

Sayan Mukherjee

Curriculum Vitæ
2008 November

1. Personal Information:

Name: Sayan Mukherjee	Addr: Duke University DSS 112 Old Chemistry Building, Box 90251 Durham, NC 27710 USA
Born: Calcutta, India	Tel: 1-919-684-4608
Citizen: USA	Fax: 1-919-684-8594
	Email: sayan@stat.duke.edu
	Web: http://www.genome.duke.edu/labs/mukherjee/

2. Education:

Fellow MIT, 2001-2004 Broad Institute and Center for Biological & Computational Learning Sloan Post Doctoral Fellow	Ph.D.. MIT, 2001 Center for Biological & Computational Learning
Advs: Tomaso Poggio, Todd Golub	Advs: Tomaso Poggio
	Diss: Application of Statistical Learning Theory to DNA Microarray Analysis
MS. Columbia University, 1995 Department of Applied Mathematics and Physics	B.S.E. Princeton University, 1992 Department of Electrical Engineering

3. Academic Appointments:

Duke University:	
2006- Assistant Professor, Primary	Department of Statistical Science
2008- Assistant Professor, Secondary	Department of Biostatistics and Bioinformatics
2005- Assistant Professor, Secondary	Department of Computer Science
2004-06 Assistant Professor, Secondary	Department of Statistical Science
2004-06 Assistant Professor, Primary	Department of Biostatistics and Bioinformatics
2004- Investigator	Institute for Genome Sciences & Policy

4. Awards, Honors, Prizes:

2008	Young Researcher Award – International Indian Statistical Association
2002	Sloan Foundation Postdoctoral Fellow – Computational molecular biology

5. Publications since 2000:

Published, in press or accepted:

1. Q Wu, F Liang, S Mukherjee, 2008. Local sliced inverse regression. (accepted Proceedings of Neural Information Processing Systems).
2. KS Garman, E Edelman, CR Acharya, M Grade, J Gaedcke, S Sud, K Walters, G Ginsburg, W Barry, AM Dieh, D Provenzale, BM Ghadimi, T Ried, D Hsu, JR Nevins, S Mukherjee, A Potti, 2008. A genomic approach to dissecting colon cancer progression yields biologic insights into therapeutic opportunities. (accepted Proceedings of the National Academy of Science).
3. CR Acharya, DS Hsu, CK Anders, A Anguiano, KH Salter, KS Walters, RC Redman, SA Tuchman, CA Moylan, S Mukherjee, WT Barry, HK Dressman, GS Ginsburg, KP Marcom, KS Garman, GH Lyman, JR Nevins, A Potti, 2008. Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. *JAMA*. **299**, 13, 1574-87.
4. E Edelman, J Guinney, JT Chi, PG Febbo, S Mukherjee, 2008. Modeling Cancer Progression via Pathway Dependencies. *PLoS Comput. Biol.* **4**,2, e28.
5. H Bonnefoi, A Potti, M Delorenzi, L Mauriac, M Campone, M Tubiana-Hulin, T Petit, P Rouanet, J Jasem, E Blot, V Farmer, S André, CR Acharya, S Mukherjee, D Cameron, J Bergh, JR Nevins, RD Iggo, 2007. Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EROTC 10994/BIG 00-01 clinical trial. *Lancet Oncol.* **8**, 12, 1071–1078.
6. F Liang, S Mukherjee, M West, 2007. Understanding the use of unlabelled data in predictive modelling. *Statistical Science* **22**, 2, 189-205.
7. JT Chi, EW Rodriguez, Z Wang, DSA Nuyten, S Mukherjee, M van de Rijn, MJ van de Vijver, T Hastie, PO Brown, 2007. Gene Expression Programs of Human Smooth Muscle cells: tissue-specific differentiation and prognostic significance in breast cancers. *PLoS Genet.* **3**, 9, e164.
8. NS Pillai, Q Wu, F Liang, S Mukherjee, RL Wolpert, 2007. Characterizing the function space for Bayesian kernel models. *J. Mach. Learn. Res.* **8**, 1769–1797.
9. L Goh, SK Murphy, S Mukherjee, TS Furey, 2007. Genomic sweeping for hypermethlyated genes. *Bioinformatics* **23**, 3, 281-288.
10. S Mukherjee and Q Wu, 2006. Estimation of gradients and coordinate covariation in classification. *J. Mach. Learn. Res.* **7**, 2481–2514.
11. S Mukherjee and DX Zhou, 2006. Learning coordinate covariances via gradients. *J. Mach. Learn. Res.* **7**, 519-549.
12. Z Wang, HF Willard, S Mukherjee, TS Furey, 2006. Evidence of influence of genomic DNA sequence on human X chromosome inactivation. *PLoS Comput. Biol.* **2**, 9, 979-988.
13. E Edelman, A Porrello, J Guinney, B Balakumaran, A Bild, PG Febbo, S Mukherjee, 2006. Analysis of sample set enrichment of sets of genes for individual genome-wide expression profiles. *Bioinformatics* **22**, 14, 108–116.
14. A Potti, S Mukherjee, R Petersen, HK Dressman, A Bild, J Koontz, R Kratzke, MA Watson, M Kelley, GS Ginsburg, M West, DH Harpole, JR Nevins, 2006. A genomic strategy to refine prognosis in early stage non-small cell lung carcinoma. *N. Engl. J. Med.* **355**, 6, 570–580.
15. D Tropea, G Kreiman, A Lyckman, S Mukherjee, H Yi, S Horng, M Sur, 2006. Gene expression changes and molecular pathways mediating activity-dependent plasticity in visual cortex. *Nat. Neuro.* **9**, 5, 660–668.
16. S Mukherjee, P Niyogi, T Poggio, R Rifkin, 2006. Statistical learning: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.* **25**, 1-3, 161 - 193.
17. A Rakhlin, S Mukherjee, T Poggio, 2005. Stability results in learning theory. *Anal. App.* **3**, 4, 397–417.
18. P Golland, F Liang, S Mukherjee, D Panchenko, 2005. Permutation tests for classification. In *Proceedings of the Conference on Learning Theory (P Auer and R Meir)*, Springer-Verlag, Heidelberg,

501–515.

19. A Rakhlin, D Panchenko, S Mukherjee, 2005. Risk Bounds for Mixture Density Estimation. *ESAIM: Prob. Stat.* **9**, 220–229.
20. A Subramanian, P Tamayo, V Mootha, S Mukherjee, BL Ebert, M Gillette, A Paulovich, SL Pomeroy, TR Golub, ES Lander, JP Mesirov, 2005. Gene Set Enrichment Analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 43, 15545–50.
21. A Sweet-Cordero, S Mukherjee, A Subramanian, H You, J Roix, C Ladd-Acosta, JP Mesirov, TR Golub, T Jacks, 2005. An oncogenic *KRAS2* expression signature identified by cross-species gene-expression analysis. *Nat. Genet.* **37**, 1, 48–55.
22. T Poggio, R Rifkin, S Mukherjee, P Niyogi, 2004. Learning Theory: general conditions for predictivity. *Nature* **428**, 419–422.
23. R Berger, PG Febbo, PK Majumder, JJ Zhao, S Mukherjee, T Campbell, WR Sellers, TM Roberts, M Loda, TR Golub, WC Hahn, 2004. Androgen-Induced differentiation and tumorigenicity of human prostate epithelial cells. *Cancer Research* **64**, 8867–8875.
24. S Mukherjee, P Tamayo, S Rogers, R Rifkin, A Engle, C Campbell, TR Golub, JP Mesirov, 2003. Estimating dataset size requirements for classifying DNA microarray data. *J. Comput. Biol.* **10**, 2, 119–142.
25. R Rifkin, S Mukherjee, P Tamayo, S Ramaswamy, CH Yeang, M Reich, T Poggio, ES Lander, TR Golub, JP Mesirov, 2003. An Analytical Method for Multi-Class Cancer Classification. *SIAM Reviews* **45**, 4, 706–723.
26. S Mukherjee, 2003. Classifying microarray data using support vector machines. In *Understanding and Using Microarray Analysis Techniques: A Practical Guide*, (D Berar, W Dubitzky, M Granzow), Springer-Verlag, Heidelberg, 166–185.
27. S Mukherjee, R Rifkin, T Poggio, 2002. Regression and classification with regularization. In *Lecture Notes in Statistics: Nonlinear Estimation and Classification, Proceedings from MSRI Workshop*, (DD Denison, MH Hansen, CC Holmes, B Mallick, B Yu), Springer-Verlag, 171, 107–124.
28. T Poggio, S Mukherjee, R Rifkin, A Rakhlin, A Verri, 2002. b. In *Uncertainty in Geometric Computations*, (M Winkler, M Niranjana), Kluwer Academic Publishers, 131–141.
29. LD Miller, PM Long, L Wong, S Mukherjee, LM McShane, ET Liu, 2002. Optimal gene expression analysis by microarrays. *Cancer Cell* **2**, 353–361.
30. SL Pomeroy, P Tamayo, M Gaasenbeek, LM Sturua, M Angelo, ME McLaughlin, JHY Kim, LC Goumnerova, PM Black, C Lau, JC Allen, D Zagzag, JM Olson, T Curran, C Wetmore, JA Biegel, T Poggio, S Mukherjee, R Rifkin, A Califano, G Stolovitzky, DN Louis, JP Mesirov, ES Lander, TR Golub, 2002. Gene expression-based classification and outcome prediction of central nervous system embryonal tumors. *Nature* **415**, 436–442.
31. O Chapelle, V Vapnik, O Bousquet, S Mukherjee, 2002. Choosing multiple parameters for support vector machines. *Mach. Learn. - Special Issue on Support Vector Machines* **46**, 131–159.
32. L Peshkin and S Mukherjee, 2001. Bounds on sample size for policy evaluation in Markov environments. In *Proceedings of the Conference on Learning Theory, (DP Helmbold and R Williamson)*, Springer-Verlag, Heidelberg, 616–630.
33. B Heisele, T Serre, S Mukherjee, T Poggio, 2001. Feature reduction and hierarchy of classifiers for fast object Detection in Video Images. In *Computer Vision and Pattern Recognition, IEEE Computer Society*, 18–24.
34. J Sadr, S Mukherjee, K Thoresz, P Sinha, 2001. The fidelity of local ordinal encoding. In *Advances in Neural Information Processing Systems, (TG Dietterich, S Becker, Z Ghahramani)*, MIT Press, 1279–1286.
35. S Ramaswamy, P Tamayo, R Rifkin, S Mukherjee, CH Yeang, M Angelo, C Ladd, M Reich, E Latulippe, JP Mesirov, T Poggio, W Gerald, M Loda, ES Lander, TR Golub, 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 26, 15149–15154.

36. CS Yeang, S Ramaswamy, P Tamayo, R Rifkin, S Mukherjee, M Angelo, M Reich, ES Lander, JP Mesirov, TR Golub, 2001. Molecular classification of multiple tumor types. *Bioinformatics* **17**, Suppl 1, 316–322.
37. J Weston, S Mukherjee, O Chapelle, M Pontil, T Poggio, V Vapnik, 2000. Feature selection for SVMs. In *Advances in Neural Information Processing Systems*, (SA Solla, TK Leen, KR Muller), MIT Press, 668–674.
38. M Pontil, S Mukherjee, F Girosi, 2000. On the noise model of support vector machine regression. In *Proceedings of Algorithmic Learning Theory*, (H Arimura, S Jain, A Sharma), Springer-Verlag, 316–324.

Under review:

39. L Abatangelo, R Maglietta, A Distaso, D Annarita, TM Creanza, S Mukherjee, Ancona, 2008. Comparative study of gene set enrichment methods. (under review in BMC Bioinformatics).
40. S Georgiev, K Jayasurya, S Mukherjee, U Ohler, 2008. (C)ERMIT: integrating genomewide evidence of binding with overrepresented sequence patterns. (under review in Proceedings of the National Academy of Science).
41. E Edelman, K Garman, A Potti, S Mukherjee, 2008. Making mountains out of molehills: moving from single gene to pathway based models of colon cancer progression. (under review in PLoS One).
42. K Mao, Q Wu, F Liang, S Mukherjee, 2008. Non-parametric Bayesian model for simultaneous dimension reduction and regression on manifolds. (under review in Biometrika).
43. F Liang, K Mao, M Liao, S Mukherjee, M West, 2008. Non-parametric Bayesian kernel models. (under review in Bayesian Analysis).
44. Q Wu, F Liang, S Mukherjee, 2008. Local sliced inverse regression. (under review in Journal of Computational and Graphical Statistics).
45. Q Wu, F Liang, S Mukherjee, 2007. Regularized sliced inverse regression for kernel models. (under review in Statistica Sinica).
46. S Mukherjee, Q Wu, and D-X Zhou, 2007. Learning gradients and feature Selection on manifolds. (under review in Bernoulli).
47. J Guinney, Q Wu, S Mukherjee, 2007. Estimating variable structure and dependence in Multi-task learning via gradients. (under review in Machine Learning).
48. Q Wu, J Guinney, M Maggioni, S Mukherjee, 2007. Learning gradients: predictive models that infer geometry and dependence. (under review in Journal of Machine Learning Research).

In preparation:

49. S Mukherjee. Statistical Learning: algorithms and theory. Monograph compiling extensive class notes for Statistical Learning lectures at Duke. *In preparation*.
50. J Guinney, PG Febbo, M Maggioni, P Magwene, S Mukherjee (2008). Decomposing Gene Networks using Multiscale Graphical Models. *In preparation*.
51. S Lunagomez, S Mukherjee, R Wolpert (2008). Conditional Independence Models via Discrete Morse Theory. *In preparation*.

6. Sponsored Research:*Current:*

- 2007-10 PI, NSF DMS-0732260: *Probabilistic Models and Geometry for High Dimensional Data.*
- 2007-12 co-PI, NIH R01 CA125618-01, PI JT Chi: *Gene Expression Programs of Lactic Acidosis in Human Cancers.*
- 2007-12 co-PI, NIH R01 CA123175-01A1, PI PG Febbo: *mTOR Therapy in Prostate Cancer: Signatures of Response and Biology Resistance.*
- 2008-12 co-PI, NIH R01 CA138265-01, PI D Kirch: *Dissecting Mechanisms of Metastasis Through Comparative Systems Genetics.*
- 2007-12 co-PI, NIH Systems Biology Center Grant, PI P Benfy: *Duke Center for Systems Biology.*
- 2005-10 Collaborator, NHGRI Center for the Study of Public Genomics, PI Huntington Willard: *CEER Grant for Ethical, Legal, and Social Issues in Genomics.*
- 2008-10 Collaborator, DOD, PI J Zhu: *Exploring the Pathogenic and Therapeutic Implications of Aberrant Splicing in Breast Cancer.*

Pending:

- Pending PI, NIH R01: *Integration of Heterogeneous Cancer Genomic Data Using Multi-task Learning.* (Submitted 2008).
- Pending PI, NSF: *Modeling complex phenotypes using geometry and statistical dependencies for high-dimensional (genomic) data.* (Submitted 2008).
- Pending co-PI, NIH R01, PI M Reichart: *Tethered Cytokines for Immune Modulation.* (Submitted 2008).
- Pending co-PI, NIH R01, PI B Capel: *Transcriptional Network Analysis of Sex Determination in Mice.* (Submitted 2008).

7. Professional Appointments and Service:*Editorial appointments:*

- 2008- Associate Editor: *Trends in Computational Biology and Bioinformatics*

Conference and workshop service:

- 2008 Program committee: Artificial Intelligence and Statistics
- 2008 Session chair and organizer: *Biological Applications of Machine Learning* in Eastern North American Region of the International Biometric Society
- 2007 Senior program committee: International Conference on Machine Learning
- 2007 Session chair: *Statistical Machine learning and High Dimensional Inference* in Joint Statistics Meeting
- 2007 Local organizer: *Random Matrices and High-dimensional Inference* at SAMSI
- 2006 Session chair: *Learning theory* in Western North American Region of the International Biometric Society
- 2006 Session chair: *Learning theory* in Institute of Mathematical Statistics annual meeting
- 2005 Co-organizer: *Random Matrices and Computer Models* at SAMSI
- 2005 Coordinator: *Random Graphs and Stochastic Computation* at SAMSI
- 2005 Instructor: *First School on Computational Cell Biology* at University of Urbino
- 2005 Instructor: *Machine Learning Summer School* at Toyota Technical Institute
- 2001 Organizer: *Workshop in bioinformatics* at Neural Information Processing Systems

Review service:

Applied Bioinformatics, Annals of Human Genetics, British Journal of Cancer, Bioinformatics, Genome Biology, Genomics, Genome Research, PLoS Computational Biology, Journal of the American Medical Association, PLoS Genetics, Nature Biotechnology, Molecular Biology and Evolution, Nature Methods, Nature, Proceedings of the National Academy of Sciences, Research in Computational Molecular Biology, Statistics in Medicine, Annals of Statistics, Algorithmic Learning Theory, Advances in Computational Mathematics, Biometrics,

Biometrika, Biostatistics, Computational Learning Theory, Complexity, Foundations of Computational Mathematics, International Conference on Machine Learning, IEEE Transactions on Information Theory, IEEE Transactions on Pattern Analysis and Machine Intelligence, Journal of the American Statistical Association, Journal of Approximation Theory, Journal of Machine Learning Research, Machine Learning, Neural Information Processing Systems, Trends in Neural Networks, Statistical Science

8. University service:

2008- Information Technology Advisory Council
2008- Computing Committee: Department of Statistical Science
2008 First Year PhD Exam Coordinator: Department of Statistical Science
2007-08 Academic Council
2007- Outreach Coordinator: Duke Center for Systems Biology
2007-08 Admissions Committee: Computer Science Department
2005 Seminar Series Coordinator: Institute of Statistics and Decision Sciences
2005- Student Advisory Committee: Computational Biology and Bioinformatics Program
2005- Curriculum Committee: Computational Biology and Bioinformatics Program
2005-07 Admissions Committee: Computational Biology and Bioinformatics Program
2005- Genome Academy Instructor: Institute for Genome Sciences & Policy

9. Invited Addresses since 2000:

2008 Mathematics Department: Georgia Institute of Technology.
2008 Computer Science Department: Rensselaer Polytechnic Institute.
2008 Statistics and Human Genetics Departments: University of Chicago.
2008 Statistics Department: North Carolina State University.
2008 SIAM Conference on the Life Sciences: Montreal, Canada.
2008 Approximation and Learning Theory: Oberwolfach, Germany.
2008 National Cancer Institute: Bethesda, MD.
2008 International Conference on Interdisciplinary Mathematical & Statistical Techniques: Memphis, TN.
2008 ENAR Spring Meeting: Arlington, VA.
2008 Statistics Department: Virginia Polytechnic University.
2008 Information Theory and Applications: University of California at San Diego.
2008 Markov Chain Monte Carlo in Theory and Practice: Bormio, Italy.
2007 Geometric and Topological Approaches to Data Analysis: Chicago, IL.
2007 Approximation and Learning in High Dimensions: College Station, TX.
2007 Nonparametrics: Columbia, SC.
2007 Duke Systems Biology Symposium
2007 A Journey Through Computation: Genoa, Italy.
2007 International Conference on Computational Harmonic Analysis: Shanghai, China.
2007 Joint Statistics Meeting: Salt Lake City, UT.
2007 Statistics Department: Ohio State University.
2007 Computer Science Department: Ohio State University.
2007 American Institute of Mathematics: Paolo Alto, CA.
2006 Indian Statistical Institute, Kolkata, India.
2006 Institute of Mathematical Sciences: Rio de Janeiro, Brazil.
2005 First School on Computational Cell Biology: Urbino, Italy.
2005 Machine Learning Summer School: Chicago, IL.
2005 Foundations of Computational Mathematics: Universidad de Cantabria, Santander, Spain.

- 2004 Institute of Mathematical Sciences: National University of Singapore, Singapore.
- 2004 Mathematical Theory of Networks and Systems: Katholieke Universiteit Leuven, Belgium.
- 2003 Institute of Pure and Applied Mathematics: Los Angeles, CA
- 2002 Foundations of Computational Mathematics: Institute for Mathematics and its Applications, Minneapolis, MN.
- 2002 Support Vector Machines 2002: Niagra Falls, Ontario, Canada.
- 2002 Cambridge Healthtech Institute: Boston MA.
- 2002 Institute of Mathematical Sciences: National University of Singapore, Singapore.
- 2000 European School of Genetic Medicine: Sestri Levante, Italy.

10. Professional Affiliations:

- American Statistical Association (ASA)
- International Statistical Institute (Bernoulli Society)
- Institute of Mathematical Statistics (IMS)

11. Teaching Experience at Duke:

- MATH 288 – Random Graphs and Statistical Inference
- STA 294 – Geometry and Random Matrices.
- STA 293 – Topics in Statistics: Statistical Learning – Algorithms and Theory.
- STA 270 – Statistical Methods for Computational Biology.
- STA 113 – Probability and Statistics for Engineering.
- STA 114 – Statistical Inference.
- STA 294 – Geometry and High-dimensional Inference, SAMSI class cross-listed as NCSU MA/ST 810G, UNC Math 891.

12. Patents and Software:

Patents:

- Molecular Cancer Diagnosis Using Tumor Gene Expression Signatures
- Estimating Dataset Size Requirements for Classifying DNA Microarray Data

Software:

<http://www.genome.duke.edu/labs/mukherjee/research/>

13. Dissertation committees:

Departmental abbreviations:

- Biology – BIO
- Biomedical Engineering – BME
- Brain and Cognitive Science – BCS
- Computational Biology and Bioinformatics – CBB
- Computer Science – CS
- Computer Science and Artificial Intelligence Laboratory – CSAIL
- Department of Statistical Science – DSS
- Electrical and Computer Engineering – ECE
- Math – MTH
- Molecular Genetics and Microbiology – MGM
- Department of Theoretical Physics – TP

PhD advisees:

1. Elena Edelman	(CBB, 2004-present)	Duke University
2. Justin Guinney	(CBB, 2005-present)	Duke University
3. Stoyan Georgiev	(CBB, 2005-present)	Duke University
4. Kai Mao	(DSS, 2005-present)	Duke University
5. Simón Lunagómez	(DSS, 2005-present)	Duke University

PhD committees (Statistical science):

1. Gavino Puggioni	(DSS, 2004-present)	Duke University
2. Melanie Wilson	(DSS, 2005-present)	Duke University
3. Kristian Lum	(DSS, 2006-present)	Duke University
4. Eric Vance	(DSS, 2008)	Duke University
5. Natesh Pillai	(DSS, 2008)	Duke University
6. Scotland Leman	(DSS, 2007)	Duke University
7. Yuhong Wu	(DSS, 2006)	Duke University
8. Jingqin Luo	(DSS, 2006)	Duke University
9. Chris Hans	(DSS, 2005)	Duke University
10. Ming Liao	(DSS, 2005)	Duke University

PhD committees (Other departments):

1. Dan Runcie	(BIO, 2006-present)	Duke University
2. Andreas Pfenning	(CBB, 2006-present)	Duke University
3. Ken Yokoyama	(CBB, 2005-present)	Duke University
4. Iulian Pruteanu-Malinici	(ECE, 2005-present)	Duke University
5. Shashidhara Ganjug	(CS, 2005-present)	Duke University
6. Amit Patel	(CS, 2005-present)	Duke University
7. Bei Wang	(CS, 2005-present)	Duke University
8. Karen Hayden	(MGM, 2005-present)	Duke University
9. Julia Chen	(MGM, 2005-present)	Duke University
10. Qi An	(ECE, 2005-present)	Duke University
11. Jake Bouverie	(BCS, 2005-present)	MIT
12. Todd Wasson	(CBB, 2004-present)	Duke University
13. David Crosslin	(CBB, 2004-present)	Duke University
14. Nicole Johnson	(CBB, 2004-present)	Duke University
16. Elizabeth Rach	(CBB, 2004-present)	Duke University
17. David Orlando	(CBB, 2003-present)	Duke University
18. Haige Shen	(CBB, 2008)	Duke University
19. Jonathan Jesneck	(BME, 2007)	Duke University
20. Yingchun Liu	(TP, 2007)	Lund University
21. Stanley Bileschi	(CSAIL, 2006)	Duke University
22. Shaorong Chang	(ECE, 2005)	Duke University
23. Savina Andonova	(Math, 2003)	Boston University