BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. DO NOT EXCEED FIVE PAGES.

NAME: LEUNG, MING-YING

eRA COMMONS USER NAME (credential, e.g., agency login): mleung

POSITION TITLE: Professor

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE (if applicable)	END DATE MM/YYYY	FIELD OF STUDY
University of Hong Kong	BS	1980	Mathematics
University of Hong Kong	MPHIL	1983	Mathematics
Stanford University, Stanford	MS	1988	Computer Science
Stanford University, Stanford	PHD	1989	Mathematics

A. Personal Statement

My research focuses on developing statistical models and computational algorithms for bioinformatics analysis of biomolecular sequence data. In particular, I have developed Markov models, scan statistics, and computational algorithms to identify unusual palindromic patterns in the nucleotide sequences and have applied them to the analysis of genomic sequences of DNA and RNA viruses like the herpesviruses and SARS coronaviruses. I have also been developing efficient computational approaches for predicting secondary structures, including pseudoknots, of long RNA sequences. We have devised methods to circumvent the extremely high demands of memory and computing time in the structure prediction problem by using grid computing technologies available in the UTEP Border Biomedical Research Center (BBRC) Bioinformatics Computing Core Facility, where I serve as director. The research in these projects has resulted in several bioinformatics software packages publicly accessible online (bioinformatics.utep.edu/BCL). For more specific implementation of bioinformatics computing tools for RNA research, we have launched the RNA virtual laboratory (rnavlab.utep.edu) for RNA sequence analysis, structure prediction, and database development. Using a whole-exome sequencing approach coupled with gene ontology analysis to identify exonic DNA variants in patients with acute lymphoblastic leukemia from the El Paso Children's Hospital, we have added a Python-based pipeline called OncoMiner (oncominer.utep.edu) to our repertoire of tools for genomic data analytics. These web-based software applications have been made available to the biomedical research community worldwide. I have served as PI and co-PI in various projects for setting up and managing computational facilities for biomedical research and as director of the Bioinformatics and Computational Science Programs for administering graduate research training at UTEP. Over the past 30 years, my research has been funded by NIH, NSF, IBM, USDA, and the Cancer Prevention and Research Institute of Texas in viral genome replication, sequence analysis, and RNA structural prediction with a long record of publication in these areas. In addition, I am involved in several collaborative research projects in identifying protein biomarkers for hepatocellular carcinoma and studying the effects of different blood flow patterns on gene expression. In this proposed U54 project, I will lead the Integrative Analytics Unit to provide computational, informatics, and statistical support in developing data analytics methods and computational pipelines for biometric, molecular, and sociobehavioral data. For the clinical research project in particular, I will also be heavily involved in the design of appropriate databases to ensure secure data storage and efficient data retrieval to facilitate subsequent analyses, as well as the implementation and testing of statistical and machine learning models to establish reliable predictive models for the proposed biological age metric and health advantage ratio. My goal is to assist biomedical and clinical researchers to arrive at meaningful conclusions in their studies leading to new discoveries that will help reduce health disparities.

1. Mohl JE, Gerken TA, Leung MY. ISOGlyP: de novo prediction of isoform-specific mucin-type O-glycosylation. Glycobiology. 2021 Apr 1;31(3):168-172. PubMed Central PMCID: PMC8022967.
2. Begum K, Mohl JE, Ayivor F, Perez EE, Leung MY. GPCR-PEnDB: a database of protein sequences and derived features to facilitate prediction and classification of G protein-coupled receptors. Database (Oxford). 2020 Nov 20;2020 PubMed Central PMCID: PMC7678784.
3. Leung M, Knapka J, Wagler A, Rodriguez G, Kirken R. OncoMiner: A Pipeline for Bioinformatics Analysis of Exonic Sequence Variants in Cancer. In: Wong K, editor. Big Data Analytics in Genomics [Internet] Cham: Springer International Publishing; 2016. Chapter Chapter 12373-396p. Available from: http://link.springer.com/10.1007/978-3-319-41279-5_12 DOI: 10.1007/978-3-319-41279-5_12
4. Leung MY, Marsh GM, Speed TP. Over- and underrepresentation of short DNA words in herpesvirus genomes. J Comput Biol. 1996 Fall;3(3):345-60. PubMed Central PMCID: PMC4076300.

B. Positions, Scientific Appointments and Honors

Positions and Scientific Appointments

2022 - 2022	External Reviewer, PhD Computational and Data Science Program, Middle Tennessee State University, Murfreesboro, TN
2016 - 2016	Member, Scientific Program Committee, International Workshop on Applied Probability, Toronto
2013 -	Director, Computational Science Program, The University of Texas at El Paso, El Paso, TX
2013 -	Mentor, National Alliance for Doctoral Studies in the Mathematical Sciences
2012 -	Organizer, Joint UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Sciences
2011 - 2011	Chair , Cluster on Computational Biology, Institute for Operations Research and Management Science (INFORMS) Annual Conference, Charlotte, NC
2011 - 2011	Chair, Session on Computational Methods in Biomolecular and Phylogenetic Analyses, International Federation of Operational Research Societies (IFORS) Conference, Melbourne
2010 - 2010	External Review Panelist, West Virginia IDeA Network for Biomedical Research Excellence, Research Competitiveness Program, American Association for the Advancement of Science
2009 -	Member, NIH-RCMI Translational Research Network Translational Informatics Subcommittee
2008 - 2010	Member, The University of Texas System Computational Biology Workgroup for the Cancer Prevention and Research Institute of Texas, Austin, TX
2008 - 2008	Chair, Invited Session on Stochastic Models for Biological Processes, Compiegne
2007 - 2013	Associate Editor, INFORMS Journal on Computing
2005 - 2010	Chair, Sessions on Bioinformatics and Stochastic Models for Biological Systems, INFORMS Annual Meeting
2003 -	Professor, Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX
2003 -	Director, Bioinformatics Program, The University of Texas at El Paso, El Paso, TX
2001 - 2002	Visiting Associate Professor, Department of Statistics, Rice University, Houston, TX
1993 - 1993	Visiting Research Fellow, Department of Statistics, University of California at Berkeley, Department of Pharmaceutical Chemistry, University of California at San Francisco

1989 - 2003 Assistant and Associate Professor, Division of Mathematics and Statistics, The University of Texas at San Antonio, San Antonio, TX

1983 - 1989 Research Assistant and Teaching Fellow, Department of Mathematics, Stanford University, Stanford, CA

1982 - 1983 Lecturer, Department of Extramural Studies, University of Hong Kong

1980 - 1983 Teaching Assistant, Department of Mathematics, University of Hong Kong

Honors

2017 - 2018 Student Choice Award for Outstanding Teaching, The University of Texas at El Paso

2014 - 2015 Outstanding Performance Award in Securing Extramural Funding, Office of Research and Sponsored Projects, UTEP

2007 - 2008 Outstanding Performance Award, Office of Research and Sponsored Programs, UTEP

2004 - 2004 Professor Y.C. Wong Visiting Lectureship, University of Hong Kong

1986 - 1987 Andrew Mellon Foundation Research Award, Institute of Population and Resource Studies, Stanford University

C. Contribution to Science

1. Virtual Labs Established for Finding Inversions in Nucleic Acid Sequences and Their Functions: With my early training and interests in designing efficient algorithms for identifying matches in multiple long molecular sequences, my research focus has been on DNA sequences of Herpesvirus genomes. The most significant contribution is the mathematical characterization of nonrandom clusters of palindromes (e.g., GCAATATTGC), which is a short DNA segment whose reverse complementary sequence is identical to itself. The probability of finding origins of replication around nonrandom clusters of palindromes has been shown to be higher. A software package called MOLPAC was developed for locating these replication origins, which are potential targets for developing vaccines against the growth and spread of viruses. Palindromes are special cases of the more general patterns of inversions in RNA sequences. Each inversion is a palindrome with a gap between the two complimentary stem sequences. With recent outbreaks of RNA viruses (e.g., Coronaviruses, West Niles), our attention has shifted to viruses with RNA molecules as their genomes. Inversions in these RNA molecules have been found to be involved in the formation of stem loops and pseudoknots, sequence patterns important for the formation of their secondary structures and functioning of the viral genomic sequences. Therefore, the Ribonucleic Acid Virtual Laboratory (RNAVLab, rnavlab.utep.edu) has been established for providing a series of software applications and online databases for analyses of RNA secondary structures. My group continues to find new approaches, such as AT excursion and least-squares support vector machine, to predict the locations of these replication origins more accurately, facilitating the efforts to find targets of vaccine development with less experimentation.

    a. Zhang B, Yehdego DT, Johnson KL, Leung MY, Taufer M. Enhancement of accuracy and efficiency for RNA secondary structure prediction by sequence segmentation and MapReduce. BMC Struct Biol. 2013;13 Suppl 1(Suppl 1):S3. PubMed Central PMCID: PMC3952952.
    b. Taufer M, Leung MY, Solorio T, Licon A, Mireles D, Araiza R, Johnson KL. RNAVLab: A virtual laboratory for studying RNA secondary structures based on grid computing technology. Parallel Comput. 2008 Nov 1;34(11):661-680. PubMed Central PMCID: PMC2714649.
    c. Chew DS, Choi KP, Leung MY. Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses. Nucleic Acids Res. 2005 Sep 1;33(15):e134. PubMed Central PMCID: PMC1197138.
    d. Leung MY, Choi KP, Xia A, Chen LH. Nonrandom clusters of palindromes in herpesvirus genomes. J Comput Biol. 2005 Apr;12(3):331-54. PubMed Central PMCID: PMC4032367.

2. Databases for RNA Pseudoknots and G Protein-Coupled Receptors (GPCR): As an extension to our RNAVLab, we have included a standalone database for analyses of RNA secondary structures called

PseudoBase++ (pseudobaseplusplus.utep.edu) with faster algorithms implemented as an updated version of the original PseudoBase. It is a searchable database of RNA sequences containing stem loops or pseudoknots, wrapped by a versatile, user-friendly interface providing scientists with a powerful engine to access, search, select, and sort data based on different fine-grained criteria. The PseudoBase++ interface allows scientists to visualize selected structures with PseudoViewer, to map existing sequences to GenBank, and to insert new pseudoknots to the PseudoBase dataset through a syntax-controlled interface that prevents structural error for long sequences. We have also developed a pipeline called GPCR Prediction Ensemble (GPCR-PEn), available at gpcr.utep.edu with an online database GPCR-PEnDB that contains more than 3000 confirmed GPCRs and 3500 non-GPCRs from more than 1200 different organisms including bacteria and viruses. It also incorporates predicted GPCRs from three organisms. Each protein, with a unique identification number, is linked to its source organism, gene name, protein name, sequence length, and other features such as amino acid and dipeptide compositions. Recently, as we investigate GPCR-ligand binding characteristics, structural data of GPCRs and their ligands are being incorporated into CPCR-PEnDB as well.

a. Gadad B, Medrano J, Ramos E, Ruiz-Velasco A, Leung M, Thompson P. 486. Protein-Coding Transcriptome in Major Depressive Disorder and Suicidality: Distinct Pathways Regulating Myelination, Lipid Biogenesis, and Extracellular Matrix Pathways. Biological Psychiatry. 2023 May; 93(9):S291-. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0006322323008004 DOI: 10.1016/j.biopsych.2023.02.726
b. Dankwah KO, Mohl JE, Begum K, Leung MY. What Makes GPCRs from Different Families Bind to the Same Ligand?. Biomolecules. 2022 Jun 21;12(7) PubMed Central PMCID: PMC9313020.
c. Munoz S, Guerrero FD, Kellogg A, Heekin AM, Leung MY. Bioinformatic prediction of G protein-coupled receptor encoding sequences from the transcriptome of the foreleg, including the Haller's organ, of the cattle tick, Rhipicephalus australis. PLoS One. 2017;12(2):e0172326. PubMed Central PMCID: PMC5322884.
d. Taufer M, Licon A, Araiza R, Mireles D, van Batenburg FH, Gultyaev AP, Leung MY. PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. Nucleic Acids Res. 2009 Jan;37(Database issue):D127-35. PubMed Central PMCID: PMC2686561.

3. RNA-seq and Next-Generation DNA Sequencing Data Analytics Using High-Performance Computing (HPC): A series of JAVA-based applications and their upgrades have been released in the last few years, e.g., InversFinder 2.0, Segmenta 2.0, and the complete bundle of RNASSA 2.0 for RNA secondary structure analysis has been made available online with the latest version with updates since 2015. The core of this software is a new RNA segmentation algorithm based on optimal cuts between inversion clusters along the RNA sequence, relying on the mathematical theory of excursion. With the use of high-throughput grid computing across a network of computers (Bioinformatics Grid) managed by the HTCondor software for task scheduling, we have been able to reduce the computing time from days to a few minutes for structure prediction of RNA over 3000 bases. To preprocess very large datasets efficiently for our OncoMiner pipeline (oncominer.utep.edu) implemented to help biomedical researchers analyze genomic sequence variants in patients with cancer, the preprocessing program for parsing data files has been parallelized on local HPC systems and the Blue Waters system at the National Center for Supercomputing Applications.

a. Patil AR, Leung MY, Roy S. Identification of Hub Genes in Different Stages of Colorectal Cancer through an Integrated Bioinformatics Approach. Int J Environ Res Public Health. 2021 May 23;18(11) PubMed Central PMCID: PMC8197092.
b. Vasquez M, Mohl J, Leung M. Parsing Next Generation Sequencing Data in Parallel Environments for Downstream Genetic Variation Analysis. The Journal of Computational Science Education. 2018 December; 9(2):37-45. Available from: https://www.jocse.org/articles/9/2/5/ DOI: 10.22369/issn.2153-4136/9/2/5

c. Leung M. Scan Statistics Applications in Genomics. In: Glaz J, Koutras M, editors. Handbook of Scan Statistics [Internet] New York, NY: Springer New York; 2017. Chapter Chapter 42-11-26p. Available from: http://link.springer.com/10.1007/978-1-4614-8414-1_42-1 DOI: 10.1007/978-1-4614-8414-1_42-1

d. Cruz-Cano R, Lee ML, Leung MY. Logic minimization and rule extraction for identification of functional sites in molecular sequences. BioData Min. 2012 Aug 16;5(1):10. PubMed Central PMCID: PMC3492099.

4. Extension of the Sequence Segmentation Algorithm for Other Computationally Intensive Problems. With the Translational Bioinformatics Lab and Structural Bioinformatics Lab housing the Bioinformatics Grid with videoconferencing capability established in 2012, we have initiated collaborative projects requiring computationally-intensive tasks and data transfer between remote sites. In collaboration with Dr. Gerken at Case Western Reserve University on a project entitled “Initiation and Regulation of Mucin Type O-Glycosylation” with specific aims to understand the processes governing O-glycan site selection and O-glycan elongation in order to address the molecular mechanisms and biology of O-glycosylation. We have already constructed a working prototype of a web-based software tool ISOGlyP (Isoform Specific O-Glycosylation Prediction), for predicting O-glycosylation sites in amino acid sequences, available at isoglyp.utep.edu. Using a more sophisticated conceptual model on the Bioinformatics Grid with the high-throughput HTCondor system, our ongoing effort is to extend the current version by incorporating recent data on the N- or C-terminal targeting of previously glycosylated peptide substrate. Other projects include the application of super-resolution techniques in improving mammograms and the implementation of a pipeline for predicting GPCRs (G-protein coupled receptors) using hidden Markov models, and the development of bioinformatics tools for ecoinformatics studies.

a. Grant AH, Rodriguez AC, Rodriguez Moncivais OJ, Sun S, Li L, Mohl JE, Leung MY, Kirken RA, Rodriguez G. JAK1 Pseudokinase V666G Mutant Dominantly Impairs JAK3 Phosphorylation and IL-2 Signaling. Int J Mol Sci. 2023 Apr 6;24(7) PubMed Central PMCID: PMC10095075.

b. Gadad B, Medrano J, Diaz-Pacheco V, Ramos E, Yang B, Leung M, Mellios N, Jha M, Trivedi M, Thompson P. Whole-Transcriptome Brain Expression and Exon-Usage Profiling in Major Depression and Suicide. Biological Psychiatry. 2021 May; 89(9):S120-S121. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0006322321004157 DOI: 10.1016/j.biopsych.2021.02.311

c. Mohl JE, Gerken T, Leung MY. Predicting mucin-type O-Glycosylation using enhancement value products from derived protein features. J Theor Comput Chem. 2020 May;19(3) PubMed Central PMCID: PMC7671581.

d. Rivas JA Jr, Mohl J, Van Pelt RS, Leung MY, Wallace RL, Gill TE, Walsh EJ. Evidence for regional aeolian transport of freshwater micrometazoans in arid regions. Limnol Oceanogr Lett. 2018 Aug;3(4):320-330. PubMed Central PMCID: PMC6284810.