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
Algorithms and Data Structures
Statistical Data Analysis and Probabilistic Modeling

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

Nov. 2022 - Present **Ultra-small ultra-fast accurate scBERT using biological priors**, Adaptive Oncology Dep., OICR.
scBERT which has recently been published is a slightly modified and adjusted BERT that provides an accurate model for representing the single cell gene expressions and uses it for downstream task of cell annotation. However, the presentation has a few flaws that stops the model from easily extending to big single cell ATLASes. It is also not tuned specifically for unsupervised tasks such as data imputation which are much harder in nature. The goal of this project is to modify the model and map the input to a more meaningful representation of the cells and genes. The new presentation would allow for much smaller inputs and therefore save memory and training time. We also use biological priors such as pathways, GRNs, and GO to enrich the model. The input would consider genes as words and cells as sentences in the mapping of biological data to NLP and have additional embeddings derived from biological knowledge. In addition to that we want to provide an interpretation of the model to use for imputation purposes. A new line of research is growing regarding interpretation of the transformer models which would allow us to have a better understanding of the underlying gene2gene interactions that results in a highly accurate prediction of cell types (or other annotations) using transformers.

Oct. 2022 - Present **ScReps**, *"Discovering novel cancer signals from retroposans in single cell unaligned reads"*, Adaptive Oncology Dep., OICR.
Nearly half of the human genome is composed of repetitive sequences. Studies show that cancer cells often misexpress retroelements due to decrease in methylation and therefore, an increased expression of repetitive elements can enable access to cancer therapeutic opportunities.There has been a tremendous attention and rise to research on single cell analysis, but all dedicated to aligned reads. In this project, we investigate the role of repeated elements in unaligned reads in differentiating between tumor cells and normal cells and utilize it for further understanding and prediction of the cell state (cell cycle etc.) and tumorous cell development and lineage. We investigate the potentials of diagnostics and therapeutic usecases of unaligned reads in a set of 45 samples of Medulloblastoma cancer primaries in collaboration with scientists from SickKids, ON.

Feb. 2021 - Present **Immunotherapeutic Targeting of the U1 snRNA Mutation in Cancer**, Adaptive Oncology Dep., OICR.
The goal of this project is to Explore targeted therapeutics against a novel class of mutation in several types of hard-to-treat cancers. The mutation happens in the U1 snRNA and is present in majority of several cancer types such as CLL (chronic lymphocytic leukemia) and pediatric cerebellar MB (medulloblastoma). Since the patterns of mis-splicing caused by this mutation are identical across different samples, this makes an opportunity for having a global therapeutic approach towards neo-antigens present in U1 mutant tumors. For that matter we have explored the primaries as well as the cell lines for both CLL and MB samples and built up a pipeline for finding interesting peptides that is highly present across a great proportion of samples. The pipeline's focus is on predicting and validating MHC class I neo-peptides (neo-MAPs).

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.

2009-2011 **Java and UI Developer** , *Tosan Co*, Tehran, Iran.
Tosan Company is a pioneer company for total banking solutions with more than fifteen Iranian financial institutes in its customer list. As a member of a team of nearly 20 people, I participated in developing the UI of Internet banking system.

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.

2013 **Teacher**, *C++ Programming Language*, Farzanegan High School [NODET].

2013 **Teacher**, *Developing simple motion detection algorithms in MATLAB*, Farzanegan High School [NODET].

Fall 2008 **Teaching Assistant**, *Artificial intelligence*, University of Tehran.

Skills

Programming Languages	**C++ (expert), Python (expert), R (familiar), Java Core(expert), MATLAB (familiar), NetLogo (familiar), C# (familiar), Rust (starter)**.
Libraries and Frameworks	**Popular Python Libraries (numpy, pandas, scipy.stats, sklearn, matplotlib, and seaborn), ShinyR, Spring Framework, Hibernate**.
Databases	**MySQL (expert), Oracle (expert)**.
Other Tools	**Git, Jupyter Notebook, IntelliJ IDEA, CLion, Atlassian Jira, Atlassian Confluence, ThoughtWorks Go, Anaconda Platform, Pycharm**.