OMB No. 0925-0001 and 0925-0002 (Rev. 10/15 Approved Through 10/31/2018)

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: Xia Jiang

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

POSITION TITLE: Associate Professor, Department of Biomedical Informatics, University of Pittsburgh, School of Medicine

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 DateMM/YYYY | FIELD OF STUDY |
| --- | --- | --- | --- |
| Rose-Hulman Institute of Technology, Terre Haute, IN | MS | 1997 | Mechanical Engineering |
| Northeastern Illinois University, Chicago, IL | MS | 1999 | Computer Science |
| University of Pittsburgh, Pittsburgh, PA | PHD | 2008 | Biomedical Informatics |

**A. Personal Statement**

I am an Associate Professor in the Department of Biomedical Informatics and the Intelligent Systems Program at the University of Pittsburgh, and in the joint CMU-Pitt Computational and Systems Biology PhD program. I am interested in applying Bayesian Network and other machine learning methods to help solve biomedical problems that are in the nature of prediction, detection, search, learning interactions, and decision making. I am also interested in algorithm design and optimization. I am one of the pioneering researchers in applying Bayesian Networks to disease outbreak early detection and characterization, and made substantial contribution in this regard. I developed BAYESNET for Cryptosporidium outbreak detection and characterization by monitoring over-the-counter antidiarrhea drug sales; I developed REFINE, a spatial cluster detection algorithm for early influenza outbreak detection; I developed a Bayesian Network method for predicting an epidemic curve early in an outbreak, and G-AMOC curves for the evaluation of outbreak detection systems; and I developed PCT, PCTS, and BNetScan, a set of Bayesian Network based spatial and temporal outbreak detection systems. By monitoring 54 different chief complaints by patients admitted to the emergent department, these systems detect 9 different disease outbreaks including influenza, cryptosporidiosis, hepatitis A, and the six CDC Category A diseases, which are anthrax (two stages/values), plague (two stages/values), smallpox, tularemia, botulism, and hemorrhagic fever (two stages/values). I am also one of the early researchers in applying Bayesian Network scoring to learning epistatic interactions from GWAS data. In that regard, I defined a class of Bayesian Network models for learning epistasis called SNP patterns; I developed a novel Bayesian Network learning score criterion for detecting epistatic biomarkers, named as the Bayesian Network Minimum Bit Length (BNMBL) score; I identified a bound for the Bayesian score of a SNP pattern; I also developed several novel and efficient epistasis learning and evaluation algorithms including Multiple Beam Search (MBS), Repeated Extended Greedy Approach to Learning (REGAL), Bayesian Network Posterior Probability (BNPP), the Interaction Strength (IS), and Information Gain based algorithms for learning interactions (IGain). These algorithms have not only been applied to the detection of epistatic biomarkers, including rare variants associated with a disease, but also been applied to detect the interaction between microRNA and mRNA in breast cancer. I have collaborated with other researchers in developing intelligent systems and algorithms in causal learning, biomedical prediction, clinical decision support, and disease mechanism discovery at the molecular level. My research has resulted in 55 publications, 41 of which are original research papers related to the development and evaluation of new algorithms and informatics tools. I am the first author on 24 of these original research papers. I have served as PI on multiple externally funded projects, and have extensive experience and prior success in mining large-scale biomedical data.

**B. Positions and Honors**

**Employment**

8/1995-4/1997 Part-time Teaching Assistant, Rose-Hulman Institute of Technology

4/1997-4/2002 Software Engineer, Warehouse Equipment Inc., Elk Grove Village, IL

5/2002-8/2005 Instructor, Department of Computer Science, Northeastern Illinois University

8/2005-11/2008 Graduate Student Researcher, Department of Biomedical Informatics, University of Pittsburgh

12/2008-9//2010 Postdoctoral Scholar, Department of Biomedical Informatics, University of Pittsburgh

10/2010-12//2011 Postdoctoral Associate, Department of Biomedical Informatics, University of Pittsburgh

1/2012-10/2016 Assistant Professor, Department of Biomedical Informatics, University of Pittsburgh

11/2017-present Associate Professor, Department of Biomedical Informatics, University of Pittsburgh

**Honors**

2008 Semi-finalist of “the Best Publication of the Year” award for the *Journal of Biomedical Informatics* paper “Bayesian Prediction of an Epidemic Curve”, International Society for Disease Surveillance Annual Conference

2008 National Library of Medicine Postdoctoral Training Scholarship

2009 National Library of Medicine Postdoctoral Training Scholarship

2010 National Library of Medicine Postdoctoral Training Scholarship

 2010 Finalist Award, student paper competition at American Medical

 Informatics Association (AMIA) 2010 Annual Symposium for paper titled

 “A Fast Algorithm for Learning Epistatic Genomic Relationships”,

 Proceedings of American Medical Informatics Association (AMIA)

 2010 Annual Symposium

**Professional Societies**

2006-Present Member, Association for the Advancement of Artificial Intelligence (AAAI)

2007-2008, 2012 Member, Association for Computing Machinery (ACM)

2007-Present Member, American Medical Informatics Association (AMIA)

**C. Contribution to Science**

C1. Developing Bayesian Network methods for disease outbreak detection, prediction, and characterization:

Even modest improvement in disease outbreak detection and early characterization could have significant impact on public health in terms of lives saved and reduced economic cost. I have ample research experience in this regard. I am one of the pioneering researchers in applying Bayesian Networks for disease outbreak detection and characterization. I developed BAYESNET for Cryptosporidium outbreak detection and characterization by monitoring over-the-counter antidiarrheal drug sales [1], developed REFINE, a spatial cluster detection algorithm for early influenza outbreak detection [2], and developed PCT, PCTS, and BNetScan [5, 6, 7], which are spatial and temporal outbreak detection systems that detect 9 different disease outbreaks by monitoring the chief complaints data collected in the emergency department. I also developed a Bayesian Network method for predicting an epidemic curve early in an outbreak [3], and G-AMOC curves for the evaluation of outbreak detection systems [4].

1. **Jiang X**, Wallstrom GL. A Bayesian network for outbreak detection and prediction. 21st Association for the Advancement of Artificial Intelligence (AAAI) Conference. 2006: 1155-60.
2. **Jiang X**, Cooper GF. A recursive algorithm for spatial cluster detection. AMIA Annu Symp Proc 2007; 369-73. PMCID: PMC2655859.
3. **Jiang X**, Wallstrom G, Cooper GF, Wagner MM. Bayesian prediction of an epidemic curve. Journal of biomedical informatics. Journal of Biomedical Informatics. 2009 Feb; 42 (1):90-9. PMID: 18593605.
4. **Jiang X**, Cooper GF, Neill DB. Generalized AMOC curves for evaluation and improvement of event surveillance. AMIA Annual Symposium; 2009:281-5.
5. **Jiang X**, Cooper GF. A real-time temporal Bayesian architecture for event surveillance and its application to patient-specific multiple disease outbreak detection. Data Mining and Knowledge Discovery. 2010; 20 (3):328-60.
6. **Jiang X**, Cooper GF. A Bayesian spatio-temporal method for disease outbreak detection. Journal of the American Medical Informatics Association: JAMIA. 2010; 17(4):462-71. PMCID: PMC2995651. PMID: 20595315.
7. **Jiang X**, Neill DB, Cooper GF. A Bayesian network model for spatial event surveillance. International Journal of Approximate Reasoning. 2010; 51(2):224-39.

C2*.* Developing Bayesian Network learning methods for epistasis detection, interactive biomarker and health risk factors discovery: The interaction of entities/attributes is a universal phenomenon and plays a critical role in many aspects of human life. For example, it is believed that interactions among genetic loci (epistasis) may contribute significantly to susceptibility to common diseases. However, learning interactions from data is an area of machine learning that has not yet been fully explored. I was the first to apply Bayesian network scoring to learning epistatic interactions from GWAS data. Specifically, I defined a class of Bayesian Network models for learning epistasis called SNP patterns, and I developed the Bayesian Network Minimum Bit Length (BNMBL) score, a scoring criterion for detecting epistatic interactions that is based on the minimum description length principle [1]. I also identified a bound for the score of a SNP pattern. The bound provides an upper limit on the Bayesian score of any pattern that could be obtained by expanding a given pattern. Finally, I developed evaluation measures for learning interactions, and a tool for ameliorating multiple hypotheses testing issues when learning from high-dimensional data [2, 3]. The interaction learning methods I developed have been applied to learning interactive causes. [4, 5].

1. **Jiang X**, Barmada MM, Visweswaran S. Identifying genetic interactions in genome-wide data using Bayesian networks. Genetic Epidemiology 2010; 34(6): 575-81. PMID: 20568290. PMCID: PMC3931553.
2. **Jiang X**, Barmada MM, Cooper GF, Becich MJ. A Bayesian method for evaluating and discovering disease loci associations. PLoS ONE 2011. 6(8):e22075.PMID: 21853025. PMCID: PMC3154195.
3. **Jiang X**, Barmada MM, Becich MJ. Evaluating de novo locus-disease discoveries in GWAS using the signal-to-noise ratio. AMIA Annual Symposium 2011; 617-24. PMID: 22195117. PMCID: PMC3243170.
4. Zeng Z. **Jiang X**, Neapolitan RE. Discovering Causal Interactions Using Bayesian Network Scoring and Information Gain. BMC Bioinformatics 2016; 17:221. PMID: 27230078. PMCID: PMC 4880828.
5. Hill SM, Heiser LM, Cokelaer T,,, **Jiang X**,,,. Empirical [Inferring causal molecular networks: empirical assessment through a community-based effort](https://na01.safelinks.protection.outlook.com/?url=https%3a%2f%2fscholar.google.com%2fcitations%3fview_op%3dview_citation%26hl%3den%26user%3dE9_aqCoAAAAJ%26cstart%3d20%26citation_for_view%3dE9_aqCoAAAAJ%3a2P1L_qKh6hAC&data=01%7c01%7cxij6%40pitt.edu%7c2ba3b47c58f94ad6c26208d3b7599812%7c9ef9f489e0a04eeb87cc3a526112fd0d%7c1&sdata=GKjyBW2BRLufQdV5e48Cbup6R7sIoBoC4N7ZD8yURss%3d). Nature Methods 2016; 13: 310–318. PMID: 26901648. PMCID: PMC4854847.

C3. Developing Bayesian Network methods for medical prediction and clinical decision support: Medical diagnosis, prognosis determination, and treatment selection require prediction and decision making. Prediction is a critical step in the effort to recommend decisions that maximize the expected utility of the outcomes to the patient. Clinical data are becoming increasingly available in electronic form, providing tremendous opportunities for developing accurate classification and prediction methods. My accomplishments in this regard include: developed new methods for predicting the survival of breast cancer patients using integrated heterogeneous data and Bayesian network modeling [1, 2]; developed decision support models [3, 6]; developed novel causation learning algorithm that performed better than well-known existing causal learning algorithms when applied to simulated data [5]; and conducted ubiquitin site prediction using physicochemical properties of protein segments, an effort that is related to the study of DNA repair mechanisms [4].

1. Neapolitan RE, **Jiang X**. Study of integrated heterogeneous data reveals prognostic power of gene expression for breast cancer survival. PLoS ONE 2015; 10 (2): PMID: 25723490. PMCID: PMC4344205.
2. **Jiang X**, Xue D, Brufsky AM, Khan SA, Neapolitan RE. A new method for predicting patient survivorship using efficient Bayesian network learning. Cancer Informatics 2014; 13 (2):47-57. PMID: 24558297. PMCID: PMC3928477.
3. Neapolitan R, **Jiang X**, Ladner DP, Kaplan B. A primer on Bayesian decision analysis with an application to a kidney transplant decision. Transplantation 2016; 100(3): 489-496. PMCID: PMC4818954.
4. Cai B, **Jiang X.** Computational methods for ubiquitination site prediction using physicochemical properties of protein sequences. BMC Bioinformatics 2016; 17:116. PubMed PMID: 26940649. PMCID: PMC4778322.
5. Rathnam, CS, Lee, S, and **Jiang X**, An algorithm for direct causal learning of influences on patient outcomes. Artificial Intelligence in Medicine, Jan 2017; DOI:10.1016/j.artmed.2016.10.003.
6. **Jiang X**, Wells A, Brufsky A, Neapolitan RE, A Clinical Decision Support System Learned From Data to Personalize Treatment Recommendations Towards Preventing Breast Cancer Metastasis, PLoS One 14(3): e0213292, March 8, 2019; doi:10.1371/journal.pone.0213292

C4. Developing efficient machine learning algorithms: A difficulty when mining interactions from high-dimensional datasets concerns the *curse of dimensionality*. With the arrival of the big data era, developing efficient search algorithms has become one of the most important and challenging tasks in the machine learning community. I helped design a dynamic programming algorithm for spatial cluster detection [1]. I developed efficient Bayesian network-based epistasis learning algorithms, namely Multiple Beam Search (MBS) and Repeated Extended Greedy Approach to Learning (REGAL). I extended this research by developing a learning tool, which uses the BDeu score and MBS to learn SNP interactions, and then uses the BNPP (an evaluation tool I developed) to evaluate the probability of that association. The complete system is called Learning and Evaluation Association Patterns (LEAP) [2]. I showed that LEAP is an effective tool for extracting candidate interacting causal patterns from high-dimensional datasets and determining their probability. Striving to improve further on the interaction learning power of LEAP, I finally developed MBS-IGain [3], which uses both Bayesian network scoring and information theory to learn interactions. MBS-IGain substantially out-performed 9 other methods including LEAP, and may become the premier method for learning interactions. I guided and participated in the development of an original artificial neural network modification that improves computational efficiency of the learning steps and overall prediction performance [4]. I guided the efforts for using these algorithms not only to detecting epistatic biomarkers, but also to detecting the interactions between microRNA and mRNA in breast cancer.

1. Sverchkov Y, **Jiang X**, Cooper GF. Spatial cluster detection using dynamic programming. BMC Medical Informatics and Decision Making 2012; 12-22. PMCID: PMC3403878.
2. **Jiang X**, Neapolitan RE. LEAP: biomarker inference through learning and evaluating association patterns. Genetic Epidemiology 2015; 39(3):173–184. PMID: 25677188. PMCID: PMC4666609.
3. **Jiang X**, Jao J, Neapolitan RE. Learning predictive interactions using information gain and Bayesian network scoring. PLoS ONE 2015; 10(12). PMCID: PMC4666609.
4. Cai B, **Jiang X**. Novel Artificial Neural Network Method for Biomedical Prediction based on Matrix Pseudo-Inversion. Journal of Biomedical Informatics 2014; 48: 114-21. PMCID: PMC4004678.
5. Lee S, and **Jiang, X**, Modeling miRNA-mRNA interactions that cause phenotypic abnormality in breast cancer patients. PLoS One, August 2017; DOI:10.1016/j.ar.

C5. Understanding disease mechanisms at the molecular level: I have made substantial contribution to developing and applying methodology to tab into the alteration of signal transduction pathways (STPs) in diseases, by collaborating with other reseachers. In [1] we applied SPIA to 10 TCGA cancer datasets, and 157 KEGG STPs. We also performed a pan cancer analysis in which all 10 datasets were merged. We obtained results confirmed by the literature and new results. In particular, in the pan cancer analysis the 4 most notable pathways learned are all known to be major players in cancer. In [2] we developed a Bayesian network based method, called CASA, for learning aberrant STPs in disease from data, and applied it to a TCGA breast cancer dataset and 22 STPs, 12 of which are believed to be implicated in cancer. The top-ranking STPs that we learned are substantiated by previous research. In [4] we applied both CASA and SPIA (a non-BN based method for learning STPs from data) to a TCGA ovarian cancer dataset and 26 STPs, 20 of which are believed to be implicated in cancer. Our results for the two methods are consistent with the literature, and both confirm and complement each other. In [3, 5] we applied causal learning and inference in discovering STP and genomic and alterations in cancer.

1. Zeng Z, **Jiang X**, Li X, Wells A, Luo Y, Neapolitan RE, Conjugated equine estrogen and medroxyprogesterone acetate are associated with decreased risk of breast cancer relative to bioidentical hormone therapy and controls. PLoS One 13(5): e0197064, May 16, 2018; https://doi.org/10.1371/journal.pone.0197064.
2. Neapolitan R, Horvath CM, Jiang X. Pan-cancer analysis of TCGA data reveals notable signaling pathways. BMC Cancer; 2015; 15: 516.
3. Neapolitan R, Xue D, Jiang, X. Modeling the altered expression levels of genes on signaling pathways in tumors as causal Bayesian networks. Cancer Informatics; 2014; 13: 77-84.
4. Neapolitan R, Jiang X. Inferring aberrant signal transduction pathways in ovarian cancer from TCGA data. Cancer Informatics; 2014; Suppl. 1: 29-36.
5. Cai C, Cooper GF, Lu K, Ma X, …, **Jiang X**, … , Lu X, Systematic Discovery of the Functional Impact of Somatic Genome Alterations in Individual Tumors through Tumor-specific Causal Inference, PLoS Comput Biol 15(7): e1007088. July 5, 2019; https://doi.org/10.1371/journal.pcbi.1007088

**Complete List of Published Work:**

### <https://scholar.google.com/citations?pli=1&authuser=1&user=SNew-uYAAAAJ>

### D. Research Support

**Active**

**W81XWH1910495 BCRP Level II (PI: Jiang)** 8/15/2019-8/14/2022 3.6 months

**DoD** $1,169,513.00

Leveraging Deep Learning and Bayesian Networks to Identify Risk Factors and Support Personalized Prediction for Metastatic Breast Cancer

The purpose of this proposal is to explore new avenues for leveraging artificial intelligence to facilitate the improvement of our capabilities in detecting risk factors and making patient-specific prediction for breast cancer metastasis.

**Completed**

**R01 LM011663 NLM (Submitting PI: Jiang, co-PI: Neapolitan at Northwestern University)** 6/1/14 – 5/31/19 (NCE) , 4.8 months

**NLM/NIH          $1,955,585 total**

A New Generation Clinical Decision Support System

This project will develop a novel decision support system that utilizes both the clinical features and the genomic profile of a breast cancer patient to assist the physician in integrating information about a specific patient (diagnostic subtype, tumor stage and grade, age, comorbidities) to make therapeutic plans for the patient. We call this system A Clinical Decision Support System for Making Personalized Assessments and Recommendations Concerning Breast Cancer Patients.

**4100070287 (PI: Cooper)**  6/1/2015-5/31/2019 (NCE), 0.48 month

**PA DOH $5,042,791 total**

Big Data for Better Health (BD4BH) in Pennsylvania

Rapidly increasing volumes of molecular and electronic health record data in health care hold great promise for predicting patient outcomes, personalizing care, reducing geo-demographic disparities, and improving health. This project focused on developing and applying machine learning methods to large complex datasets to predict cancer outcomes based on constructed features that are biologically meaningful.

**K99/R00 LM010822 (PI: Jiang)** 10/1/2010-2/28/2015

**NIH/NLM. $764,810 total**

Detecting Genome-Wide Epistasis with Efficient Bayesian Network Learning

The goal of this career development award was to obtain training in the domains of genomics and cancer and to build a foundation for future R00 work in translational bioinformatics.

**U54HG008540 (PIs: Cooper, Bahar, Berg)**  9/15/2014-8/31/2018

**NIH/NLM. $12,799,098 total**

Center for Causal Modeling and Discovery of Biomedical Knowledge (CCMD) from Big Data.