Yun William Yu

CONTACT INFORMATION

Harvard Medical School
Department of Biomedical Informatics
Countway Library 3rd Floor
10 Shattuck St.
Boston, MA 02115

Mobile: +1.812.250.1248
E-mail: william_yu@hms.harvard.edu
WWW: yunwilliamyu.net

EDUCATION

Ph.D. Mathematics/Computer Science & AI Lab, 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

HMS BIRT T15 Fellow, Weber Group, DBMI, Harvard Medical School, since June 2018.
Research Fellow, Weber Group, DBMI, Harvard Medical School, June 2017-May 2018.
Senior Intern, D.E. Shaw Research, Summer 2016
Organizer and Recitation Leader, 18.095 Mathematics Lecture Series, MIT, IAP 2016
Software Engineering Intern, Ads Data Infrastructure, Google, Summer 2015
Recitation Leader, 18.03 Differential Equations, MIT, Spring 2015
Teaching Assistant, BE3-HMIB Modelling in Biology, Imperial College London, Fall 2010
Teaching Intern, BIOL-L311 Genetics, Indiana University Bloomington, Fall 2007 & Fall 2008

AWARDS & HONORS

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

FEATURED WORK

- **Yu YW** and Weber GM (2017). HyperMinHash: MinHash in LogLog space. arXiv:1710.08436. [cs.DS]. *Paper submitted to journal*.

  One of the standard problems in the database literature is measuring the similarity between two given sets. A standard metric for this is Jaccard index, which measures the ratio of the intersection cardinality to the union cardinality, and can be easily approximated using a min-wise hashing technique (MinHash). Naively, given a universe of size n from which the sets are drawn, this method requires $\theta(\log n)$ space to store summary sketches of sets that can be combined to compute Jaccard index, to ensure that accidental hash collisions are rare. **I realized that we only need to prevent accidental min-wise hash collisions, rather than all hash collisions; by using techniques from the LogLog family of cardinality estimators, HyperMinHash reduces the space requirement to $\theta(\log \log n)$.**

- **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**

  The similarity search problem is pervasive in computational biology. Though sophisticated heuristics and engineering are employed to make these tools (e.g. BLAST, Diamond) quite fast, they still don't keep pace with the exponential growth of biological data as experimental techniques become better. In this paper, we lay out a rigorous mathematical justification for

'compressive genomics', originally introduced by my advisor, which computes on compressed data to achieve orders of magnitude speedup. Using this insight, we were able to accelerate by orders of magnitude also small-molecule, metagenomic, and protein similarity search. **My major contribution was in the synthesis of mathematical ideas from fractal geometry and a new formulation of biological data space, the basis for the runtime proofs and algorithms we present.**

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

  Read-mapping—determining the most likely origin locations for substrings (reads) in the genome—is a basic problem in genomics and a manifestation of the similarity search problem. In this paper, we use clever clustering of reads and a homology table of the human genome to greatly accelerate standard read-mappers (such as Bowtie 2 or BWA) for both best- and all-mapping, improving all-mapping speeds by several orders of magnitude. **I played a key role in getting this paper published in Nature Biotech by framing the problem as an example of entropy-scaling search, going beyond the engineering details to the underlying nature of the problem, providing important context for a broader audience and enabling us to show that our CORA software was superior to other mapping tools.**

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

  As next-generation sequencing technologies rapidly improve, we are being inundated with vast amounts of raw sequencing data. One large part of these data are the quality scores that come with nucleotide calls in a read. These empirical confidence scores are much less compressible than the nucleotides themselves. In this paper, we show that by putting a human prior onto the quality scores using the k-mer landscape of each read, we can rapidly identify quality scores which are superfluous because the corresponding nucleotide calls are almost certainly correct. Not only does filtering out these quality scores drastically reduce storage and transmission requirements, but in our experiments, our software package Quartz even improve accuracy. **I was the primary developer of Quartz and coded a fast, parallelized implementation of it, as well as provided the mathematical justification for its efficacy.**

- 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**

  Here, we surveyed the modern algorithmic genomic landscape, providing an introduction to state-of-the-art methods that take advantage of our knowledge of biological data structure. Along those lines, we describe how our entropy-scaling search framework is revolutionizing the field of compressive genomics by providing **sublinear runtime algorithms that scale with the explosion of data**.

PEER-REVIEWED PUBLICATIONS

Luo Y*, **Yu YW***, Zeng J, Berger B, Peng J (2018). Metagenomic binning through low density hashing. *Bioinformatics*, 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*, doi:10.1089/cmb.2018.0032.

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**

Shajii A, Yorukoglu D, **Yu YW**, Berger B (2016). Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics*, 32(17): i538–i544.

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

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]

POSTER PRESENTATIONS

Luo Y, Yu YW, Zeng J, Berger B, Peng J. (2017). Metagenomic binning through low density hashing. *Biology of Genomes, Cold Spring Harbor, NY, May 9-13, 2017*

Yu YW, Yorukoglu D, Berger B. (2014). Compressive acceleration of approximate search in genomics. *Biological Data Science, Cold Spring Harbor, NY, November 5-8, 2014*

Yu YW, Barahona M, Yaliraki SN. (2010). Protein dynamics: a multiscale approach using community detection and graph comparison. *50th Anniversary Symposium of the British Biophysical Society, Cambridge, UK, July 16-18,2010*

Talks presented

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

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

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

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).

Professional Activities

Reviewer, *Bioinformatics*, 2017, 2018

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