MOSES STAMBOULIAN

PhD student at Indiana University
700 N Woodlawn Ave.
Bloomington Indiana 47405

mossigstamboulian@gmail.com
(812)-327-1498
Linkedin, Github, Website

Objective

- **An informatics PhD student with extensive research experience.**

Education

- **Indiana University (IU)** — Bloomington, Indiana, U.S.A
  PhD Bioinformatics GPA: 4.0/4.0 — *2015 - 2020*
- **University of Newcastle (NCLU)** — Newcastle, United Kingdom
  MSc Bioinformatics Score: 78/100 — *2013 - 2014*
  - 10 month Erasmus Mundus student exchange program
- **Lebanese American University (LAU)** — Beirut, Lebanon
  MSc Computer Science GPA: 3.5/4.0 — *2012 - 2015*
- **Haigazian University (HU)** — Beirut, Lebanon
  BSc Computer Science Score: 87.67/100 — *2008 - 2012*
- **C. Gulbenkian Secondary school** — Hoch Moussa, Lebanon
  Secondary classes and Lebanese Baccalaureate certificate — *2005 - 2008*

Work Experience

- **Bioinformatics Scientist intern** — 06/2020– 08/2020
  Guardant Health
  - worked under one of the core Bioinformatics teams at Guardant Health, developing validation pipelines for Guardant Health 360 tumor identification and analysis pipeline.
  - Developed a quick and efficient tool for Sequence Alignment Map file comparisons
- **Research Associate** — 01/2020– present
  Microbiome and Me LLC
  - Basic tasks include data analysis, visualization and summary statistics for the company
  - Company website development and maintenance
  - implementation of customized pipelines for data analysis over multi-omics data (genomics, transcriptomics and proteomics)
- **Research Assistant** — 08/2018– present
  Indiana University Bloomington
  - Conducting research in a human microbiome lab. Projects involve multi-omics data, (metagenomics, metatranscriptomics and metaproteomics)
- **Associate Instructor** — 08/2015– present
  Indiana University Bloomington

  - Teaching Python programming, Advanced Python programming, data-mining, search informatics, machine learning for bioinformatics

- **Teaching Assistant** 02/2013 – 06/2015
  Lebanese American University
  - Taught introduction to Computing for non CS Majors

- **Research Assistant** 09/2012 – 07/2015
  Lebanese American University
  - Worked on Graph algorithms such as minimizing the complexity of Vertex Cover Algorithm, Published a paper on Computational Protein Structure Prediction

- **Junior C developer** 04/2012 – 10/2012
  Information Management Limited
  - Developed and Maintained Apps for Accounting, Warehouses and Point of Sales.

- **Summer Internship** 08/2011 – 09/2011
  Vivacell MTS Armenia
  - Worked in a team that got dispatched to on-site network infrastructure installation/maintenance, also performed server side software maintenance and management.

Projects

A Tree of Human Gut Bacterial Species and its Applications to Metagenomics and Metaproteomics Data Analysis (Stamboulian et. al)

02/2019 - 07/2020

- In this project we constructed a large database of human gut bacterial species (over 3,000 genomes) from recent studies and proposed a new approach to build a phylogenetic tree which we show works well under conditions of incomplete genomes and or genomes with distant ancestry. We also demonstrate some applications of our constructed bacterial tree in the context of metagenomics and metaproteomics.
- *Skills: Python, Bash, phylogenetics, tree construction, information retrieval, data-mining, data analysis.*
  - Collected complete and near complete human gut bacterial genomes, from recently published studies.
  - Predicted protein coding genes across all genomes.
  - Extracted and annotated marker genes across all genomes using hidden markov models.
  - Constructed a large phylogenetic tree using all the genomes with the proposed marker gene averaging method.
  - Annotated genomes up to species level in most cases (when possible) using average nucleotide identity approaches.
  - demonstrated two useful applications of having a comprehensive gut bacterial gene tree, in the context of metagenomics and metaproteomics.
- Paper under review at BMC Evolutionary Biology

Using high abundance proteins as guides for fast and effective peptide/protein identification from human gut metaproteomic data (Stamboulian et. al)

02/2019 - 06/2020

- Project's objective is two folds: First to develop a pipeline to profile and characterize gut bacterial samples through a metaproteomic perspective. The objective will be to identify and report the abundance of representative species in these samples and identify some of the protein functions that are expressed across the different samples. The second outcome of this pipeline is to construct a targeted database to speed up the spectral search, improve it's accuracy and its throughput.
- *Skills: Python, Bash, peptide identification, information retrieval, data-mining, data-analysis*
  - Compiled a relatively large bacterial database based on the recently published genome catalogs.
  - A representative bacterial tree of life was created using these representative binned genomes.
  - Peptide search and identification using mass spectrometry data.
  - building a pipeline that will first profile a metaproteomic sample extract representative species from these samples, report their abundances and perform targeted search.
- Paper under review at the Journal of Microbiome

Pathway-based and phylogenetically adjusted quantification of metabolic interaction between microbial species (Lam et. al)

01/2020 - 07/2020

- Project's objective is to develop a novel approach for estimating the competition and complementarity indices for a pair of microbial species, adjusted by their phylogenetic distance. An automated pipeline, PhyloMint, was implemented to construct competition and complementarity indices from genome scale metabolic models derived from microbial genomes.
- *Skills: Python, Bash, assembly, abundance quantification, Information Retrieval,Data-mining, Data-analysis, phylgenetics, evolutionary distance calculation*
  - Competition and Cooperation between microbial entities were calculated using CarveMe, starting from seed metabolites.
  - metabolic interaction indices were normalized using evolutionary distances calculated by constructing a comprehensive bacterial species tree.
  - a discretization approach, was used to detect pairs of bacterial species with cooperation scores significantly higher than the average pairs of bacterial species with similar phylogenetic distances.
  - a network community analysis of high metabolic cooperation but low competition reveals distinct modules of bacterial interactions.
- Paper under review at Plos Computational Biology.

Protein Functional Landscapes of human gut bacteria using metaproteomics

01/2019 - present

- Project's objective is to characterize functional landscapes of different individuals gut, having different underlying conditions, by profiling them at the metaproteome level and studying highly expressed bacterial species and their expressed protein functions using metaproteomics datasets.

- *Skills: Python, Bash, Metaproteomics, abundance quantification, Information Retrieval,Data-mining, Data-analysis, protein-protein-interaction networks (ppi)*
  - Numerous human gut metaproteomics samples are gathered, profiled and peptides are identified.
  - Top 5/10 bacterial species for each individual were studied for functional landscapes.
  - Protein-protein interaction networks were constructed based on protein expression profiles.
  - expressed peptides were mapped back to genomes and gene prediction was revised.

Protein sequence to functional annotations using deep learning 05/2020 - present

- Project's objective is to develop and train a Long Short term Memory deep neural network based predictor to accurately fragment the protein into functional sites and assign the necessary functional annotations to these individual domains.
- *Skills: Python, Bash, Deep learning, neural networks, tensor flow, pytorch, Data-analysis*
  - Exhaustive list of protein sequences from uniprot database was collected and manually curated for annotations.
  - Pfam annotations associated with each sequence were collected and maintained.
  - Sequence to vector and annotation to vector encodings are engineered.
  - Deep neural networks are trained and tested over different subsets of our training data.

Predicting venous thromboembolism risk from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges (McInnes et. al) 10/2018 - 05/2019

- Project's objective was to predict venous thromboembolism (VTE) status in exome data from African American subjects for 103 unlabeled exomes from patients treated with warfarin for nonVTE causes or VTE and asked to predict which disease each subject had been treated for.
- *Skills: Python, VCF tools, SNP calling, GATK Bash, assembly, reads mapping, Data-mining, Data-analysis*
  - protein coding variants from raw VCF files were annotated using ANNOVAR.
  - Mutpred2 and Mutpred-LOF were use to assign pathogenecity prediction scores per individual exome.
  - Confirmed risk genes were used as seed in the human PPI -network to run network propagation algorithm.
  - Beta distributions for Mutpred pathogenecity scores were generated using top 100 scoring genes for each phenotype.

Tree guided bining of phylgenetic markers 11/2018 - present

- Project's objective is developing a tool that will group short reads from metagenomics samples into binned genomes by taking advantage of time series datasets, through abundance profile correlations and evolutionary relationships of these markers.

- *Skills: Python, Bash, assembly, abundance quantification, Information Retrieval, Data-mining, Data-analysis*
  - Metagenomic reads in each sample get assembled into contigs.
  - Marker genes get predicted from each of the contigs.
  - abundance quantification profiles are extracted.
  - a separate gene tree gets created from each of the gene markers
  - A tree guided and abundance correlation based genome marker binning is performed.

Pathogenecity Prediction for NFS-indels 06/2017 - 02/2018

- Project's objective is developing a machine learning model to predict whether or not a given non-frame shifting insertion or deletion is pathogenic.
- *Skills: Python, Matlab, Bash, information retrieval, machine learning, data-mining, data-analysis*
  - Extracted Pathogenic and putatively neutral non-frame shifting variants from HGMD and gnomAD databases. Cleaned data-sets and prepossessed for analysis.
  - Performed an exploratory data-analysis over the data to study for patterns and properties.
  - Performed Feature engineering to extract features and attributes of data that showed interesting signals and patterns during data-analysis.
  - Created two classifiers, Supprot Vector Machine based and Random Forrest based, trained, tested them and reported results.

The Ortholog Conjecture Revisited: the Value of Orthologs and Paralogs in Function Prediction, (Stamboulian et. al)
10/2015 - 6/2018

- Project's objective: Understanding the evolution of protein function between species, human, mouse in particular and *S.* cerevisiae, *Sch* pombe, and assessing values of orthologs and paralogs in the context of protein function prediction.
- *Skills: Python, Matlab, R, perl, Information Retrieval, SQL, Bash, Data-mining, Data-analysis*
  - Developed Scripts to download data from different databases and data-sources.
  - Developed scripts to annotate protein functions, calculate functional similarity.
  - Developed scripts to calculate functional similarity using different distance metrics/measures.
  - Performed Data-Analysis over annotation data and protein homology data.
  - Paper accepted at the Journal of Bioinformatics.

Predicting House Prices in Ames, Iowa (Kaggle competition) 9/2016 - 12/2016

- Project's objective: Employed different regression techniques to correctly estimate prices of real estates in Ames Iowa.
- *Skills: Python, Information Retrieval, SQL, Bash, Data-mining, Data-analysis, Machin learning*
  - Data-retrieval, data-normalization cleaning, predicting missing values.
  - Data-analysis and feature extraction

- Feature Selection and implementation of regression approaches for real estate price prediction problem.
- Converting the problem to a classification and developing a classifier for the task.

Scatter Search for Homology Modeling (Stamboulian et. al) 01/2015 - 06/2015

- Project's objective was to improve protein's three dimensional structure prediction by optimizing the target-template sequence alignments using scatter search meta-heuristic
- *Skills: Python, Modeller, Information Retrieval, SQL, Bash, Java, Heuristics, optimization, genetic programming, scatter search*
  - Developed a wrapper to interact with the Modeller program for comparative modeling
  - Developed a Scatter Search meta-heuristic approach sequence alignment optimization to iteratively improve sequence alignments.
  - Predicted three dimensional structures for proteins, and compared accuracy with other approaches.

Awards, Grants & Honours

Vatche and Tamar Manoukian Student Grant . . . . . 2008 - 2011
Jorjorian Student Grant 2008-2011
Haigazian University Financial Aid 2008-2012
Dean's List and President's list award (6 out of 8 undergraduate semesters) 2008-2012
Lebanese American University Graduate Assistantship 2012-2013; 2014-2015
Lebanese American University Research Assistantship 2012-2013; 2014-2015
Haigazian University Financial Aid 2008-2012
Erasmus Mundus Welcome Study Abroad Scholarship 2013-2014
Indiana University PhD Scholarship 2015-2020

Technical Skills

- **Programming languages:** Python, Java, Bash/shell scripting, Matlab, SQL, R, C, C++.
- **Bioinfromatics:** Computational genomics, Comparative genomics, DNA/RNA sequence analysis, Differential gene expression, Cancer genomics, Marker gene identifications, Proteomics, Peptide-sequence identification and quantification, Sequence assembly, Sequence alignment, Genome annotations, Gene and protein prediction, Variant calling, Gene annotation, Protein functional annotation, Computational genome binning, Microbiology.
- **Computer Science:** Data analytics, Data visualization, Computational statistics, Exploratory data analysis (EDA), Data mining, Machine learning.
- **Tools:** Linux, Unix, HPC systems, Git/version Control, APIs, Pandas, Numpy, NetworkX, sickit-learn, SciPy, Json, Biopython, Multi-processing, Keras, TensorFlow, Pytorch, Jupyter Notebook, Rmarkdown, Flair, Command line, LaTeX,

Languages (Speaking, Reading and Writing)

- Armenian (*Native*), English (*Fluent*), Arabic (*Fluent*).