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: Ming-Ying Leung

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

POSITION TITLE: Professor of Mathematical Sciences

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)*	Completion Date MM/YYYY	FIELD OF STUDY
University of Hong Kong	B.Sc.	1980	Mathematics
University of Hong Kong	M.Phil.	1983	Mathematics
Stanford University	M.S.	1988	Computer Science
Stanford University	Ph.D.	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 ALL (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 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 20 years, my research has been funded by NIH, NSF, IBM, and the Texas NHARP 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 project, I will manage the bioinformatics research support component of the Bioinformatics and Biostatistics Unit in the Research Infrastructure Core, implementing and maintaining necessary bioinformatics computing tools, including searchable databases and automated software pipelines for mining large datasets, for BBRC researchers and the biomedical research community.

1. Leung, M.Y., Burge, C., Blaisdell, B.E., and Karlin, S., (1991) An Efficient Algorithm for Identifying Matches with Errors in Multiple Long Molecular Sequences, *J. Mol. Biol.* 221, 1367-1378. **PMCID: PMC4076298**
2. Leung, M.Y., Marsh, G.M., and Speed, T.P. (1996) Over and Under Representation of Short Oligonucleotides in Herpes Virus Genomes. *J. Computational Biology,* 3(3), 345-360. **PMCID: PMC4076300**

3. Chew, D.S.H., Choi, K.P., Heidner, H., and Leung, M.Y. (2004) Palindromes in SARS and Other Coronaviruses, *INFORMS J. Comp.* 16(4), 331-340. **PMCID: PMC4066412**
4. Leung, M.Y., Knapka, J.A., Wagler, A.E., Rodriguez, G., Kirken, R.A. (2016) OncoMiner: A pipeline for bioinformatics analysis of exonic sequence variants in cancer. In: Big Data Analytics in Genomics, Wong, K.C. (Ed.), pp. 373-396, Springer, New York. Available at link.springer.com/chapter/10.1007/978-3-319-41279-5_12. Accessed 2/21/2019.

B. Positions and Honors

Positions and Employment

1980-1983 : Teaching Assistant, Department of Mathematics, University of Hong Kong
1982-1983 : Lecturer, Department of Extramural Studies, University of Hong Kong
1983-1989 : Research Assistant and Teaching Fellow, Department of Mathematics, Stanford University
1989-2003 : Assistant and Associate Professor, Division of Mathematics and Statistics, The University of Texas at San Antonio
1993 : Visiting Research Fellow, Department of Statistics, University of California at Berkeley and Department of Pharmaceutical Chemistry, University of California at San Francisco
2001-2002 : Visiting Associate Professor, Department of Statistics, Rice University
2003- : Professor, Department of Mathematical Sciences, and Director, Bioinformatics Program, The University of Texas at El Paso
2013- : Director, Computational Science Program, The University of Texas at El Paso

Other Experience and Professional Memberships

1990-1991 : Consultant for the mathematical molecular biology groups at the University of Southern California and Stanford University
2002-2003 : Editorial Board Member, Advances and Applications in Statistics
2002-2004 : Co-chair, Organizing Committee for the International Workshop on Statistical Methods in Microarray Data Analysis, Institute of Mathematical Sciences, National University of Singapore
2005 : Chair, Joint Session in Bioinformatics, 2005 INFORMS Annual Meeting
2005-2009 : External Advisory Board Member, NSF and HHMI funded "Talent Expansion in Quantitative Biology" project at East Tennessee State University
2007-2013 : Associate Editor, INFORMS Journal on Computing
2008 : Chair, Invited paper session in Stochastic Models for Biological Processes, International Workshop on Applied Probability, July 2008, Compiegne, France
2008-2010 : Member, The University of Texas System Computational Biology Workgroup for the Cancer Prevention and Research Institute of Texas
2009- : Member, NIH-RCMI Translational Research Network Translational Informatics Subcommittee
2010 : External Review Panelist for West Virginia IDeA Network for Biomedical Research Excellence, Research Competitiveness Program, American Association for the Advancement of Science
2010 : Chair, Session on Stochastic Models for Biological Systems, 2010 INFORMS Annual Meeting
2011 : Chair for invited Session Computational Methods in Biomolecular and Phylogenetic Analyses, International Federation of Operational Research Societies (IFORS) Conference, July 2011, Melbourne, Australia.
2011 : Chair for invited Cluster on Computational Biology, Institute for Operations Research and Management Science (INFORMS) Annual Conference, November 2011, Charlotte.
2012- : Organizer, Joint UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Sciences
2013- : Mentor, National Alliance for Doctoral Studies in the Mathematical Sciences
2016 : Member, Scientific Program Committee, International Workshop on Applied Probability, Toronto

Honors

1986-1987 : Andrew Mellon Foundation Research Award, Institute of Population and Resource Studies, Stanford University
2004 : Professor Y.C. Wong Visiting Lectureship, University of Hong Kong
2007-2008 : Outstanding Performance Award, Office of Research and Sponsored Programs, UTEP
2014-2015 : Outstanding Performance Award in Securing Extramural Funding, Office of Research and Sponsored Projects, UTEP
2017 : Student Choice Award for Outstanding Teaching, Department of Mathematical Sciences, UTEP

C. Contribution to Science

1. Palindromes in Searches for Replication Origins: 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. These replication origins are potential targets for developing vaccines against the growth and spread of viruses. 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. Leung, M.Y., Choi, K.P., Xia, A. and Chen, L.H.Y. (2005) Nonrandom Clusters of Palindromes in Herpesvirus Genomes, *J. Computational Biology* 12(3), 331-354. **PMCID: PMC4032367**
   b. Chew, D.S.H., Choi, K.P., and Leung, M.Y. (2005) Scoring Schemes of Palindrome Clusters for More Sensitive Prediction of Replication Origins in Herpesviruses, *Nucleic Acids Research* 33 (15), e134**. PMCID: PMC1197138**
   c. Chew, D.S.H., Leung, M.Y., and Choi, K.P. (2007) AT Excursion: a New Approach to Predict Replication Origins in Viral Genomes by Locating AT-rich Regions. *BMC Bioinformatics* 8, 163-174. **PMCID: PMC1904460**
   d. Cruz-Cano, R., Chew, D.S.H., Choi, K.P., and Leung, M.Y. (2010) Least-Squares Support Vector Machine Approach to Viral Replication Origin Prediction, *INFORMS J. Computing*, 22(3), 457-470. **PMCID: PMC2923853**

2. Establishment of RNA Virtual Lab with Databases for RNA Pseudoknots: 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., SARS, 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) has been established for providing a series of software applications and online databases for analyses of RNA secondary structures, such as PseudoBase++ (pseudobaseplusplus.utep.edu), with faster algorithms implemented as an extension to the original PseudoBase.
   a. Taufer, M., Leung, M.Y., Solorio, T., Licon, A., Mireles, D., and Johnson, K.L. (2008) RNAVLab: A Virtual Laboratory for Studying RNA Secondary Structures Based on Grid Computing Technology, *Parallel Computing* 34: 661-680. **PMCID: PMC2714649**
   b. Taufer, M., Licon, A., Araiza, R., Mireles, D., Gultyaev, A., Van Batenburg, F.H.D., and Leung, M.Y. (2009) PseudoBase++: An Extension of PseudoBase for Easy Searching, Formatting, and Visualization of Pseudoknots. *Nucleic Acids Research* 37(Database Issue):D127-135. **PMCID: PMC2686561**
   c. Licon, A., Taufer, M., Leung, M.Y., Johnson, K.L. (2010) A Dynamic Programming Algorithm for Finding the Optimal Segmentation of an RNA Sequence in Secondary Structure Predictions. In: *Proceedings of the 2nd International Conference on Bioinformatics and Computational Biology 2010 (BICoB-2010)*, Honolulu, Hawaii, pp.165-170. **PMCID: PMC4335647**
   d. Yehdego, D. T., Zhang, B., Kodimala, V. K.R., Johnson, K. L., Taufer, M., Leung, M.Y. (2013) Secondary Structure Predictions for Long RNA Sequences Based on Inversion Excursions and MapReduce. In: *Proceedings of the 12th IEEE International Workshop on High Performance Computational Biology (HiCOMB 2013)*, Boston, MA: pp. 1-10. Available at hicomb.org/ HiCOMB2013/papers/HICOMB2013-03.pdf. Accessed 2/21/2019.

5. RNA Structure and Next-Generation 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 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 using a multiprocessing approach.

a. Rosskopf, J.J.; Upton, J.H.III, Rodarte, L., Romero, T.A., Leung, M.Y., Taufer, M. and Johnson, K.L. (2010) A 3' terminal stem-loop structure in Nodamura virus RNA2 forms an essential cis-acting signal for RNA replication. *Virus Research* 150(1-2):12-21. **PMCID: PMC3017585**
b. Mohl, J., Licon, A., Viswakula, S., Kelley, P., Araiza, R., Kodimala, V., Vegesna, R., Saldivar, L., Yehadego, D., Cardenas, G., Vest, E., Taufer, M., Fuentes, O., Johnson, K. L., and Leung, M.Y. (2012) RNASSA 2.0: RNA Secondary Structure Analysis (Version 2.0.121208). Available at rnavlab.utep.edu/rnassa. Accessed 2/21/2019.
c. Zhang, B., Yehdego, D. T., Johnson, K. L., Leung, M.Y., Taufer, M., (2013). Enhancement of accuracy and efficiency for RNA secondary structure prediction by sequence segmentation and MapReduce. *BMC Structural Biology, 13 (Suppl. 1)* (S3), 1-24. Available at biomedcentral.com/1472-6807/13/S1/S3. Accessed 2/21/2019. **PMCID: PMC3952952**
d. Leung, M.Y. (2017) Scan Statistics Applications in Genomics. In: Handbook of Scan Statistics, Glaz, J. and Koutras, M.V. (Eds.); Springer, New York. Available at doi.org/10.1007/978-1-4614-8414-1_42-1. Accessed 2/21/2019.

3. 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. Cruz-Cano, R., Lee, M.-L. T., Leung, M.Y., (2012). Logic Minimization and Rule Extraction for Identification of Functional Sites in Molecular Sequences. *BioData Mining, 5*(10), 1-21. Available at biodatamining.org/content/5/1/10. Accessed 2/21/2019. **PMCID: PMC3492099**
b. Guerrero, F., Kellogg, A., Ogrey, A. N., Heekin, A. M., Barrero, R., Bellgard, M. I., Dowd, S. E., Leung, M.Y. (2016) Prediction of G protein-coupled receptor encoding sequences from the synganglion transcriptome of the cattle tick, *Rhipicephalus microplus*. *Ticks and Tick-borne Diseases*. 7(5), 670-677. Available at doi.org/10.1016/j.ttbdis.2016.02.014. Accessed 2/21/2019. **PMCID: PMC5138280**
c. Munoz, S., Guerrero, F., Kellogg, A., Heekin, A. M., Leung, M.Y. (2017) 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* 12(2):e0172326. Available at doi.org/10.1371/journal.pone.0172326. Accessed 2/21/2019. **PMCID: PMC5322884**
d. Rivas, J.A., Mohl, J.E., Van Pelt, R.S., Leung, M.-Y., Wallace, R.L., Gill, T.E. and Walsh, E.J. (2018) Evidence for regional aeolian transport of freshwater micrometazoans in arid regions. Limnology and Oceanography Letters. Available at doi.org/10.1002/lol2.10072. Accessed 2/21/2019.

D. Additional Information: Research Support and/or Scholastic Performance

Complete List of Published Work in MyBibliography

https://www.ncbi.nlm.nih.gov/myncbi/browse/collection/47215158/?sort=date&direction=ascending

Ongoing Research Support

Grant: 1U01GM113534-01 (Gerken) 12/1/2014 – 6/30/2019
Source: NIH/Case Western Reserve University
Title: Initiation and Regulation of Mucin Type O-Glycosylation
Major Goal: To study the specificity of the large family (20 members in humans) of polypeptide-GalNAc transferases (ppGalNAcTs) that initiate mucin-type O-glycosylation. The results of this work will provide specific information on the activities and biological roles of the enzymes that perform mucin-type glycosylation and will allow the eventual development of specific therapeutics for the treatment of a range of hormonal, metabolic, inflammatory, cardiovascular and neoplastic diseases.
Role: Subcontract PI

Grant: 2G12MD007592 (Kirken) 7/1/2014 – 3/31/2019
Source: NIH
Title: Border Biomedical Research Center
Major Goal: Achieving the RCMI-BBRC goal and objectives will help create a healthier Paso de Norte border region by spearheading, sustaining and leveraging biomedical research and practice, and increasing the work force to develop countermeasures to health problems concentrated in the area but of national consequence, and provide a highly competitive faculty and training setting at UTEP to prepare minority students for entry into the biomedical research mainstream of the nation.
Role: Director, Bioinformatics Core Facility

Completed Research Support

Grant: 1R15AI105823-01A1 (Johnson) 4/1/2014 – 3/31/2017
Source: NIH
Title: Mechanisms in Viral RNA Replication Complex Assembly: Novel Targets For Antivira
Major Goal: This study will uncover essential processes in viral RNA replication common to many pathogenic viruses and will identify new targets for antiviral therapies.
Role: Co-I

Grant: 2012-38422-19910 (Leung) 9/1/2012 – 8/31/2016
Source: USDA
Title: Bioinformatics Education for Agricultural Science
Major Goal: The major goals of this project are to support undergraduates and graduate students, and train them to conduct bioinformatics analysis of data from USDA research. The project will generate educational materials packaged in the form of web-based modules that can be readily utilized for lecture presentation and class projects in bioinformatics-related courses in biology, biochemistry, computer science, and mathematics.
Role: PI

Grant: DUE 0966151 (Pannell) 9/1/2010 – 8/31/2016
Source: NSF
Title Recruiting and Keeping Undergraduate Students in the Sciences
Major Goal: The major goal of this project is to offer scholarships to talented freshmen majoring in biology, chemistry, mathematics, geology, or physics.
Role: Co-PI

Grant: DUE 0926721 (Leung) 9/1/2009 – 8/31/2016
Source: NSF Interdisciplinary Training for Undergraduates in Biological and Mathematical Sciences (UBM)
Title: UBM Institutional: Undergraduate Training in Bioinformatics
Major Goal: The major goal of this project is to establish an undergraduate training program of bioinformatics research at UTEP.
Role: PI