Yun William Yu

Contact Information

Carnegie Mellon University / School of Computer Science
Ray & Stephanie Lane Computational Biology Department
Gates-Hillman 7703
4902 Forbes Ave
Pittsburgh, PA, 15213
United States of America

Mobile: +1.812.250.1248
E-mail: ywyu@cmu.edu
WWW: yunwilliamyu.net

Research Mission

I am generally interested in developing novel algorithms for bioinformatics applications and translating existing tools from the math/CS literature to biology. My primary research tools include probabilistic sketches and compressive algorithms, which I have applied to biological database search, medical privacy, and metagenomic classification.

Education

Ph.D. Mathematics, Massachusetts Institute of Technology, 2017
Dissertation: *Compressive algorithms for search and storage in biological data*
Advisor: Bonnie Berger

MPhil Applied Mathematics, Imperial College London, 2014
Thesis: *Local, multi-resolution detection of network communities by Markovian dynamics*
Supervisors: Mauricio Barahona & Sophia Yaliraki

MRes Biomedical Physical Chemistry, Imperial College London, 2010
Thesis: *Protein Dynamics: Applications of Graph Theory and Community Detection*
Supervisors: Sophia Yaliraki & Mauricio Barahona

B.S. Mathematics, Indiana University, 2009

B.A. Chemistry & Germanic Studies, Indiana University, 2009

Professional Experience

Assistant Professor, Computational Biology, Carnegie Mellon University, since August 2023.
Assistant Professor, Applied Mathematics, University of Toronto, July 2019-August 2023.
HMS BIRT T15 Fellow, Weber Group, DBMI, Harvard Medical School, June 2018-June 2019.
Research Fellow, Weber Group, DBMI, Harvard Medical School, June 2017-May 2018.
Senior Intern, D.E. Shaw Research, Summer 2016
Software Engineering Intern, Ads Data Infrastructure, Google, Summer 2015

Featured Work

Shaw J and **Yu YW** (2023). Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nature Methods*, 20(11): 1661-1665.

By combining theoretical insights into k-mer subsampling with efficient sparse chaining, we built skani, a tool for quickly determining average nucleotide identity (ANI) and align fraction (AF) between metagenome-aligned genomes (MAGs). skani is 20x faster than the prior state-of-the-art FastANI, and has better robustness against contamination and incompleteness of MAGs.

Shaw J and **Yu YW** (2023). Proving sequence aligners can guarantee accuracy in almost O(m log n) time through an average-case analysis of the seed-chain-extend heuristic. *Genome Research*, 33(7): 1175-1187.

Worst-case time complexity for string alignment is known to have a quadratic lower bound. However, in practice, there are many fast heuristics that in practice are much faster. In this manuscript, we prove that under a suitable fixed random string substitution model for related strings, the popular seed-chain-extend heuristic has an expected runtime that is $O(m \log n)$, where $m \leq n$ are the string lengths.

Yu YW and Weber GM (2022). HyperMinHash: MinHash in LogLog space. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):328-339.

The MinHash probabilistic sketch was invented to approximate Jaccard index, a measure of set similarity. Using any min-wise independent hash function, assuming no hash collisions, the probability that the minimum hash values of two sets are equal is exactly the Jaccard index. MinHash thus appears to require $O(\log n)$ bits in the hash output to ensure that sets of order n to not have any hash collisions. However, we noticed that MinHash does not require the hash function to be collision-free; instead, only the minimum values have to be free from collision.

Thus, we prove in this manuscript that by using a floating-point encoding, we can reduce the space complexity of storing the minimum hashes to $O(\log\log n)$.

Peer-reviewed Publications

Yu YW (in press). On minimizers and convolutional filters: theoretical connections and applications to genome analysis. *Journal of Computational Biology*. arXiv:2111.08452 [cs.LG]

Leighton AT and **Yu YW** (accepted). Secure Federated Boolean count queries using fully-homomorphic cryptography. *Research in Computational Molecular Biology (RECOMB)*. bioRxiv:10.1101/2021.11.10.46809

Shaw J and **Yu YW** (2023). Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nature Methods*, 20(11): 1661-1665.

Staniscia L and **Yu YW** (2023). Image-centric compression of protein structures improves space savings. *BMC Bioinformatics*, 24(1): 437.

Shaw J and **Yu YW** (2023). Proving sequence aligners can guarantee accuracy in almost O(m log n) time through an average-case analysis of the seed-chain-extend heuristic. *Genome Research*, 33(7): 1175-1187. An earlier version was presented at RECOMB 2023.

Ippolito D, Carlini N, Lee K, Nasr M, and **Yu YW** (2023). Reverse-Engineering Decoding Strategies Given Blackbox Access to a Language Generation System. In *Proceedings of the 16th International Natural Language Generation Conference (INLG)*, 396–406. **Alphabetical author ordering**

Deol K, Weber GM, **Yu YW** (2022). SlowMoMan: A web app for discovery of important features along user-drawn trajectories in 2D embeddings. bioRxiv:10.1101/2022.08.23.505019. Accepted for proceedings/oral presentation at GIW/ISCB-Asia 2022 joint conference.

Yu YW and Weber GM (2022). HyperMinHash: MinHash in LogLog space. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):328-339. doi: 10.1109/TKDE.2020.2981311

Shaw J and **Yu YW** (2022). Theory of local k-mer selection with applications to long-read alignment. *Bioinformatics*, 38(20):4959-4669.

Shaw J and **Yu YW** (2022). flopp: Extremely Fast Long-Read Polyploid Haplotype Phasing by Uniform Tree Partitioning. *Journal of Computational Biology*, 29(2): 195-211. An earlier version was presented at RECOMB 2021.

Koppel J and **Yu YW** (2022). Skiing is Easy, Gymnastics is Hard: Complexity of Routine Construction in Olympic Sports. *11th International Conference on Fun with Algorithms*, Leibniz International Proceedings in Informatics (LIPIcs), 226(17):1-20.

Tao Z, Weber GM, **Yu YW** (2021). Expected 10-anonymity of HyperLogLog sketches for federated queries of clinical data repositories, *Bioinformatics*, 37(1), i151-i160. doi: 10.1093/bioinformatics/btab292. Also a proceedings talk at ISMB/ECCB 2021.

Bengio Y, Ippolito D, Janda R, Jarvie M, Prud'homme B, Rousseau JF, Sharma A, **Yu YW** (2021). Inherent privacy limitations of decentralized contact tracing, *Journal of the American Medical Informatics Association*, 28(1), 193-195. doi: 10.1093/jamia/ocaa153

Berger B, Waterman M, **Yu YW** (2021). Levenshtein Distance, Sequence Comparison and Biological Database Search. *IEEE Transactions on Information Theory*, 67(6), 3287-3294 doi: 10.1109/TIT.2020.2996543. **Alphabetical author ordering**

Yu YW and Weber GM (2020). Balancing Accuracy and Privacy in Federated Queries of Clinical Data Repositories: Algorithm Development and Validation, *Journal of Medical Internet Research*, 22(11), e18735. doi: 10.2196/18735

Bengio Y, Janda R, **Yu YW**, Ippolito D, Jarvie M, Pilat D, Struck B, Krastev S, Sharma A (2020). The need for privacy with public digital contact tracing during the COVID-19 pandemic, 2(7), E342-E344. *The Lancet Digital Health*, doi: 10.1016/S2589-7500(20)30133-3

Nazeen S, **Yu YW**, Berger B (2020). Carnelian uncovers hidden functional patterns across diverse study populations from whole metagenome sequencing reads. *Genome Biology*, 21(1), 1-18.

Luo Y*, **Yu YW***, Zeng J, Berger B, Peng J (2019). Metagenomic binning through low density hashing. *Bioinformatics*, 35(2), 219-226. doi:10.1093/bioinformatics/bty611. ***Joint first-author**

Orenstein Y, **Yu YW**, Berger B (2018). Joker de Bruijn: Covering k-Mers Using Joker Characters. *Journal of Computational Biology*, 25(11), 1171-1178.

Demaine ED, Demaine ML, Eisenstat S, Hesterberg A, Lincoln A, Lynch J, **Yu YW** (2017). Total Tetris: Tetris with Monominoes, Dominoes, Trominoes, Pentominoes,... *Journal of Information Processing*, 25, 515-527. **Alphabetical author ordering**

Berger B, Daniels NM, **Yu YW** (2016). Computational biology in the 21st century: Algorithms that scale. *Communications of the ACM*, 59(8): 72–80. **Cover article, alphabetical author ordering**

Yorukoglu D, **Yu YW**, Peng J, Berger B (2016). Compressive Mapping for Next-generation Sequencing. *Nature Biotechnology*, 34(4): 374–376.

Shajii A, Yorukoglu D, **Yu YW**, Berger B (2016). Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics*, 32(17): i538–i544. Also was an oral presentation/proceedings paper at ISMB/ECCB 2015.

Yu YW*, Daniels NM*, Danko DC, Berger B (2015). Entropy-scaling search of massive biological data. *Cell Systems*, 1(2): 130–140. **Cover article, *joint first-author**

Yu YW, Yorukoglu D, Peng J, Berger B (2015). Quality score compression improves genotyping accuracy. *Nature Biotechnology*, 33(3): 240–243.

Yu YW, Yorukoglu D, Berger B (2014). Traversing the k-mer Landscape of NGS Read Datasets for Quality Score Sparsification, *Research in Computational Molecular Biology*, 385–399.

Tan B, O'Dell DK, **Yu YW**, Monn MF, Hughes HV, Burstein S, Walker JM (2010). Identification of endogenous acyl amino acids based on a targeted lipidomics approach. *Journal of Lipid Research*, 51, 112–119.

Tan B, **Yu YW**, Monn MF, Hughes HV, O'Dell DK, Walker JM (2009) Targeted lipidomics approach for endogenous N-acyl amino acids in rat brain tissue, *Journal of Chromatography B*, 877(26): 2890–2894.

Tan B, Bradshaw HB, Rimmerman N, Srinivasan H, **Yu YW**, Krey JF, Monn MF, Chen JS, Hu SS, Pickens SR, Walker JM (2006). Targeted lipidomics: discovery of new fatty acyl amides, *The AAPS Journal*, 8(3), 461–465.

Other articles and preprints (unrefereed)

Shaw J, Gounot JS, Chen H, Nagarajan N, **Yu YW** (2024). Floria: Fast and accurate strain haplotyping in metagenomes. bioRxiv:10.1101/2024.01.28.577669

Shaw J and **Yu YW** (2023). Skani enables accurate and efficient genome comparison for modern metagenomic datasets. *Nature Methods* (Research Briefing), 20: 1633-1634.

Zheng A, Shaw J, and **Yu YW** (2022). Mora: abundance aware metagenomic read re-assignment for disentangling similar strains. bioRxiv:10.1101/2022.10.18.512733

Yu YW, Delvenne J-C, Yaliraki SN, Barahona M (2020). Severability of mesoscale components and local time scales in dynamical networks. arXiv:2006.02972 [physics.soc-ph]

Hammoud A and **Yu YW** (2020). Privacy-accuracy trade-offs in noisy digital exposure notifications. arXiv:2011.03995 [cs.CR]

Alsdurf H, Bengio Y, Deleu T, Gupta P, Ippolito D, Janda R, Jarvie M, Kolody T, Krastev S, Maharaj T, Obryk R, Pilat D, Pisano V, Prud'homme B, Qu M, Rahaman N, Rish I, Rousseau JF, Sharma A, Struck B, Tang J, Weiss M, **Yu YW** (2020). COVI White Paper. arXiv:2005.08502. [cs.CR]

Cho H, Ippolito D, **Yu YW** (2020). Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. arXiv:2003.11511. [cs.CR]

Yu YW (2015). Approximation hardness of Shortest Common Superstring variants. arXiv:1602.08648. [cs.CC]

Yu YW. Appendix to the paper "Representation Varieties of Fuchsian Groups" by M. Larsen & A. Lubotzky (from *Fourier Analysis and Number Theory to Radon Transforms and Geometry* (2012) pp 375-398) arXiv:1203.3408 [math.AG]

Conference Talks not listed elsewhere

Deol K, Weber GB, **Yu YW**. *SlowMoMan: A web app for discovery of important features along user-drawn trajectories in 2D embeddings*. December 12, 2022. GIW XXXI/ISCB-Asia V (International Conference on Genome Informations and International Society of Computational Biology-Asia Joint conference), Tainan, Taiwan.

Leighton A, **Yu YW**. *Secure Federated Aggregate-Count Queries on Medical Patient Databases Via Fully-Homomorphic Cryptography*. July 12, 2022. Intelligent Systems for Molecular Biology (ISMB), Madison, WI, USA.

Berger B, **Yu YW**. *Insights from compression for biological data analysis*. July 25, 2019. Intelligent Systems for Molecular Biology (ISMB) Genome Privacy and Security Session, Basel, Switzerland.

Yu YW. *HyperMinHash: MinHash in LogLog space*. July 7, 2018. Intelligent Systems for Molecular Biology (ISMB), Chicago, IL.

Yu YW. *Entropy-scaling Search of Massive Biological Data*. September 5, 2016. European Conference on Computational Biology (ECCB), the Hague.

Yu YW. *Traversing the k-mer Landscape of NGS Read Datasets for Quality Score Sparsification*. April 3, 2014. Research in Computational Molecular Biology (ReCOMB), Philadelphia.

Invited talks

Yu YW. *Augmenting k-mer sketching for (meta)genomic sequence comparisons*. October 26, 2023. Pennsylvania State University Computer Science and Engineering Department Colloquium.

Yu YW. *Augmenting k-mer sketching for (meta)genomic sequence comparisons*. October 18, 2023. Massachusetts Institute of Technology Bioinformatics seminar.

Yu YW. *Visualizing and operating on high-dimensional biological data*. December 19, 2022. Chung Yuan Christian University Information and Computer Engineering Graduate Seminar.

Yu YW. *Compressive Hash-Based Feature Selection in Bio(Medical) Informatics*. April 25, 2022. CMU Computational Biology Seminar.

Yu YW. *Compressive Hash-Based Feature Selection in Bio(Medical) Informatics*. April 8, 2022. Toyota Technological Institute of Chicago Special Seminar.

Yu YW. *On the relationship between MinHash, Minimizers, and CNNs: an exploration through hash functions*. February 25, 2022. University of Rhode Island CS Seminar.

Yu YW. *The shape and shaping of data: a genomics tour*. November 5, 2021. Toronto Data Workshop.

Yu YW. *Distributed clinical queries: trade-offs in privacy, accuracy, and computation*. October 23, 2019. University of Toronto Math Departmental Colloquium.

Yu YW. *Compression for biological data analysis*. February 15, 2019. Indiana University Bloomington SICE CS Colloquium Series.

Yu YW. *Compression for biological data analysis*. February 1, 2019. Stanford ISL Colloquium and IT-Forum.

Yu YW. *Compression for biological data analysis*. January 16, 2019. University of Toronto Math Special Colloquium.

Yu YW. *From non-uniform data distributions to asymptotic speed and memory improvements: compressive genomics and HyperMinHash*. March 14, 2018. Brown CCMD and CS Seminar.

Yu YW. *From non-uniform data distributions to asymptotic speed and memory improvements: compressive genomics and HyperMinHash*. February 12, 2018. CMU Computational Biology Seminar.

Yu YW. *Compressive Metagenomics: Scaling Faster than Light*. June 2, 2017. Berger group Work-in-Progress presentation for the MIT Center for Microbiome Informatics and Therapeutics.

Yu YW. *Using insights from fractal geometry to accelerate similarity search in biological data*. April 6, 2016. MIT, Student Colloquium for Undergraduates in Mathematics (SCUM).

Yu YW. *Shortest Common Superstring and friends: approximation hardness*. February 12, 2015. MIT, Simple Person's Applied Math Seminar (SPAMS).

Yu YW. *Exploiting structure in biological datasets to accelerate similarity search*. September 25, 2014. MIT, Simple Person's Applied Math Seminar (SPAMS).

Yu YW. *Mesoscale structures and the coexistence of time scales in dynamical networks*. May 1, 2014. MIT, Simple Person's Applied Math Seminar (SPAMS).

Yu YW. *Local, multi-resolution detection of network communities by Markovian dynamics*. March 10, 2014. Universite Catholique Louvain, Large Graphs and Networks Seminar.

Yu YW. *Traversing the k-mer Landscape of NGS Read Datasets for Quality Score Sparsification*. October 24, 2013. MIT, Simple Person's Applied Math Seminar (SPAMS).

Grants

NSERC Discovery Grants Program / Data compression for biomedical data analysis / 2022-April-01 to 2027-March-31 / 145,000 CAD

DND/NSERC Discovery Grants Supplement / Data compression for biomedical data analysis / 2022-April-01 to 2025-March-31 / 120,000 CAD

NSERC Discovery Launch Supplement / Data compression for biomedical data analysis / 2022-April-01 to 2023-March-31 / 12,500 CAD

2020-2021 Connaught New Researcher Award / Development of long-read haplotype phasing methods and theory / 2021-April-01 to 2023-March-31 / 20,000 CAD

TCAI (Toronto COVID-19 Action Initiative) / Privacy-preserving contact tracing app / 2020-April-16 to 2021-April-15 / 50,000 CAD

University of Toronto at Scarborough Center for Teaching and Learning / 2019-Nov-01 to 2020-Oct-31 / 166 CAD

Teaching

02-510/02-710: Computational Genomics: Carnegie Mellon University, Spring 2024
Co-Instructor (w/ Prof. Jian Ma), 1 lecture section, 50 students
A graduate/advanced undergraduate-level survey course covering a mix of techniques used in modern computational genomics, designed for our PhD Computational Biology students, but also an option for certain undergraduate majors.

15-351/15-650/02-613: Algorithms and advanced data structures:
Carnegie Mellon University, Fall 2023
Instructor, 1 lecture section, 100 students
A masters/advanced undergraduate-level algorithms course, designed for our MS Computational Biology students, but also an option for certain undergraduate majors.

MATC58: An Introduction to Mathematical Biology: University of Toronto, Winter 2023
Instructor, 1 lecture section, 4 students
A 3rd-year undergraduate course covering biological modelling using difference and differential equations, taught using Linda Allen's textbook of the same title. I redesigned the course as a flipped classroom.

MATA35: Calculus II for Biological Sciences: University of Toronto, Winter 2023
Instructor, 1 lecture section, 257 students
A 1st-year undergraduate course covering basic integration, linear algebra, multivariable calculus, ordinary differential equations, and biological modelling.

MATA02: The Magic of Numbers: University of Toronto, Winter 2022
Instructor, 1 lecture section, 125 students
A 1st-year undergraduate distribution requirement course for non-majors covering basic number theory, discrete math, and building up to RSA encryption.

MAT1841: Mathematics of Massive Data Analysis: University of Toronto, Fall 2021
Instructor, 1 lecture section, 16 students
A topics course I designed on the mathematics of data analysis, with a focus on the shape of high-dimensional data. Topics including concentration inequalities, Markov chains, hash functions, streaming/sketching algorithms, random graph theory, percolation theory, wavelets, and computational topology.

MATA35: Calculus II for Biological Sciences: University of Toronto, Summer 2021
Instructor, 1 lecture section, 45 students
A 1st-year undergraduate course covering basic integration, linear algebra, multivariable calculus, ordinary differential equations, and biological modelling.

MATC58: An Introduction to Mathematical Biology: University of Toronto, Winter 2021
Instructor, 1 lecture section, 9 students
A 3rd-year undergraduate course covering biological modelling using difference and differential equations, taught using Linda Allen's textbook of the same title. I redesigned the course as a flipped classroom.

MAT1850: Linear Algebra and Optimization: University of Toronto, Fall 2020
Instructor, 1 lecture section, 16 students
A new graduate core course I designed covering advanced methods in linear algebra and introducing the theory of optimization.

MAT1801: Methods of Applied Mathematics 2: University of Toronto, Winter 2020
Instructor, 1 lecture section, 7 students
A topics course I designed on the mathematical underpinnings of modern data science. Topics included Markov chains, streaming/sketching algorithms, high-dimensional space, matrix factorization, random graph theory, percolation theory, wavelets, and computational topology.

MATB44: Differential Equations 1: University of Toronto, Fall 2019
Instructor, 1 lecture section, 173 students
A 2nd-year introduction to differential equations, taught using a combination of Tenenbaum & Pollard and Teschl's textbooks.

18.O95: Mathematics Lecture Series: MIT, IAP 2016
Course organizer, 1 lecture section, 1 recitation, 17 students, 16 listeners.
A topics course with rotating lectures by faculty members on various topics of interest, with an associated recitation covering problems on each topic.

BE3-HMIT Modelling in Biology: Imperial College London, Fall 2010
Teaching Assistant, lab section of 70 students with 4 TAs

A 3rd-year Bioengineering course on modelling dynamical systems using Matlab

BIOL-L311 Genetics: Indiana University Bloomington, Fall 2007 & Fall 2008
Teaching Intern, learning group sections of 12-15 students
A 3rd-year introduction to genetics.

Awards & Honors

Connaught New Researcher Award, University of Toronto, 2020-2021
Charles W and Jennifer C Johnson Prize, MIT Math, 2016
Fannie and John Hertz Foundation Fellowship, 2012-2017
NSF Graduate Research Fellowship Honorable Mention, 2010
Lord Porter Prize, Imperial College Chemical Biology Centre, 2010
Imperial British Marshall Scholarship, 2009-2011
Barry M Goldwater Scholarship, 2008-2009
Herman B Wells Scholarship, 2005-2009

Service

Graduate admissions committee, Carnegie Mellon University Schoolf of Computer Science Ray and Stephanie Lane Computational Biology Department, 2023-2024
Program Committee, Intelligent Systems and Molecular Biology (ISMB) conference 2020-2024
Program Committee, Research in Computational Molecular Biology (RECOMB) conference, 2021-2024
Organizer, University of Toronto Analysis and Applied Math Seminar Series, 2019-2023
Applied math faculty hiring committee, University of Toronto Department of Mathematics, 2022-2023
Computer and Data Security Committee Chair, University of Toronto Department of Mathematics, 2021-2022
Diversity and Equity Committee, University of Toronto Department of Mathematics, 2020-2021
Math teaching stream faculty hiring committee, University of Toronto at Scarborough Department of Computer and Mathematical Sciences, 2020-2021
Graduate Committee, University of Toronto Department of Mathematics, 2019-2020
Reviewer, *Proceedings of the National Academy of Sciences*, 2022
Reviewer, *Nature Biotechnology*, 2020
Reveiwer, *Institut national de la santé et de la recherche médicale*, 2020
Reviewer, *American Journal of Preventative Medicine*, 2020
Reviewer, *Journal of Chemical Information and Modeling*, 2020
Reviewer, *IEEE Transactions on Information Theory*, 2020
Reviewer, *Patterns*, 2020
Reviewer, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020
Reviewer, *Bioinformatics*, 2017, 2018

Professional Affiliations

Member, American Medical Informatics Association, since 2018
Member, i2b2 tranSMART Foundation, since 2018
Member, International Society for Computational Biology, since 2014
Member, American Mathematical Society, 2012-2017