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: Wu, Cathy Huey-Hwa, PhD, ACM Fellow, IEEE Fellow

eRA COMMONS USERNAME (credential, e.g., agency login): CATHYWU

POSITION TITLE: Unidel Edward G. Jefferson Chair in Engineering and Computer Science
Director, Center for Bioinformatics & Computational Biology; Director, Data Science Institute

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
National Taiwan University, Taiwan	B.S.	06/1978	Plant Pathology
Purdue University, West Lafayette, Indiana	M.S./Ph.D.	12/1984	Plant Pathology
Michigan State University, East Lansing, Michigan	Postdoc.	08/1986	Molecular Biology
University of Texas at Tyler, Texas	M.S.	08/1989	Computer Science

A. Personal Statement

I have conducted bioinformatics data science research for 30 years in areas encompassing artificial neural networks, genomic and protein informatics, text mining, natural language processing, ontology, semantic data integration, and knowledge network analysis. I have led/co-led the development of major bioinformatics resources including the international UniProt Consortium and the Protein Ontology Consortium, as well as community engagement effort such as the BioCreative (Critical Assessment of Text Mining in Biology). Recognized as a "Highly Cited Researcher" (top 1%) for 7 consecutive years (2014-2020), I have published more than 290 peer-reviewed papers, with 50,000 citations and an h-index of 71.

As Founding Director of the Center for Bioinformatics and Computational Biology (CBCB) since 2010 and the Data Science Institute (DSI) since 2018, I have provided a nucleating effort to catalyze multidisciplinary research collaborations and address big data problems across fields impacting our society. I launched several Bioinformatics degree programs, including the MS/PSM programs in Bioinformatics and Computational Biology, PhD program in Bioinformatics Data Science, as well as Graduate and Online Graduate Certificate programs. The interdisciplinary graduate programs now have more than 70 students from five colleges. I also established the Bioinformatics Core at CBCB and the Data Intensive & Computational Science Core at DSI to provide cutting-edge research cyberinfrastructure and data analytics capabilities to support research and educational programs.

I have mentored more than 200 undergraduate and graduate students, trainees, junior scientists and young investigators throughout my academic career, and received grants with significant mentoring effort, such as NSF IGERT and NRT-HDR training grants and several NIH Diversity/Re-Entry Supplement grants. As a mentor, I am fully committed to creating an inclusive, supportive and safe scientific research environment with utmost scientific integrity and will strive to promote diversity and equity. My mentees have received many awards and fellowships and have successful careers in academic, industry and government organizations. I serve as the faculty advisor of the UD Bioinformatics Student Association and supported the launching of the ACM-W (Women in Computing) chapter at UD. I am a mentor in the National Research Mentoring Network (NRMN) and has participated in the Culturally Aware Mentoring (CAM) study. A strong advocate for scientific rigor, transparency and reproducibility, I served on the Board on Research Data and Information of the National Research Council (NRC) and was an early adopter of the FAIR (Findable, Accessible, Interoperable and Reusable) data principles.

<u>Ongoing Research Projects</u>

1R35GM141873-01, NIH/NIGMS
Wu (PI)
08/25/21 – 07/31/26
Protein Knowledge Networks and Semantic Computing for Disease Discovery

2U24HG007822-08, NIH/NHGRI

Bateman, Bridge, Wu (MPI)
09/17/21 – 05/31/26
UniProt: A Protein Sequence and Function Resource for Biomedical Science

2125703, NSF/DGE
Jayaraman (PI), Role: Co-PI
09/01/21 – 08/31/26
NRT- HDR: Computing and Data Science Training for Materials Innovation, Discovery, Analytics

1919839, NSF/OAC
Eigenmann (PI), Role: Co-PI
09/01/19 – 08/31/22
MRI: Acquisition of a Big Data and High-Performance Computing System to Catalyze Delaware Research and Education

20-2020EPSCoR-0024, NASA
Matthaeus (PI), Role: Co-Science PI
10/01/20 – 09/30/23
NASA EPSCoR Research Project: Building a Competitive and Sustainable Delaware Remote Sensing Big Data Center for Cutting-Edge Coastal and Climate Change Research and Workforce Development

U54GM104941-07, NIH/NIGMS
Binder-Macleod (PI), Role: Biostatistics, Epidemiology & Research Design (BERD) Core
09/05/18 – 06/30/23
Delaware Clinical and Translational Research ACCEL Program

1736123, NSF/OIA
Harcum (PI), Role: Co-I and Mentor
08/01/17 – 07/31/22
RII Track-2 FEC: Advanced Biomanufacturing: Catalyzing Improved Host Development and High-Quality Medicines through Genome to Phenome Predictions

Recently Completed Research Projects

2P20GM103446-15, NIH/NIGMS
Stanhope (PI), Role: Program Coordinator
05/01/19 – 08/31/21
Delaware INBRE

2R01GM080646-10, NIH/NIGMS
Wu (PI)
09/01/15 – 08/31/20
PRO: A Protein Ontology in OBO Foundry for Scalable Integration of Biomedical Knowledge

1U01GM120953-01, NIH/NIGMS
Wu, Shanker (MPI)
08/05/16 – 07/31/20
Semantic Literature Annotation and Integrative Panomics Analysis for PTM-Disease Knowledge Network Discovery

1144726, NSF/DGE
Lee (PI), Role: Co-PI
07/01/12 – 06/30/19
IGERT: Systems Biology of Cells in Engineered Environments (SBE2)

Citations

1. Zhang X, Maity TK, Ross KE, Qi Y, Cultraro CM, Bahta M, Pitts S, Keswani M, Gao S, Nguyen KDP, Cowart J, Kirkali F, **Wu CH**, Guha U (2021) Alterations in the global proteome and phosphoproteome in third-generation EGFR TKI resistance reveal drug targets to circumvent resistance. *Cancer Research* 81(11):3051-3066. doi: 10.1158/0008-5472.CAN-20-2435. [PMC8182571]

2. Ma M, Zhao L, Ren J, Tulyakov S, **Wu CH**, Peng X. (2021). SMIL: Multimodal learning with severely missing modality. *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (3), 2302-2310.
3. Chen C, Ross KE, Gavali1 S, Cowart JE, **Wu CH**. (2021) COVID-19 knowledge graph from semantic integration of biomedical literature and databases. *Bioinformatics* 37(23), 4597-4598. doi: 10.1093/bioinformatics/btab694
4. McGarvey PB, Nightingale A, Luo J, Huang H, Martin MJ, **Wu CH**, UniProt Consortium. (2019) UniProt genomic mapping for deciphering functional effects of missense variants. *Human Mutation* 2019: 1-12. doi: 10.1002/humu.23738 [PMC6563471]

B. Positions, Scientific Appointments, and Honors

Positions

2018-present	Founding Director, Data Science Institute, University of Delaware (UD)
2010-present	Founding Director, MS in Bioinformatics & Computational Biology; PSM in Bioinformatics; Graduate Certificate in Bioinformatics (since 2010); Founding Director, PhD program in Bioinformatics Data Science (since 2012); Founding Director, Online Graduate Certificate in Applied Bioinformatics (since 2017); in Biomedical Informatics & Data Science (since 2019), UD
2009-present	Unidel Edward G. Jefferson in Engineering and Computer Science; Founding Director, Center for Bioinformatics & Computational Biology; Professor, Department of Computer & Information Sciences and Department of Biological Sciences, UD
2008-2010	Founding Co-Director, Bioinformatics Track, MS in Biochemistry and Molecular Biology, GUMC
2001-present	Professor (2001-2008), Adjunct Professor (since 2009), Department of Biochemistry and Molecular & Cellular Biology; Member, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center (GUMC)
1999-present	Director of Bioinformatics (1999-2001), Director (since 2001), Protein Information Resource (PIR), Georgetown University (since 1999) and University of Delaware (since 2009)
1990-1999	Assistant Professor (90-94), Associate Professor (94-98), Professor (98-99), Department of Biomathematics, University of Texas Health Center at Tyler
1989-1994	Assistant Professor, Department of Computer Science, University of Texas at Tyler
1986-1987	Research Scientist, Department of Plant Pathology & Microbiology, Texas A&M University
1985-1986	Postdoctoral Fellow, MSU-DOE Plant Research Laboratory, Michigan State University (Advisor: Christopher R. Somerville, Member, National Academy of Sciences)

Scientific Appointments

2021-Present	Board of Regents Comparative Genomics Resource Working Group, National Library of Medicine (NLM), NIH
2021-2022	Council of Councils (CoC) Working Group for the Common Fund Data Ecosystem, National Institutes of Health (NIH)
2020	Board of Scientific Counselors (Ad hoc member), National Library of Medicine (NLM), NIH
2019-Present	External Advisory Committee, Center of Quantitative Biology COBRE (Centers of Biomedical Research Excellence), Geisel School of Medicine at Dartmouth, NH
2017-present	Senior Member (2017-2021), Fellow (2022-present), The Institute of Electrical and Electronics Engineers (IEEE)
2017-2021	Advisory Council, National Institute of General Medical Sciences (NIGMS), NIH
2016-present	Editorial Board, *Current Opinion in Systems Biology*
2015-present	Advisory Board, NIAID Bioinformatics Integration Support Contract (BISC), NIH
2014-present	Associate Editor (2014-present), Steering Committee (2018-2021), IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)
2014	External Scientific Panel, Library of Integrated Network-Based Cellular Signatures (LINCS), NIH
2013-2015	Informatics Advisory Committee, Joint Genome Institute (JGI), Department of Energy (DOE)
2012	External Advisory Panel, NHLBI Proteomics Program, NIH
2010-present	Board of Directors, SIGBio (2010-present), Distinguished Member (2017-2021), Fellow (2021-present), Association for Computing Machinery (ACM)
2008-2018	Board of Directors; Chair of Bioinformatics and Biostatistics Subcommittee of US HUPO Initiative; Executive Committee (2010-2013), US Human Proteome Organization (US HUPO)
2008-2013	Executive Editor, Journal of Proteomics and Bioinformatics

2008-2010 Board on Research Data and Information (BRDI), National Research Council (NRC)
2006-2010 TeraGrid Scientific Advisory Board (2006-2010), Grand Challenge Communities (GCC) Task Force, Office of Cyberinfrastructure (OCI) (2009-2010), National Science Foundation (NSF)
2005-2015 Advisory Board, Protein Data Bank (PDB)
2005-2014 Council (2012-2014; 2005-2008), Human Proteome Organization (HUPO)
2002-2013 Protein Structure Initiative Advisory Committee, NIGMS, NIH
2000-present Board of Directors (2000-2004), Education Committee (since 2003), Senior Member (2016-present), International Society for Computational Biology (ISCB)

Honors

2022 Fellow, Institute of Electrical and Electronics Engineers (IEEE)
2021-2026 MIRA (R35) Award, National Institute of General Medical Sciences, NIH
2021 Fellow, Association for Computing Machinery (ACM)
2021 Fellow, Asia-Pacific Artificial Intelligence Association (AAIA)
2020 Distinguished Agriculture Alumni Award, Purdue University
2019 Recognition of Service Award, Association for Computing Machinery (ACM)
2014-2020 Recognized as a "Highly Cited Researcher" (top 1%) by Thomson Reuters/Clarivate Analytics
1993-1999 FIRST (R29) Award, National Library Medicine (NLM), NIH
1988 President's Academic Scholarship, University of Texas at Tyler
1983 Du Pont Graduate Student Award, Purdue University
1975-1978 Book Coupon Award (top 5% of class), National Taiwan University (1975, 1977, 1978)

C. Contribution to Science

[>290 peer-reviewed publications, Google Scholar: 50,000 citations, h-index: 71, i10-index: 207]

1. *Artificial Neural Networks and Deep Learning (1990-present):* I have developed artificial neural networks for molecular sequence analysis, resulting in about 40 refereed papers, an NIH FIRST (R29) Award (1993-1999), a US patent (#5845049, 1998) and licenses, and a book "*Neural Networks and Genome Informatics*" (ISBN 0080428002, 2000). My recent research further employs deep learning algorithms.
    a. **Wu CH**, Whitson G, McLarty J, Ermongkonchai A, Chang TC. (1992) Protein classification artificial neural system. *Protein Science* 1(5): 667-677. [PMC2142223] (158 citations; 1500 citations from related artificial neural network papers)
    b. Book: **Wu CH**, McLarty J. (2000). *Neural Networks and Genome Informatics*. Elsevier. (161 citations)
    c. Huang L, Liao L, **Wu CH**. (2018) Completing sparse and disconnected protein-protein networks by deep learning. *BMC Bioinformatics* 19(1):103. doi: 10.1186/s12859-018-2112-7. [PMC5863833]
    d. Su P, Li G, **Wu CH**, Vijay-Shanker K. (2019) Using distant supervision to augment manually annotated data for relation extraction. *PLoS One* 14: e0216913. doi: 10.1371/journal.pone.0216913 [PMC6667146]

2. *Protein Information Resource (PIR) (1999-present) and UniProt Consortium (2002-present):* I have led the development of PIR to become a major bioinformatics resource, as profiled in *The Scientist* (10/15/2001) "*Cathy Wu at the Crossroads: She saved the Protein Information Resource database and now aims to restore it to the world's best*," and co-founded the UniProt Consortium with Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Resource (EBI). Many PIR resources have been integrated with the UniProt, including the PIRSF protein families into InterPro. Funded by the NLM P41 and NHGRI U01/U24 grants, the PIR/UniProt resources now receive >8 million pageviews monthly from nearly 1 million users worldwide.
    a. **Wu CH**, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, Barker WC. (2003) The Protein Information Resource. *Nucleic Acids Res.* 31: 345-347. [PMC165487] (562 citations; 2000 citations from related PIR papers)
    b. UniProt Consortium (**Wu CH**). (2015) UniProt: a hub for protein information. *Nucleic Acids Res.* 43: D204-D212 [PMC4384041] (3834 citations; 20,000 citations from UniProt papers)
    c. Hunter S, Apweiler R, Attwood TK, Bairoch A, et al., **Wu CH**, Yeats C. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37: D211-215 [PMC2686546] (1899 citations; 5000 citations from related PIRSF and InterPro papers)
    d. Suzek BE, Wang Y, Huang H, McGarvey P, **Wu CH**, UniProt Consortium. (2015) UniRef Clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31(6): 926-932. [PMC4375400] (728 citations, 2000 citations from UniRef papers).

3. *Natural Language Processing and BioCreative Text Mining Challenge Evaluations (2002-present):* I have established several collaborations in text mining and natural language processing research for full-scale information extraction from PubMed/PMC papers. To improve the utility, usability and interoperability of text mining tools, I have co-led the BioCreative Workshops to introduce the Interactive Text Mining Track, and co-edited 3 BioCretive Conference Proceedings and 3 journal special/virtual issues featuring best-performing text mining systems along with Workshop and Track overviews.
    a. Hirschman L, Park JC, Tsujii J, Wong L, **Wu CH.** (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics* 18(12): 1553-1561. doi: 10.1093/bioinformatics/18.12.1553. [PMID: 12490438] (395 citations)
    b. Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, **Wu CH.** (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics* 21(11): 2759-2765. doi: 10.1093/bioinformatics/bti390. [PMID: 15814565] (121 citations; 1000 citations from related text mining/NLP papers)
    c. Arighi CN, Wang Q, **Wu CH**. (Editors) (2017). Proceedings of the BioCreative VI Challenge Evaluation Workshop, October 18-20, 2017 (1000 citations from related BioCreative Consortium papers)
    d. Ren J, Li G, Ross KE, Arighi CA, McGarvey PB, Rao S, Cowart JC, Madhavan S, Vijay-Shanker K, **Wu CH**. (2018) iTextMine: integrated text-mining system for large-scale knowledge extraction from literature. *Database (Oxford)* 2018. doi: 10.1093/database/bay128 [PMC6301332]

4. *Biomedical Ontology, Semantic Computing and Knowledge Networks (2007-present):* I have led the Protein Ontology Consortium to develop PRO within the Open Biomedical Ontologies (OBO) Foundry for semantic knowledge integration, and have developed integrative approach combining data mining, text mining and ontologies for knowledge network construction from scientific literature and omics data.
    a. Natale DA, Arighi CN, Barker WC, Blake J, Chang TC, Hu Z, Liu H, Smith B, **Wu CH.** (2007) Framework for a protein ontology. *BMC bioinformatics* 8 (Suppl 9): S1. [PMC2217659] [147 citations; 2500 citations from ontology papers]
    b. Huang LC, Ross KE, Baffi TR, Drabkin H, Kochut KJ, Ruan Z, D'Eustachio P, McSkimming D, Arighi CN, Chen C, Natale DA, Smith C, Gaudet P, Newton AC, **Wu CH**, Kannan N. (2018) Integrative annotation and knowledge discovery of kinase post-translational modifications and cancer-associated mutations through federated protein ontologies and resources. *Scientific Reports* 8(1): 6518 [PMC5916945]
    c. Gavali S, Cowart J, Chen C, Ross KE, Arighi CN, **Wu CH**. (2020) RESTful API for iPTMnet: A resource for protein post-translational modification network discovery. Database (Oxford) 2020. doi: 10.1093/database/baz157 (>100 citations from iPTMnet papers)
    d. Chen C, Huang H, Ross KE, Cowart JE, Arighi CN, **Wu CH**, Natale NA. (2020) Protein ontology on the semantic web for knowledge discovery. *Scientific Data* 7(1): 337. doi: 10.1038/s41597-020-00679-9.

5. *Clinical Genomics/Proteomics and Data Analytics (2009-present):* I have founded the Center for Bioinformatics and Computational Biology and developed research infrastructure including next-generation sequencing and big data analytics capabilities for clinical and translational research in precision medicine.
    a. Chen C, Khaleel SS, Huang H, **Wu CH**. (2014) Software for pre-processing Illumina next-generation sequencing short read sequences. *Source Code Biol Med.* 9: 8. doi: 10.1186/1751-0473-9-8. [PMC4064128] (176 citations)
    b. Selvanathan SP, Graham GT, Erkizan HV, Dirksen U, Natarajan TG, Dakic A, Yu S, Liu X, Paulsen MT, Ljungman ME, **Wu CH**, Lawlor ER, Üren A, Toretsky JA. (2015) Oncogenic fusion protein EWS-FLI1 is a network hub that regulates alternative splicing. *Proc Natl Acad Sci USA* 112(11): E1307-1316. [PMC4371969] (114 citations)
    c. Lu C, Sidoli S, Kulej K, Ross K, **Wu CH**, Garcia BA. (2019) Coordination between TGF-β cellular signaling and epigenetic regulation during epithelial to mesenchymal transition. *Epigenetics & Chromatin* 12:11. doi: 10.1186/s13072-019-0256-y [PMC6368739]
    d. Ovadia EM, Pradhan L, Sawicki LA, Cowart J, R. E. Huber, Polson SW, Chen C, van Golen KL, Ross KE, **Wu CH**, Kloxin AM. (2020) Understanding ER+ breast cancer dormancy using bioinspired synthetic matrices for long-term 3D culture and insights into late recurrence. Adv Biosyst. 4(9): e2000119. doi: 10.1002/adbi.202000119. [PMC7807552]

List of Published Work in My Bibliography:
http://www.ncbi.nlm.nih.gov/sites/myncbi/cathy.wu.1/bibliography/40319526/public/?sort=date&direction=descending