering techniques**. We encorporated manual reference construction and fact checking through text semantic similarity search between results and source documents. It generates Reactome-like pathway summaries, handles complex pathway queries, and formulates hypotheses for uncharted pathways. Ongoing work includes expanding to more databases and optimizing performance.

Oct. 2022 - Present **ScReps**, *"Discovering novel cancer signals from retroposans in single cell unaligned reads"*, Adaptive Oncology Dep., OICR.

In this project, I explore the therapeutic potential of retroelement misexpression in cancer cells for cancer prognosis. The computational pipeline involves multiple key steps, including

- Aligning 10X reads with Cellranger
- Integrating samples using Harmony
- Annotating cells via Seurat
- Selecting unaligned reads
- Aligning these reads to Repbase repetitive elements database using Salmon
- and Conducting single-cell differential expression analysis between Healthy and Condition samples for each cell type using ZINB_DESeq2 and AggregateBioVar3.

Applying the data on diverse datasets, including control datasets like 10X-PBMC8k, Medulloblastoma Tumor Cells (G3, G4, SHH, CPA, PFA-PFB), and Sarcoma Immune Cells (across various conditions and individuals). I found significant expression-based signals differentiating cancer and healthy samples. The pipeline is written in bash. Next steps are creating a Snakemake for the pipeline and expanding initial analysis to more datasets to confirm the results.

Feb. 2021 - Present **Immunotherapeutic Targeting of the U1 snRNA Mutation in Cancer**, Adaptive Oncology Dep., OICR.

I led a pioneering project across multiple institutions targeting novel U1 snRNA mutations prevalent in hard-to-treat cancers like CLL and pediatric cerebellar MB. Our pipeline, starting from Nanopore long reads, involved rigorous sequence filtering, read assembly, and transcript consolidation. We first dscard sequences with imbalance-primer using Pychopper and then utilized advanced tools such as minimap2, and StringTie2 for alignment and assembly of the rest of the reads. With access to both Mutant and Wildtype samples, we filtered out potential sequencing artifacts using paired QC-based analysis. By focusing on mutant-specific transcripts and selecting key MHC-binding peptides, we aimed to develop targeted therapeutics for these U1 mutant tumors, offering a promising approach to improve treatment outcomes in challenging cancer cases.

Aug. 2017 - 2021 **Mantis**, *"A fast, small, and exact large-scale sequence-search index"*, Computational Biology Lab., SBU, UMD, `https://github.com/splatlab/mantis/tree/mergeMSTs`.

Mantis is a space and time efficient data structure to index and query large collections of raw sequencing read experiments. The index is based on colored de Bruijn graph representation and therefore supports graph-based operations such as graph traversal, and bubble calling useful for assembly and variation detection. In our recent work we have advanced the index to more than 30,000 raw read sequencing samples and enabled the nice feature of gradual growth by making the index incrementally updatable.

Jun. 2017 - Aug. 2020 **Pufferfish and Puffaligner**, *"A space and time-efficient compacted de Bruijn graph index and aligner"*, Computational Biology Lab., SBU, UMD, `https://github.com/COMBINE-lab/pufferfish/tree/develop`.

Pufferfish is an efficient data structure for indexing colored compacted de Bruijn graphs. This index achieves a balance between time and space resources by making use of succinct data structures and minimum perfect hash function. PuffAligner, our recent work, is a highly sensitive aligner on top of Pufferfish for aligning different types of short sequencing reads to a huge population of references, specifically good in the representation of high similarity in the reference sequences.

Apr. - Jun. 2017 **Rainbowfish**, *"A succinct colored de Bruijn graph data structure"*, Computational Biology Lab., SBU, `https://github.com/COMBINE-lab/rainbowfish`.

This tool provides a new data structure to store and query colored de Bruijn graphs that in case of large data sets improves storage by more than twenty times compared to state-of-the-art tools without hurting performance of the queries.

Nov. 2016 - Apr. 2018 **Grouper**, *An extension to "Rapid Clustering" tool*, Computational Biology Lab., SBU, `https://github.com/COMBINE-lab/grouper`.

Grouper is a tool for clustering contigs of a de novo transcriptome assembly. We improved the accuracy of clustering by making use of orphan reads, for which each end of the pair is mapped to a different reference sequence.

Aug 2016-Jan 2017 **MLDD**, *"Multi-Level Distribution Detection"*, Data Science Lab., SBU.

Using statistical tests and classification models such as NaiveBayes we show how distribution of NLP features in social media changes in different levels of analysis (county, user, and message). This can highly affect prior assumptions for further text analysis as we show that central-limit theorem could be applied in social media language analysis as well.

2013-2014 **AutismFD**, *"A game to improve face emotion detection in children with Autism"*.

Beside collaboration with psychology students to design the method, I also implemented the idea as a tool in C# language. This package was used in a treatment center to help children with Autism to identify face emotions and track their progress over time.

Publications

[1] Fatemeh Almodaresi, Giorgos Skoufos, Mohsen Zakeri, Joseph N Paulson, Rob Patro, Artemis G Hatzigeorgiou, and Ioannis S Vlachos. Agamemnon: an accurate metagenomics and metatranscriptomics quantification analysis suite. *Genome biology*, 23(1):1–27, 2022.

[2] Fatemeh Almodaresi, Jamshed Khan, Sergey Madaminov, Michael Ferdman, Rob Johnson, Prashant Pandey, and Rob Patro. An incrementally updatable and scalable system for large-scale sequence search using the bentley–saxe transformation. *Bioinformatics*, 2022.

[3] Fatemeh Almodaresi, Mohsen Zakeri, and Rob Patro. Puffaligner: A fast, efficient, and accurate aligner based on the pufferfish index. *Bioinformatics*, 2021.

[4] Avi Srivastava, Laraib Malik, Hirak Sarkar, Mohsen Zakeri, Fatemeh Almodaresi, Charlotte Soneson, Michael I Love, Carl Kingsford, and Rob Patro. Alignment and mapping methodology influence transcript abundance estimation. *Genome biology*, 21(1):1–29, 2020.

[5] Fatemeh Almodaresi, Prashant Pandey, Michael Ferdman, Rob Johnson, and Rob Patro. An efficient, scalable and exact representation of high-dimensional color information enabled via de bruijn graph search. In *International Conference on Research in Computational Molecular Biology*, pages 1–18. Springer, 2019. A slightly modified version appeared in Journal of Computational Biology 27 - 2020.

[6] Prashant Pandey, Fatemeh Almodaresi, Michael A Bender, Michael Ferdman, Rob Johnson, and Rob Patro. Mantis: A fast, small, and exact large-scale sequence-search index. *Cell Systems*, 2018.

[7] Laraib Malik, Fatemeh Almodaresi, and Rob Patro. Grouper: Graph-based clustering and annotation for improved de novo transcriptome analysis. *Bioinformatics*, 1:8, 2018.

[8] Fatemeh Almodaresi, Hirak Sarkar, Avi Srivastava, and Rob Patro. A space and time-efficient index for the compacted colored de bruijn graph. *Bioinformatics*, 34(13):i169–i177, 2018. (appeared in the proceedings of ISMB 2018).

[9] Mohsen Zakeri, Avi Srivastava, Fatemeh Almodaresi, and Rob Patro. Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics*, 33(14):i142–i151, 2017. (appeared in the proceedings of ISMB 2017).

[10] Fatemeh Almodaresi, Lyle Ungar, Vivek Kulkarni, Mohsen Zakeri, Salvatore Giorgi, and H Andrew Schwartz. On the distribution of lexical features at multiple levels of analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 79–84, 2017.

[11] Fatemeh Almodaresi, Prashant Pandey, and Rob Patro. Rainbowfish: A Succinct Colored de Bruijn Graph Representation. In *17th International Workshop on Algorithms in Bioinformatics (WABI 2017)*, volume 88 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 18:1–18:15, 2017.

Work Experiences

Jun-Aug 2016 **Member of the NLP Team**, *Third Frederick Jelinek Memorial Summer Workshop (JSALT)*, MD, US.
JSALT is a well-known summer workshop in Language and Speech organized by JHU at Baltimore each year. .
During the project, we worked on analyzing and forecasting social media user's psychological state based on their language in their posts using statistical methods such as significance tests and time series models such as ARMA and ARIMA.

Jan-Aug 2015 **Senior Designer and Developer**, *Nexeven AB*, Tehran, Iran.
Nexeven AB is a Swedish company and a niche player in the online video broadcasting field.

2011-2015 **Team Supervisor, Senior Designer and Developer**, *Tosan Intelligent Data Miners Co. (TIDM)*, Data Mining Development Team, Tehran, Iran.
TIDM is the first solution provider for fraud detection and anti-money laundry in banking section in Iran, a Subsidiary of Tosan Company.

- **Customer Relationship Management System** [2014]
  In this project we use statistical and data mining methods to calculate customer's RFM, CLV, and churn probability.
- **Data mining Module, Operational Intelligence System** [2014]
  This module, developed in PLSQL, uses Statistical and Mining Methods such as regression models, error functions, k-means, and SVM to detect fraudulent transactions online in the stream of transactions.
- **Customer Name Similarity Detection Module** [2013]
  As a part of Anti-money Laundry System, this module uses natural language algorithms to detect accounts with similar names. The whole system is developed in PLSQL and now operational in many private banks in Iran including Eghtesad-Novin and Ansar Bank.
- **Unsupervised Fraud Detection System, Version 1 & 2** [2011-2014]
  Version 1 which is fully designed and developed by myself is now operational in Saman Bank, Ansar Bank, and Mehr-e-Eghtesad Bank in Iran. Version 2 is now installed in Eghtesad-Novin Bank.

Honors & Awards

2020 **Larry Davis Dissertation Award**, *UMD*.
2019 **Grace Hopper Conference 2019 (GHC19) Scholarship**.
2019 **Catacosinos Fellowship for Excellence in Computer Science**, *SBU*.
2019 **RECOMB2019 Conference Travel Fellowship**.
2018 **ISMB2018 Conference Travel Fellowship**.
2016 **CS Department Best TA Award**, *SBU*.
2015 **Special CS Department Chair Fellowship**, *SBU*.

Teaching Experiences

Fall 2017 **Teaching Assistant**, *Computational Biology*, Stony Brook University.
Spring 2017 **Teaching Assistant**, *Machine Learning*, Stony Brook University.