

# Gaurav Pandey

University of California, Berkeley  
Department of Plant and Microbial Biology  
461 Koshland Hall  
Berkeley, CA 94720  
(612) 701-2494  
gaurav@compbio.berkeley.edu  
<http://compbio.berkeley.edu/people/gaurav>

1716 University Avenue  
Apartment 9  
Berkeley, CA 94704  
(612) 701-2494

RESEARCH INTERESTS Computational Biology, Data Mining, Machine Learning, Functional Genomics.

EDUCATION

- ◇ **University of Minnesota, Twin Cities**  
Degree: Ph.D. (Computer Science and Engineering)  
Date of Graduation: May, 2010  
Thesis: Data Mining Techniques for Enhancing Protein Function Prediction  
Advisor: Vipin Kumar

- ◇ **Indian Institute of Technology, Kanpur, India**  
Degree: B.Tech. (Computer Science and Engineering)  
Date of Graduation: May, 2004  
Thesis: Exploratory Learning of Geometrical Concepts

PROFESSIONAL EXPERIENCE

- ◇ **Post-doctoral Associate, University of California, Berkeley**  
*Duration:* June 2010-current  
*Advisor:* Steven Brenner, Department of Plant and Microbial Biology  
*Co-advisor:* Michael Jordan, Departments of Statistics and Electrical Engineering and Computer Science  
*Project:* Development and application of phylogenetics-based protein function methods
- ◇ **Research Assistant, University of Minnesota, Twin Cities, USA**  
*Duration:* September 2004-May 2010  
*Advisor:* Vipin Kumar, Department of Computer Science  
*Project:* Data Mining Techniques for Enhancing Protein Function Prediction
- ◇ **Summer Intern, Rosetta Inpharmatics (a wholly owned subsidiary of Merck & Co., Inc.), USA**  
*Duration:* May 2008 - August 2008  
*Advisor:* Bin Zhang, Department of Genetics (now at Sage Bionetworks)  
*Project:* An Integrative Multi-Network and Multi-Classifer Approach to Predict Genetic Interactions
- ◇ **Summer Intern, SAP Labs Inc, USA**  
*Duration:* May 2006 - August 2006  
*Advisor:* Rakshit Daga, Design Services Team  
*Project:* Extraction of Structured Knowledge from Unstructured Business Documents
- ◇ **Summer Intern, University of Sydney, Australia**  
*Duration:* May 2003 - July 2003  
*Advisor:* Sanjay Chawla and Joseph Davis, School of Information Technologies  
*Project:* Design and analysis of association rule networks
- ◇ **Summer Intern, Pertinence Data Intelligence, Paris**  
*Duration:* May 2002 - July 2002  
*Advisor:* Michele Sebag, LRI, Universite Paris-Sud  
*Project:* Clustering on the basis of frequent itemsets

PUBLICATIONS ◇ **Book**

1. Gaurav Pandey, Chad L. Myers, Michael Steinbach and Vipin Kumar, *Computational Approaches for Protein Function Prediction*, to be published by John Wiley & Sons, Inc in 2011.

◇ **Book Chapter**

1. Chandrika Kamath, Nikhil Wale, George Karypis, Gaurav Pandey, Vipin Kumar *et al*, *Scientific Data Analysis*, in *Scientific Data Management: Challenges, Existing Technology, and Deployment*, Editors Arie Shoshani and Doron Rotem, CRC Press, 2009.

◇ **Journal Articles****Published/Accepted**

1. Gaurav Pandey\*, Bin Zhang\*, Aaron Chang, Chad L. Myers, Jun Zhu, Vipin Kumar and Eric E. Schadt, *An Integrative Multi-Network and Multi-Classifer Approach to Predict Genetic Interactions*, PLoS Computational Biology, 6(9): e1000928, 2010 (\* Equal contribution) [Cited by Nature Biotech feature article as "one of the notable breakthroughs in computational biology from 2010".]
2. Gang Fang, Gaurav Pandey, Manish Gupta, Michael Steinbach and Vipin Kumar, *Mining Low-Support Discriminative Patterns from Dense and High-dimensional Data*, Accepted by IEEE Transactions on Knowledge and Data Engineering (TKDE), 2010.
3. Gaurav Pandey, Chad L. Myers, and Vipin Kumar, *Incorporating Functional Interrelationships into Protein Function Prediction Algorithms*, BMC Bioinformatics, 10:142, 2009 (Determined "Highly Accessed" by the publisher).
4. Gaurav Pandey, Sanjay Chawla, Simon Poon, Bavani Arunasalam and Joseph Davis, *Association Rules Network: Definition and Applications*, Statistical Analysis and Data Mining, Vol 1, No 4, pp. 260-279, 2009.
5. Hui Xiong, Gaurav Pandey, Michael Steinbach and Vipin Kumar, *Enhancing Data Analysis with Noise Removal*, in IEEE Transactions on Knowledge and Data Engineering (TKDE), Vol 18, No 3, pp. 304-319, 2006.

**Under review**

1. Gaurav Pandey\*, Yanji Xu\*, Ross Haller, Konjit Betre, Joel W. Slaton, Akhouri A. Sinha, LeeAnn Higgins, Lorraine B. Anderson and Michael J. Wilson, *Proteomic Analysis of Human Prostate Secretion*, under review (\* Equal contribution).
2. Gaurav Pandey\*, Sahil Manocha\* and Vipin Kumar, *Systematic Evaluation of Common Neighborhood Similarity Measures for Protein Interaction Networks*, under review (\* Equal contribution).
3. Jeremy Bellay, Gowtham Atluri, . . . , Gaurav Pandey, . . . , Vipin Kumar and Chad L. Myers, *Epistasis in context: Understanding genetic interactions through an exhaustive modular decomposition*, under review.

◇ **Conferences Articles**

1. Gang Fang, Rui Kuang, Gaurav Pandey, Michael Steinbach, Chad L. Myers and Vipin Kumar, *Subspace Differential Coexpression Analysis: Problem Definition and a General Approach*, Proceedings of the Pacific Symposium on Biocomputing (PSB), pp. 145-156, 2010.
2. Gaurav Pandey, Gowtham Atluri, Michael Steinbach, Chad L. Myers and Vipin Kumar, *An Association Analysis Approach to Biclustering*, Proceedings of ACM SIGKDD, pp. 677-685, 2009.
3. Gowtham Atluri, Rohit Gupta, Gang Fang, Gaurav Pandey, Michael Steinbach and Vipin Kumar, *Association Analysis Techniques For Bioinformatics Problems*, Proceedings of the 1<sup>st</sup> International Conference on Bioinformatics and Computational Biology (BICoB), pp. 1-13, 2009 (Invited paper).

4. Gaurav Pandey, Lakshmi N. Raamakrishnan, Michael Steinbach and Vipin Kumar, *Systematic Evaluation of Scaling Methods for Gene Expression Data*, Proceedings of IEEE BioInformatics and BioMedicine (BIBM), pp 376-381, 2008 (Extended version published as UMN CS Technical Report 07-015).
5. Gaurav Pandey, Michael Steinbach, Rohit Gupta, Tushar Garg and Vipin Kumar, *Association Analysis-based Transformations for Protein Interaction Networks: A Function Prediction Case Study*, Proceedings of ACM SIGKDD, pp. 540-549, 2007 (Also selected for presentation at the ISMB 2008 Highlights Track).
6. Gaurav Pandey, Himanshu Gupta and Pabitra Mitra, *Stochastic Scheduling of Active Support Vector Learning Algorithms*, Proceedings of the ACM Symposium on Applied Computing (SAC), pp. 38-42, Santa Fe, USA, 2005.
7. Gaurav Pandey, Ankit Anand and Harish Karnick, *EUCLID: A System for the Exploratory Discovery of Geometrical Properties of Triangles*, Proceedings of the 2<sup>nd</sup> Indian International Conference on Artificial Intelligence (IICAI), pp. 2759-2776, Pune, India, 2005.
8. Sanjay Chawla, Joseph Davis and Gaurav Pandey, *On Local Pruning of Association Rules Using Directed Hypergraphs*, Proceedings of the 20<sup>th</sup> IEEE International Conference on Data Engineering (ICDE), pp. 832, Boston, USA, 2004.
9. Gaurav Pandey, Chaitanya Mishra and Paul Ipe, *Tansen : A System for Automatic Raga Identification*, Proceedings of the 1<sup>st</sup> Indian International Conference on Artificial Intelligence (IICAI), pp. 1350-1363, Hyderabad, India, 2003.

◇ **Workshop/Meeting Articles**

1. Gowtham Atluri, Jeremy Bellay, Gaurav Pandey, Chad Myers and Vipin Kumar, *Discovering Coherent Value Bicliques In Genetic Interaction Data*, Proceedings of the 9th International Workshop on Data Mining in Bioinformatics (BIOKDD), 2010.
2. Maneesh Bhargava, Trisha L. Becker, Pratik D. Jagtap, LeeAnn Higgins, Lorraine Anderson, Gaurav Pandey, Michael S. Steinbach, Vipin Kumar, Gary L. Nelstuen, David H. Ingbar and Christine H. Wendt, *Alveolar Epithelial Cell Proteome in Hyperoxic Lung Injury*, International Conference of the American Thoracic Society, 2009.
3. Gaurav Pandey, Gowtham Atluri, Gang Fang, Rohit Gupta, Michael Steinbach and Vipin Kumar, *Association Analysis Techniques For Analyzing Complex Biological Data Sets*, Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS), 2009.
4. Gaurav Pandey\*, Bin Zhang\*, Aaron Chang, Jun Zhu, Vipin Kumar and Eric E. Schadt, *Integrative Multi-Network Approach to Predict Synthetic Lethal Interactions*, Proceedings of the 5th Annual RECOMB Satellite Meeting on Regulatory Genomics and Systems Biology, 2008 (\* Equal contribution).
5. Gaurav Pandey, Gowtham Atluri, Michael Steinbach and Vipin Kumar, *Association Analysis Techniques for Discovering Functional Modules from Microarray Data*, Proceedings of the ISMB special interest group meeting on Automated Function Prediction, 2008 (Also published as Nature Precedings 10.1038/npre.2008.2184.1).
6. Gaurav Pandey and Vipin Kumar, *Incorporating Functional Inter-relationships into Protein Function Prediction Algorithms*, Proceedings of the ISMB special interest group meeting on Automated Function Prediction, 2007.
7. Rohit Gupta, Tushar Garg, Gaurav Pandey, Michael Steinbach and Vipin Kumar, *Comparative Study of Various Genomic Data Sets for Protein Function Prediction and Enhancements using Association Analysis*, Proceedings of the Workshop on Data Mining for Biomedical Informatics, help in conjunction with the SIAM International Conference on Data Mining, 2007.

8. Gaurav Pandey and Rakshit Daga, *On Extracting Structured Knowledge from Unstructured Business Documents*, in Proc IJCAI Workshop on Analytics for Noisy Unstructured Text Data, pp. 155-162, 2007.

◇ **Technical Reports**

1. Gaurav Pandey, Vipin Kumar and Michael Steinbach, *Computational Approaches for Protein Function Prediction: A Survey*, TR 06-028, Department of Computer Science and Engineering, University of Minnesota, Twin Cities, 2006.

◇ **Patents**

1. Rakshit Daga and Gaurav Pandey, *Structure Extraction from Unstructured Documents*, US Patent 7562088, 2009.
2. Rakshit Daga and Gaurav Pandey, *Determination of Document Similarity*, US Patent Application 20080162455.

◇ **Availability:** All the published papers are available from my homepage.

ACADEMIC  
HONORS

- ◇ Awarded the Doctoral Dissertation Fellowship by the University of Minnesota Graduate School for 2009-10.
- ◇ KDD 2007 paper selected for presentation at the ISMB 2008 Highlights track.
- ◇ Finalist for the IBM PhD fellowship 2008-09 (one of about 30 candidates across all CS departments), and one of two candidates nominated by the CS Department, UMN as a candidate for the same award in 2007-08.
- ◇ Awarded travel grants to attend several conferences and research schools:
  1. NSF-sponsored ISCB travel grant for attending ISMB 2008.
  2. Travel grant for attending a workshop on Search and Knowledge Building for Biological Datasets organized by IPAM, UCLA.
  3. Minnesota Plant Genomics Institute (MPGI) travel grants for attending ISMB 2007 and KDD 2009.
  4. International Joint Conference on Artificial Intelligence (IJCAI), 2007.
  5. Quantitative Approaches to Gene Regulatory Systems, Center for Theoretical Biological Physics, UCSD (Could not attend).
  6. Modeling and Mining of Networked Information Spaces Winter School, Banff Research Center, Canada (Could not attend).
- ◇ Recipient of the ACM SIGMOD undergraduate scholarship for attending the SIGMOD/PODS 2004 joint conference at Paris.
- ◇ Recipient of the prestigious NTSE scholarship awarded by National Council of Educational Research and Training in 1998.
- ◇ Secured 10th rank in the Indian National Mathematics Olympiad, 1999 and was a contender for a place in the Indian team for IMO-1999 and IMO-2000.
- ◇ Secured 5th rank in the Indian National Chemistry Olympiad, 2000 and was a contender for a place in the Indian team for IChO-2000.
- ◇ Recipient of a scholarship from the National Board for Higher Mathematics (India) in appreciation of a good performance in a mathematics nurture camp held in ISc, Chennai in 2001.

MAJOR  
PROJECTS

- ◇ **Development and Application of Phylogenetics-based Protein Function Prediction Methods (June 2010 - Current)** *Advisors:* Profs. Steven Brenner and Michael Jordan, University of California, Berkeley, USA. *Summary:* In this project, we are building on the SIFTER (Statistical Inference of Function Through Evolutionary Relationships) platform to integrate phylogenetics-based protein function approaches with other

approaches based on systems-level data, such as protein interaction networks and next-generation sequencing data. In addition, we are also exploring the applicability of SIFTER to a large set of unannotated proteins, particularly by participating in the CAFA (Critical Assessment of Functional Annotations) initiative.

◇ **Incorporation of Gene Ontology Information into Protein Function Prediction Algorithms** (Summer 2007 - Spring 2009)

*Advisors:* Profs. Vipin Kumar and Chad L. Myers, Department of CSE, University of Minnesota, USA

*Summary:* Most protein function prediction algorithms, particularly data mining and machine learning-based ones, assume the functional classes being used for prediction to be disjoint. However, with the growing use of Gene Ontology, which is a hierarchical DAG-based arrangement of function classes, it is clear that these algorithms do not incorporate the complete biological truth about functional classes. Thus, in this project, we developed methods to explicitly incorporate the structure and semantics of Gene Ontology into standard protein function prediction algorithms, in order to enhance the accuracy of the predictions made. This work was published in BMC Bioinformatics in 2009.

◇ **Development of Novel Association Analysis Frameworks for Analyzing Complex Biological Data Sets** (Fall 2007 - Spring 2010)

*Advisor:* Prof. Vipin Kumar, Department of CSE, University of Minnesota, USA

*Summary:* Association analysis is one of the most popular analysis paradigms in data mining. Despite the solid foundation of association analysis techniques and their potential applications, they face some challenges when used to analyze biological data. In this work, we are developing different types of association patterns that can address some of these challenges. In our work published in ACM SIGKDD 2009, we developed a novel biclustering method based on these techniques, that can be used to effectively find functional modules enriched by specific GO terms. We are also currently developing similar techniques for identifying specific types of network motifs in genetic interaction data.

◇ **An Integrative Multi-Network and Multi-Classifer Approach to Predict Genetic Interactions** (Summer 2008 - Summer 2010)

*Advisor:* Dr. Bin Zhang, Department of Genetics, Rosetta Inpharmatics, LLC, Seattle, USA

*Summary:* Genetic interactions occur when a combination of mutations results in a surprising phenotype. These relationships capture functional redundancy, and thus are important for predicting function, dissecting protein complexes into functional pathways, and exploring the mechanistic underpinnings of common human diseases. Synthetic sickness and lethality are the most studied types of genetic interactions in yeast. However, even in yeast, only a small proportion of gene pairs have been tested, due to the large number of possible combinations of gene pairs. To expand the set of known synthetic lethal (SL) interactions, we have devised an integrative multi-network and multi-classifier approach for predicting these interactions, that significantly improves upon the performance of the existing approaches. We have also applied this approach to predict the genetic interactions between the known transcription factors (TFs) and uncovered a number of novel SL interactions between TFs, which are well supported by the available knowledge about these TFs. The paper discussing this study is currently under review by PLoS Computational Biology.

◇ **Biomarker Discovery from Quantitative Mass Spectrometry Data** (Spring 2007 - Current)

*Collaborator:* Prof. Michael Wilson, Department of Laboratory Medicine and Pathology, University of Minnesota, USA

*Summary:* A recent advance in the field of proteomics is the introduction of the iTRAQ technology, which enables not only the identification of the proteins in a sample, but also enables the quantification of their abundance. This new data gives important indications towards the differential behavior of proteins in different classes of individuals, such as

diseased versus healthy classes. In this particular project, we are working on identifying potential biomarkers for detecting prostate cancer occurrence from data generated using this technology. We have adopted a comprehensive statistical methodology that addresses several challenges with this data, such as a large number of missing values, and between-samples variations in the scales of expression values.

- ◇ **Pre-processing of Genomic Data to Enhance Functional Information** (Fall 2006 - Spring 2010)  
*Advisor:* Prof. Vipin Kumar, Department of CSE, University of Minnesota, USA  
*Summary:* Recently, the availability of a wide variety of high-throughput biological data has provided very useful insights into the functions of various proteins and the mechanisms leading to their accomplishment. However, this data often contains noise and unwanted artifacts such as spurious interactions and inconsistencies between different components of a given data set. Here, we have applied data mining-based pre-processing techniques such as normalization and graph transformations to improve protein function prediction from microarray data and protein interaction networks. The results of these techniques have been published in the ACM SIGKDD 2007 and IEEE BIBM 2008 conferences.
- ◇ **Data Cleaning Techniques for Very Noisy Data** (Fall 2004 - Spring 2005)  
*Advisor:* Prof. Vipin Kumar, Department of CSE, University of Minnesota, Twin Cities, USA  
*Summary:* In this project, a new technique, HCleaner, for noise removal from very noisy data was developed and was compared with several outlier detection techniques using novel validation methodologies. Experiments on a variety of data sets proved the efficacy of this approach for obtaining more accurate results from unsupervised learning techniques, namely clustering and association analysis.
- ◇ **Extraction of Structured Knowledge from Unstructured Business Documents** (Summer 2006)  
*Advisor:* Rakshit Data, SAP Labs Inc, Palo Alto, CA  
*Summary:* Efficient management of unstructured text data is a major concern of business organizations. In this direction, we proposed a novel approach to extract structured knowledge from large corpora of unstructured business documents. The approach is based on the observation that a significant fraction of these documents are created using the cut-copy-paste method, and thus, it is important to factor this observation into business document analysis projects.
- ◇ **Exploratory Learning of Geometrical Concepts** (B.Tech. Project)  
*Advisor:* Prof. Harish Karnick, Department of CSE, IIT, Kanpur, India  
*Summary:* In this project, we designed an exploratory learning algorithm using which an agent can automatically learn geometrical concepts like the geometrical properties of a triangle, starting from certain primitive categories like points, lines and angles, and operators and relations defined over them. Experiments generated several valid geometrical properties. This research was published in the proceedings of the IICAI 2005 conference.
- ◇ **Active Learning for Support Vector Machines** (Spring 2004)  
*Advisor:* Prof. Pabitra Mitra, Department of CSE, IIT, Kanpur, India  
*Summary:* In this project, we used game theoretic approaches to design two stochastic scheduling algorithms for active learning with support vector machines. The algorithms produced encouraging results on benchmark UCI datasets, and this work was presented at the ACM Symposium on Applied Computing 2005.
- ◇ **Association Rules Networks and Their Application in the Analysis of Open Source Data** (Summer 2003)  
*Advisors:* Profs. Sanjay Chawla and Joseph Davis, School of IT, University of Sydney, Australia  
*Summary:* This was a research project aimed at designing an algorithm for laying out a set of association rules into a network so as to explain a certain goal. These networks

were used to explain the reasons for the radical growth of the Open Source movement. A preliminary version of this work was accepted as a poster at the IEEE ICDE 2004 conference, and an extended version was published in 2009 in the Statistical Analysis and Data Mining journal.

- ◇ **Tansen-A System for Automatic Raga Identification** (Spring 2003)  
*Advisors:* Prof. Harish Karnick and Prof. Amitabha Mukherjee, Department of CSE, IIT, Kanpur  
*Summary:* This was a project involving considerable research into Indian classical music and machine learning techniques suitable for learning and identifying a special class of compositions known as the Ragas. This research was published in the proceedings of the ICAI 2004 conference.

- SERVICES
- ◇ Member, Program committee of ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB) 2011
  - ◇ Reviewing services
    - *Journals:* Bioinformatics, KAIS, IEEE TKDE, Statistical Analysis and Data Mining, Pattern Analysis and Algorithms, Nucleic Acids Research, Amino Acids, Information Sciences.
    - *Conferences:* ACM KDD, IEEE ICDM, IEEE ICDE, AAI, PAKDD, IEEE/ACM SuperComputing, ICAI, SIAM DMBio, ICPP, IEEE BIBM, IEEE BIBE.
  - ◇ Administrative services
    - Part of a department committee to establish a PhD student evaluation process for the CSE department at UMN.
    - Part of a CSE@UMN department committee to discuss and decide on general issues affecting graduate students.

- ADDITIONAL SKILLS
- ◇ Programming Languages: Matlab, Java, C++, C, Perl.
  - ◇ Operating Systems: Mac OS, Linux, Windows, Solaris.
  - ◇ Well familiar with the use of both commercial and open source software.
  - ◇ Experienced in working with teams and excellent people and communication skills.

- REFERENCES
- |                                                                                                                                                                                 |                                                                                                                                                                                    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Prof. Vipin Kumar</b><br/>Computer Science and Engineering<br/>University of Minnesota, Twin Cities<br/>Minneapolis, MN 55414<br/>kumar@cs.umn.edu<br/>(612) 625-0726</p> | <p><b>Prof. Chad L. Myers</b><br/>Computer Science and Engineering<br/>University of Minnesota, Twin Cities<br/>Minneapolis, MN 55414<br/>cmyers@cs.umn.edu<br/>(612) 624-8306</p> |
| <p><b>Dr. Eric Schadt</b><br/>Pacific Biosciences<br/>Menlo Park, CA 94025<br/>eschadt@pacificbiosciences.com<br/>(650) 521-8000</p>                                            | <p><b>Dr. Bin Zhang</b><br/>Sage Bionetworks<br/>Seattle, WA 98109<br/>bin.zhang@sagebase.org<br/>(206)667-2119</p>                                                                |

**More references available if required.**