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

eRA COMMONS USER NAME (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

With background in both biology and computer science, I have conducted bioinformatics and data science research for 25 years in areas encompassing genomic and protein annotation, biomedical text mining, biomedical ontology, gene-disease-drug network modeling, and big data analytics. I have received more than 60 grants as the PI, Consortium PI or Co-PI from NIH, NSF, DOE, and other agencies. I have led the development of major bioinformatics resources, including the Protein Information Resource and the international UniProt Consortium with 6 million pageviews per month from over 750,000 unique sites worldwide. Recognized as a "Highly Cited Researcher" (top 1%), I have published more than 260 peer-reviewed papers, with over 33,000 citations and an h-index of 60. I established the Center for Bioinformatics and Computational Biology in 2009 at UD to foster collaborative research, presently with over 70 affiliate faculty from across the university and the region. The Center has developed cutting-edge bioinformatics and data analytics infrastructure, including genomic analytics capabilities for precision medicine, and is the home of the Bioinformatics Master's, PhD, and graduate certificate programs, interdisciplinary programs now with more than 60 students. I was recently appointed the Founding Director of the Data Science Institute, serving as a nucleating effort to catalyze and coordinate data science activities at UD, connecting researchers across seven colleges to foster multidisciplinary research collaborations in both foundations and applications of data sciences.

1. **Wu CH**, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research* 34: D187-191. [PMC1347523]
2. Natale DA, Arighi CN, Barker WC, Blake JA, Bult CJ, Caudy M, Drabkin HJ, D'Eustachio P, Evsikov AV, Huang H, Nchoutmboube J, Roberts NV, Smith B, **Wu CH.** (2011) The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Research* 39: D539-545. [PMC3013777]
3. Huang H, Arighi CN, Ross KE, Ren J, Li G, Chen SC, Wang Q, Cowart J, Vijay-Shanker K, **Wu CH**. (2018) iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic Acids Research* 46: D542-550. [PMC5753337]
4. 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]

B. Positions and Honors

Positions and Employment

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

Honors, Professional Appointments and Activities

1975-1978 Book Coupon Award (top 5% of class), National Taiwan University (1975, 1977, 1978)
1983 Du Pont Graduate Student Award, Purdue University
1988 President's Academic Scholarship, University of Texas at Tyler
1993-1999 NIH FIRST (R29) Award, National Library Medicine (NLM), National Institutes of Health (NIH)
2000-present Board of Directors (2000-2004), Education Committee (since 2003), Senior Member (since 2016), International Society for Computational Biology (ISCB)
2002-2013 Protein Structure Initiative Advisory Committee, National Institute of General Medical Sciences (NIGMS), National Institutes of Health (NIH)
2004-present Editor, SEBI (Science, Engineering, and Biology Informatics) Book Series (ISSN: 1793-3692)
2005-2014 Council (2012-2014; 2005-2008), Human Proteome Organization (HUPO)
2005-2015 Advisory Board, Protein Data Bank (PDB)
2006-2010 TeraGrid Scientific Advisory Board, National Science Foundation (NSF)
2008-present 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)
2009-2010 Grand Challenge Communities (GCC) Task Force, Office of Cyberinfrastructure (OCI), National Science Foundation (NSF)
2010-present Board of Directors, SIGBio, Association for Computing Machinery (ACM)
2012 External Advisory Panel, NHLBI Proteomics Program, National Institutes of Health (NIH)
2013-2015 Informatics Advisory Committee, Joint Genome Institute (JGI), Department of Energy (DOE)
2014 Ad hoc Member, Advisory Council, NIGMS, NIH
2014 External Scientific Panel, Library of Integrated Network-Based Cellular Signatures (LINCS), NIH
2014-present Recognized as a "Highly Cited Researcher" (top 1%) by Thomson Reuters/Clarivate Analytics
2014-present Associate Editor, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*
2015-present Advisory Board, NIAID Bioinformatics Integration Support Contract (BISC), NIH
2016-present Editorial Board, *Current Opinion in Systems Biology*
2017-Present Advisory Council, NIGMS (National Institute of General Medical Sciences), NIH

Conference Organizing/Scientific Committees (>60): ACM BCB-2018 (Co-Chair); BioCreative VI-2017; BioCreative V-2015; IPG PPPS-2014 (Co-Chair); ACM BCB-2013 (Co-Chair); BioCreative IV-2013; BioCuration-2012; HUPO-2012; BioCreative-2012 (Chair), BioCreative III-2010 (Chair); BIBM-2012 (Co-

Chair), BIBM-2009 (Co-Chair); USHUPO-2008 (Co-Chair); BIBM-2008; ISMB-2008, 2007, 2004, 2002; HUPO-2006; PSB-2006, 2003, 2002 (since 2002)

Grant Review Panels/Site Visit Teams/Ad hoc Reviews (>60): NIGMS council; NIH (CSR, NIGMS, NLM, NCI, NIDA, NCRR), including BD2K and PRISMS panels; NSF (BIO/BDI, BIO/PGRP, CISE/IIS); DOE (BER)

Invited Presentations (>170): Invited talks and panelist at international conferences, universities, companies

Bioinformatics Resources: The Protein Information Resource (proteininformationresource.org/) [Director] and UniProt (www.uniprot.org) [Consortium PI] to support genomic, proteomic and systems biology research

Students/Junior Researchers Mentored (>100): Advised/mentored over 100 graduate students, postdoctoral researchers and junior investigators to date

Books (4), Journal Special/Virtual Issues and Conference Proceedings (11)

C. Contribution to Science
[>260 peer-reviewed publications, Google Scholar: >33,000 citations, h-index: 60, i10-index: 165]

1. *Artificial Neural Networks, Machine Learning and Deep Learning (1990-present):* My early computational biology research and publications involved the development of artificial neural networks for molecular sequence analysis, which led to ~40 refereed papers, an NIH FIRST (R29) Award (1993-1999), a US patent and license agreement, and a book "*Neural Networks and Genome Informatics*" (ISBN 0080428002, 2000). Recent years, my team has employed machine learning and deep learning for applications ranging from natural language processing to protein sequence and network analysis
    a. **Wu CH**, Whitson G, McLarty J, Ermongkonchai A, Chang TC. (1992) Protein classification artificial neural system. *Protein Science* 1(5): 667-677. [PMC2142223] (144 citations; >1000 citations from related artificial neural network papers)
    b. Book: **Wu CH**, McLarty J. (2000). *Neural Networks and Genome Informatics*. Elsevier. (129 citations)
    c. Du T, Liao L, **Wu CH**, Sun B. (2016) Prediction of residue-residue contact matrix for protein-protein interaction with Fisher score features and deep learning. *Methods* 110:97-105. [PMID: 27282356]
    d. 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]

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, an activity 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 Amos Bairoch (SIB-Swiss Institute of Bioinformatics) and Rolf Apweiler (EBI-European Bioinformatics Resource). Many research resources developed at the PIR have been integrated into the Consortium resources, including the integration of the PIRSF protein family classification into the InterPro signature database. Funded by the early NLM P41 and recent NHGRI U01 and U24 grants, the PIR/UniProt resources now receive >6 million pageviews.
    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. *Nucl. Acids Res.* 31: 345-347. [PMC165487] (496 citations; >1800 citations from related PIR papers)
    b. Bairoch A, Apweiler R, **Wu CH**, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. (2005) The Universal Protein Resource (UniProt). Nucleic acids research. 33: D154-159. [PMC540024] (1808 citations; >15000 citations from related UniProt papers)
    c. Hunter S, Apweiler R, Attwood TK, Bairoch A, et al., **Wu CH**, Yeats C. (2009) InterPro: the integrative protein signature database. *Nucl. Acids Res.* 37: D211-215 [PMC2686546] (1446 citations; >4500 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] (741 citations for the original 2007 UniRef paper).

3. *Text Mining/Natural Language Processing (2002-present) and BioCreative Consortium (2009-present):* I have established several collaborations in text mining and natural language processing research for full-scale information extraction from PubMed abstracts and PMC Open Access full-text articles, with over 40 papers. To improve the utility, usability and interoperability of text mining tools, I have co-led the BioCreative (Critical Assessment of Text Mining in Biology) Workshops to introduce the Interactive Text Mining Track and

the BioC exchange format, 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. [PMID: 12490438] (368 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. [PMID: 15814565] (114 citations; >400 citations from related text mining/NLP papers)
c. Peng Y, Torii M, **Wu CH**, Vijay-Shanker K. (2014) A generalizable NLP framework for fast development of pattern-based biomedical relation extraction systems. *BMC Bioinformatics* 15**,** 285. [PMC4262219] (Highly Accessed)
d. Arighi CN, Wang Q, **Wu CH**. (Editors) (2017). Proceedings of the BioCreative VI Challenge Evaluation Workshop, October 18-20, 2017 (>450 citations from related BioCreative Consortium papers)

4. *Protein Ontology (PRO) Consortium, Systems Biology and Knowledge Networks (2007-present):* I have led the NIGMS-funded Protein Ontology Consortium to develop PRO within the Open Biomedical Ontologies (OBO) Foundry for semantic knowledge integration, and have developed integrative bioinformatics approach combining data mining, text mining and ontologies for system biology, including protein and PTM 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] [Cited 110 times – Google Scholar; Total Cited >300 times from PRO Consortium papers]
    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]
    c. Celen I, Ross KE, Arighi CN, **Wu CH**. (2015) Bioinformatics knowledge map for analysis of beta-catenin function in cancer. *PLoS One* 10(10): e0141773. [PMC4624812]
    d. 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]

5. *Clinical Genomics and Data Analytics (2009-present):* I have founded the Center for Bioinformatics and Computational Biology and developed research infrastructure for the Delaware-INBRE (IDeA Network of Biomedical Research Excellence), including next-generation sequencing (NGS) 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. [PMC4064128] (Highly Accessed)
    b. Crowgey E, Stabley D, Chen C, Huang H, Polson SW, Sol-Church K, **Wu CH**. (2015) An integrated approach for analyzing clinical genomic variant data from next generation sequencing. *J Biomol Tech.* 26(1): 19-28. [PMC4310222]
    c. Crowgey EL, Kolb A, **Wu CH**. (2015) Development of bioinformatics pipeline for analyzing clinical pediatric NGS data. *AMIA Jt Summits Transl Sci Proc.* 2015: 207-211. [PMC4525226]
    d. Mahmood AS, Rao S, McGarvey PB, **Wu CH**, Madhavan S, Vijay-Shanker K (2017) eGARD: Extracting associations between genomic anomalies and drug responses from text. *PLoS One* 12(12):e0189663. [PMC5738129]

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

D. Research Support

Ongoing Research Support

NIH/NHGRI 2U24HG007822-05 PI: Bateman; Site PI: Wu 06/01/18 – 05/31/21
UniProt: A Centralized Protein Sequence and Function Resource
The goal of the UniProt Consortium is to provide a centralized curated, accurate, stable, and comprehensive protein sequence and function resource by enhancing the UniProt Knowledgebase and

ensuring that the rich and diverse information in UniProt will be of use to a broad scientific user community. Role: Consortium PI (PI of the PIR Component)

NIH/NIGMS 2R01GM080646-10 PI: Wu 09/01/15 – 08/31/19
PRO: A Protein Ontology in OBO Foundry for Scalable Integration of Biomedical Knowledge
The goal of this project is to deepen and broaden the Protein Ontology for scalable semantic integration of biomedical data to facilitate protein-disease knowledge discovery and clinical applications.

NIH/NIGMS 1U01GM120953-01 MPI: Wu, Shanker 08/05/16 – 07/31/19
Semantic Literature Annotation and Integrative Panomics Analysis for PTM-Disease Knowledge Network Discovery
The goal is to develop a collaborative knowledge environment for semantic annotation of scientific literature and integrative omics analysis for PTM-disease discovery to support Big Data to Knowledge in biomedicine.

NIH/NIGMS 2P20GM103446-14 PI: Stanhope; PC: Wu 05/01/19 – 04/30/24* (*funding in progress)
Delaware INBRE
The goal is to further develop the research infrastructure for the IDeA Network of Biomedical Research Excellence in Delaware. Role: Program Coordinator

NSF/DGE 1144726 PI: Lee 07/01/12 – 06/30/19
IGERT: Systems Biology of Cells in Engineered Environments (SBE2)
The goal of this interdisciplinary training program is to prepare students to carry out research on Cells in Engineered Environments using the technologies and approaches of Systems Biology. Role: Co-PI

NSF/OIA 1736123 PI: Harcum 08/01/17 – 07/31/21
RII Track-2 FEC: Advanced Biomanufacturing: Catalyzing Improved Host Development and High Quality Medicines through Genome to Phenome Predictions
The goal is to bring together several EPSCoR collaborating institutions to develop new biological approaches to better understand the Chinese hamster ovary (CHO) cell line used to manufacture most biopharmaceuticals. Role: Senior Investigator and Mentor

Completed Research Support (during past 3 years)

NIH/NIGMS 2P20GM103446-14 PI: Stanhope; PC: Wu 08/01/14 – 04/30/19
Delaware INBRE
The goal is to further develop the research infrastructure for the IDeA Network of Biomedical Research Excellence in Delaware. Role: Program Coordinator and Bioinformatics Core Director

NIH/NHGRI 1U41HG007822-01 PI: Bateman; Site PI: Wu 08/01/14 – 05/31/18
UniProt: A Centralized Protein Sequence and Function Resource
The goal of the UniProt Consortium is to provide a centralized curated, accurate, stable, and comprehensive protein sequence and function resource by enhancing the UniProt Knowledgebase and ensuring that the rich and diverse information in UniProt will be of use to a broad scientific user community. Role: Consortium PI (PI of the PIR Component)

NIH/NHGRI 1U01HG008390-01 PI: Madhavan; Site PI: Wu 09/22/15 – 05/31/18
MACE2K-Molecular and Clinical Extraction: A Natural Language Processing Tool for Personalized Medicine
The goal is to build an innovative software stack (MACE2K) to adapt and extend widely tested BioCreative natural language processing (NLP) tools to automatically retrieve and pre-process targeted therapy information from clinicaltrials.gov and scientific literature. Role: UD Subcontract PI

NIH/Common Fund 5U54HL127624-03 PI: Ma'ayan; Site PI: Wu 05/01/16 – 04/30/18
Data Coordination and Integration Center for LINCS-BD2K: eDSR-Collaborative Resource for LINCS Panomics PTM Knowledge Network
The goal is to develop a post-translational modification (PTM)-centric research resource with the Data Coordination and Integration Center to support scientific discovery from panomics LINCS data.
Role: UD Subcontract PI

NSF/DBI 1062520 PI: Wu 07/01/11 – 06/30/16
ABI Development: Integrative Bioinformatics for Knowledge Discovery of PTM Networks
The goal of this project is to develop a bioinformatics research infrastructure for integrated understanding of plant post-translational modifications (PTMs) in systems biology context.