

Aug 2006

CURRICULUM VITAE
Jane Fridlyand, Ph.D.

PERSONAL INFORMATION

Date and Place of Birth:

08/29/1973, Leningrad, Russia

Current Title and Department:

*Assistant Professor
Department of Epidemiology and Biostatistics
Center for Bioinformatics and Molecular Biostatistics
Breast Oncology Program
Cancer Center Biostatistics Core
UCSF Comprehensive Cancer Center*

Business Address (Mt. Zion / China Basin)

*2340 Sutter Str, N412 / 185 Berry street
University of California, San Francisco
San Francisco CA 94143-0128 / 94107-0560
Phone: (415)476-0168
Fax: (415)514-8159
jane@cc.ucsf.edu
<http://www.biostat.ucsf.edu/jane>*

EDUCATION

1993-1995	UC Berkeley	BA	Applied Mathematics
1995-1996	UC Berkeley	MA	Statistics
1996-2001	UC Berkeley	PhD	Statistics

EMPLOYMENT

Principal Positions Held

Sep 1995-May 1996	Life Science Division Lawrence Berkeley National Laboratory	<i>Graduate Student Researcher</i>
Sep 1996-May 1997	Department of Statistics University of California at Berkeley	<i>Graduate Teaching Assistant</i>
Sep 1998-Dec 1998	Bioinformatics Group Walter and Eliza Hall Institute Melbourne, Australia	<i>Graduate Student Researcher</i>
Sep 1997-Aug 1999	Sequence Analysis Group Human Genome Project Lawrence Berkeley National Laboratory	<i>Graduate Student Researcher</i>
Sep 1999-May 2001	Department of Statistics University of California at Berkeley	<i>Graduate Student Researcher</i>
Aug 2001-Mar 2004	Cancer Research Institute University of California at San Francisco	<i>Programmer Analyst IV</i>
Apr 2004-present	Dept. of Epidemiology and Biostatistics University of California at San Francisco	<i>Assistant Professor</i>

Ancillary Positions Held

1999-2001 Roche Molecular Systems *Statistical Consultant*
Alameda, CA

HONORS AND AWARDS

2006 ENAR Junior Investigator Workshop Award
2000 ENAR Student Competition award -- XXth International Biometrics Conference
1998-2001 Program for Mathematics in Molecular Biology, -- Burroughs Welcome
Fellowship
1995-1997 Eugene Cobes Fellowship
1993-1995 UC Regents' Scholarship

KEYWORDS/AREAS OF INTEREST

Statistics, cancer, computational biology, microarray data, data mining, classification, genomics, identifying interactions, prediction, clustering, array CGH, gene expression, biomarkers, meta-analysis, magnetic resonance spectroscopy

PROFESSIONAL ACTIVITIES

Manager, Cancer Center Biostatistics Core, UCSF. The main responsibility is to supervise the master's level biostatistician in the core and to provide consulting and guidance for the projects in genomics and cancer. Co-organizer of invited session at ENAR conference, Florida, 2006.

PROFESSIONAL ORGANIZATIONS

Memberships

2001-present American Statistical Association

INVITED MEETINGS AND PRESENTATIONS

International and National

International Biometrics Conference, Berkeley, CA, July 2000 (contributed, award-winning talk)

Statistical Methods in Microarray Analysis, National University of Singapore, Singapore, Jan 2004 (invited talk)

Joint Statistical Meetings, New York City, Aug 2002 (invited talk)

3rd Early Detection Research Network Scientific Workshop, Bethesda, MD, June 2004 (invited talk)

National Biospecimen Network, Baltimore, MD, July 2004 (Prostate Cancer SPORE Task Force Informatics Meeting, invited meeting participant)

Assessing Human Germ Cell Mutagenesis in the Post-Genome Era, Bar Harbor, MN, Sep 2004 (invited talk)

3rd Breast SPORE Roundtable Meeting, Nashville, Tennessee, Nov 2005 (Organs System Branch of the National Cancer Institute, invited meeting participant).

ENAR Biometric Conference, Tampa, FL, March 2006 (session organizer and presenter)

Statistics for Gene and Protein Expression Workshop, Gotheburg, Sweden, May 2006 (invited speaker and tutorial lecturer)

10th European Workshop of Molecular Cytogenetics in Human Solid Tumors, La Grande Motte, France, June 2006 (invited speaker)

International Biometric Conference, Montreal, FL, July 2006 (invited speaker)

Regional and other invited presentations

Genetics Seminar, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, Sep, Nov, 1998 (invited talks)

UC Berkeley Biostatistics Department Seminar, Berkeley, CA, May 2000 (invited talk)

Special Seminar, Roche Molecular Systems, Alameda, CA, May 2001 (invited talk)

Stanford University Workshop in Biostatistics, Stanford, CA, Oct 2001 (invited talk)

UC Berkeley Biostatistics Department Seminar, Berkeley, CA, Nov 2001 (invited talk)

Cancer Research Institute Retreat, Santa Cruz, CA, Oct 2003 (invited talk)

UC Berkeley Biostatistics Department Seminar, Berkeley, CA, Sep 2003 (invited talk)

Workshop in Biostatistics, Stanford University, Stanford, CA, Oct 2003 (invited talk)

BIRS Workshop: Statistical Science for Genome Biology, Banff, Canada, Aug 2004 (discussion moderator)

Breast Outcome Program meeting, San Francisco, CA, Jan 2005 (invited poster)

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, Feb 2005 (invited talk)

Department of Biostatistics, John Hopkins University, March 2005 (invited talk)

17th Annual California Association of Regional Cancer Registries Annual Scientific Conference, San Francisco, CA, March 2005 (invited talk)

Department of Biostatistics, University of Pittsburg, April 2005 (invited talk)

Biomarkers in HIV and Cancer Research Workshop, Mathematical Biosciences Institute, Ohio State University, April 2005 (invited talk)

Terry Speed's group meeting, UC Berkeley, April 2005 (invited talk)

Visiting scholar at Oncology Department, University of Cambridge, UK, July 2005 (invited visit, two seminars)

ASA San Francisco Chapter Meeting, San Francisco, CA, Oct 2005 (invited talk)

Breast Outcome Program meeting, San Francisco, CA, Jan 2006 (invited poster)

Department of Biostatistics, UC Berkeley, Berkeley, April 2006 (invited talk)

Department of Biostatistics, Stanford University, Stanford, June 2006 (invited talk)

Bioinformatics workshop, MDACC, Houston, September 2006 (invited talk)

SOFTWARE

March 2004-present

aCGH R-package for the analysis of array CGH data. Contains tools for visualization, testing and breakpoint determination.

Bioconductor release 1.4.0 +.

WORKSHOPS AND LECTURES

Program in Mathematics and Molecular Biology Short Course, Berkeley, CA, June 2000
Teaching Assistant.

Analysis of Gene Expression Microarray Data Short Course, CBMB/QB3,
San Francisco, CA, Nov 2003. Lecturer

Analysis of Microarray Data Short Course, Cancer Center Biostatistics Core/Center for Biology and

Molecular Biostatistics (CBMB), San Francisco, CA, March 2004. Lecturer.

Statistics 246, UC Berkeley Statistics Department, April 2004. Guest Lecturer

Advanced Microarray Analysis Course, Denmark Technical University
Elsinore, Denmark, May 2004. Lecturer

Statistical Methods for Microarray Analysis, Cancer Center, UCSF, June 2004. Lecturer.

Analysis of Gene Expression Microarray Data Short Course, CBMB/QB3,
San Francisco, CA, Nov 2004. Lecturer

Cancer Bioinformatics for “*Cancer Biology: Carcinogenesis & Cancer Progression*” module for
Hematology/Oncology Fellows, UCSF, May 2006, Lecturer

Introduction to the analysis of expression and array CGH data for *Mathematical Systems Biology of
Cancer MSRI Workshop*, Berkeley, May 2006, Lecturer

Introduction to the analysis of the array CGH data for *Statistics for Gene and Protein Expression
Workshop*, Gotheburg, Sweden, May 2006, Lecturer

PEER REVIEWED PUBLICATIONS

1. International Human Genome Sequencing Consortium. Initial Sequencing and Analysis of the Human Genome. (*Nature* , 860 – 921, 2001). **Listed as a member of JGI sequencing team.**
2. Symons, R. C. A., Daly, M. J., **Fridlyand, J.**, Speed, T. P., Cook, W. D., Gerondakis, S., Harris, A. W. and Foote, S. J. Multiple genetic loci modify susceptibility to plasmacytoma-related morbidity in Emu-v-abl transgenic mice. (*PNAS*, 2002, No. 99).
3. **Fridlyand J.**, Dudoit S. and Speed, T.P. Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. (*Journal of American Statistical Association*, March 2002, Vol. 97, No. 457). **First two authors contributed equally**
4. **Fridlyand J.** and Dudoit S. A prediction-based resampling method to estimate the number of clusters in a dataset. (*Genome Biology*, 2002, Vol. 3, No. 7). **Both authors contributed equally**
5. Veltman, J. A., **Fridlyand, J.**, Pejavar, S., Olshen, A. B., Korkola, J. E., DeVries, S., Carroll, P., Kuo, W. , Pinkel, D., Albertson, D., Cordon-Cardo, C., Jain, A. N. and Waldman, F. W. Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors (*Cancer Research*, 2003, Vol, 63, No. 11).
6. Paris, P. L., Albertson, D. G., Alers, J. C., Andaya, A., Carroll, P., **Fridlyand, J.**, Jain, A. N., Kamkar, S., Kowbel, D., Krijtenburg, P. J., Pinkel, D., Schroder, F. H., Vissers, K. J., Watson, V. J., Wildhagen, M.F., Collins, C. and Van Dekken, H. High-resolution analysis of paraffin-embedded and formalin-fixed prostate tumors using comparative genomic hybridization to genomic microarrays . (*American Journal of Pathology*, 2003, Vol. 162, No. 3).
7. Snijders, A. M., Nowee, M. E., **Fridlyand, J.**, Piek, J. M. J., Dorsman, J. C., Jain, A. J., Pinkel, D., van Diest, P. J., Verheijen, R. H. and Albertson D. G. Genome-wide-array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing CCNE1 in Falopian tuba carcinoma. (*Oncogene*, 2003, Vol. 22, No. 27).

8. Snijders, A. M., **Fridlyand, J.**, Mans, D., Segraves, R., Jain, A. N., Pinkel, D. and Albertson D. G. Shaping of tumors and drug-resistant genomes by instability and selection. (Oncogene, 2003, Vol. 22, No. 28).
9. Hackett, C. S, Hodgson, G., Law, M. E., **Fridlyand, J.**, Osoegawa, K., de Jong, P. L., Nowak, N. J., Pinkel, D., Alberston, D. G., Jain, A. N., Jenkins, R., Gray, J. W. and Weiss, W. A. Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. (Cancer Research, 2003, Vol. 63, No. 17).
10. Korkola, J. E., DeVries, H., **Fridlyand, J.**, Hwang, E. S., Estep, A. I. H., Chen, Y., Dairkee, S. H., Jensen, R. M. and Waldman, F. M. Differentiation of lobular vs ductal breast carcinoma by expression microarray analysis. (Cancer Research, 2003, Vol. 63, No. 21)
11. Maldonado J.L., **Fridlyand J.**, Patel H., Jain A. N., Busam K., Kageshita T., Ono T., Albertson D. G., Pinkel D. and Bastian B. C. Determinants of BRAF mutations in primary melanomas. (Journal of National Cancer Institute, 2003, Vol. 95, No. 24).
12. Nelson, D. O. and **Fridlyand, J.** Designing meaningful measures of read length for data produced by DNA sequencers. (IMS Lecture Notes -- Monograph Series, 2003, Vol. 40).
13. **Fridlyand J.** and Dudoit S. Bagging to improve the accuracy of a clustering procedure. (Bioinformatics, 2003, vol. 19, No. 9). **Both authors contributed equally.**
14. Maldonado J. L., Timmerman L., **Fridlyand J.** and Bastian B. C. Mechanisms of cell cycle arrest in spitz nevi with constitutive activation of the MAP-Kinase pathway. (American Journal of Pathology, May 2004, V.164(5), pp. 1783-7).
15. Hager, J. H., Hodgson, G., **Fridlyand, J.**, Hariono, S., Gray, J. W. and Hanahan D. Oncogene expression and genetic background influence the frequency of DNA copy abnormalities in mouse pancreatic islet cell carcinomas. (Cancer Research, 2004, Vol. 64 No. 7).
16. Nakao, K., Mehta, K. R., **Fridlyand, J.**, Moore, D. H., Jain, A. N., Lafuente, A., Wiencke, J. W., Terdiman, J. P. and Waldman, F. M. High resolution analysis of colorectal cancer by array-based comparative genomic hybridization. (Carcinogenesis, Aug; 2004, V. 25(8), pp. 1345-57)
17. Paris, P. L., Andaya, A., **Fridlyand, J.**, Jain, A. N., Weinberg, V., Kowbel, D., Brebner, J. H., Simko, J., Watson, J.E., Volik, S., Albertson, D. G., Pinkel, D., Alers, J.C., Van Der Kwast, T. H., Vissers, K. J., Schroder, F. H., Wildhagen, M. F., Febbo, P. G., Chinnaiyan, A. M., Pienta, K. J., Carroll, P. R., Rubin, M. A., Collins, C., Van Dekken, H. Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. (Hum Mol Genet., Jul 2004 Jul 1, V. 13, pp. 1303-13)
18. **Fridlyand J.**, Snijders A., Pinkel, D., Albertson D. G. and Jain, A. Application of Hidden Markov Models to the analysis of the array CGH data. (Special Genomic Issue of Journal of Multivariate Analysis, June 2004, V. 90, pp. 132-153)
19. Snijders AM, Nowak NJ, Huey B, **Fridlyand J.**, Law S, Conroy J, Tokuyasu T, Demir K, Chiu R, Mao JH, Jain AN, Jones SJ, Balmain A, Pinkel D, Albertson DG. Mapping segmental and sequence variations among laboratory mice using BAC array CGH. (Genome Research, 2005 Feb;15(2):302-11)

20. Snijders AM, Schmidt BL, **Fridlyand J**, Dekker N, Pinkel D, Jordan RC, Albertson DG. Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. (*Oncogene*. 2005 Jun 16;24(26):4232-42.)
21. Ching T-T; Maunakea AK; Jun P; Hong C; Zardo G; Pinkel D; Albertson DG; **Fridlyand J**; Mao J- H; Shchors K; Weiss WA; Costello JF. Epigenome analyses using BAC microarrays identifies evolutionary conservation of tissue-specific methylation of SHANK3. (*Nature Genetics*, 37(6), 645-51, 2005,)
22. Willenbrock H. and **Fridlyand J**. A comparison study: applying segmentation to array CGH data for downstream analysis. (*Bioinformatics*, 21(22):4084-91,2005). **Fridlyand is a senior author**
23. Curtin JA, **Fridlyand J**, Kageshita T, Patel H, Busam K, Kutzner H, Cho KH, Aiba S, Brocker EB, LeBoit PE, Pinkel D and Bastian BC. Distinct sets of genetic alterations in melanoma. (*New England Journal of Medicine*, 353(20):2135-47, 2005)
24. Blaveri E., Brewer J. L., Royds Gupta R., **Fridlyand J.**, DeVries S., Koppie T., Pejevar S., Mejat K., Carroll P., Simko JP and Waldman FM . Bladder cancer stage and outcome defined by array based comparative genomic hybridization. (*Clinical Cancer Research*, 11 (19 Pt 1), 7012-22, 2005)
25. Chao RC, Pyzel U., **Fridlyand J.**, Teel L., Haaga J., Borowsky A., Horvai A., Kogan S., Bonfias J., Huey B., Jacks TE, Albertson D., Shannon K. Therapy induced malignant neoplasms in *Nf1* mutant mice. (*Cancer Cell*, 2005, 8(4):337-48, 2005)
26. Rubinstein J., **Fridlyand J.**, Shen A., Aldape K., Ginzinger D., Batchelor T., Treseler P., Bergere M., McDermott M., Prados M., Krach J., Okada C., Hyun W., Parikh S., Haqq C., Shuman M. Gene expression and angiotropism in primary CNS lymphoma. (*Blood*, Jan 2006)
27. Albertson DG, Snijders AM, **Fridlyand J**, Jordan R, Pinkel D, Schmidt B. .Comprehensive analysis of tumors by array comparative genomic hybridization: more is better. (*Cancer Research*, 66(7):3955-6, 2006)
28. **Fridlyand J**, Snijders AM, Ylstra B, Li H, Olshen AB, Segraves R, Dairkee S, Tokuyasu TA, Ljung BM, Jain AN, McLennan J, Ziegler J, Chin K, Devries S, Feiler H, Gray JW, Waldman F, Pinkel D, Albertson DG. Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer* 6:96, 2006. **First three authors contributed equally**
29. Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo WL, Gwadry F, Ajay, Kouros Mehr H, **Fridlyand J**, Jain A, Collins C, Nishizuka S, Tohon G, Roschke A, Gehlhaus K, Kirsch I, Scudiero DA, Gray JW, Weinstein JN. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Molecular Cancer Therapy* 5(4):853-67, 2006.
30. Reynolds PA, Sigaroudinia M, Zardo G, Wilson MB, Benton GM, Miller CJ, Hong C, **Fridlyand J**, Costello JF, Tlsty TD. Tumor suppressor P16INK4A regulates polycomb-mediated DNA hypermethylation in human mammary epithelial cells. *Journal of Biological Chemistry*, 2006 Aug 25;281(34):24790-802.
31. Chen Y., Toland AE, McLennan J, **Fridlyand J**, Crawford B, Costello JF, Ziegler JL. Lack of germline promoter methylation in BRCA1 negative families with familial breast cancer. *Genetic Testing*, 2006, in press.

NON-PEER REVIEWED PUBLICATIONS

Book Chapters

1. Dudoit, S. and **Fridlyand, J.** Classification in microarray experiments. (Analysis of microarray experiments, Chapman & Hall/CRC, 2003, edited by T. P. Speed).
2. Dudoit, S. and **Fridlyand, J.** Classification in microarray experiments (Understanding and Using Microarrays Analysis Techniques: A Practical Guide, Kluwer, 2003, edited by D. Berrar, W. Dubitzky and M. Grnazow).

Ph.D Dissertation

1. **Fridlyand, J.** Resampling methods for variable selection and classification: applications to genomics. (<http://www.stat.berkeley.edu/users/janef/index.html>, June 2001)

Abstracts

1. **Fridlyand J.** , Snijders A., Pinkel, D., Albertson D. G. and Jain, A. N. Statistical issues in the analysis of the array CGH data. (Computations Systems Bioinformatics 2003 Conference Proceedings, IEEE Computer Society).

PATENTS ISSUED OR PENDING

4 patents issued or pending

OTHER CREATIVE ACTIVITIES

All teaching lectures and workshops given are posted on the WWW for use by other investigators. Freely available software for the analysis of the aCGH data is available from the Bioconductor website.

RESEARCH PROGRAM

The main focus of my research has been on the development of statistical methodology and applications in biological and medical problems. More specifically, my work has been centered on developing and applying data mining techniques to the high-dimensional datasets arising in gene expression and copy number microarray, methylation and genotyping studies. I am also interested in developing methods for combining different types of data across genomic platforms (e.g., gene expression, array CGH and methylation) and tumor imaging (MRS) data. My ongoing research includes:

- Development of the methods for variable selection.

Identifying the loci responsible for variation in quantitative or binary traits such as tumor size, cancer subtype or survival status, is a problem of great importance to biologists. One of the main features of the genomic datasets is their unfavorable “p” (number of variables) to “n” (number of samples) ratio. Many important genetic variables affect the trait of interest via epistasis rather than on their own, and

identification of such genes is notoriously hard when p/n ratio is large. In my PhD thesis, I have developed a novel approach for discovering interacting loci in the context of mouse linkage studies. There, I have utilized binary decision trees in combination with powerful aggregation approaches by shifting the focus of the analysis from prediction to variable selection. This approach has been successfully used as an exploratory tool in the study of plasmacytoma-related morbidity in Emu-v-abl transgenic mice (Symon et al, 2002). I am interested in applying these ideas for determining variable importance to variable selection in even more complicated situations of microarray studies and generalizing them for pathway discovery.

- Development of the methods for accurate class prediction.

Accurate class prediction is a problem of the utmost importance in cancer classification. There, biologists are often interested in developing genetic methods for tumor subtype identification and prognosis. Given the complexity of the microarray tumor data involving unfavorable ratio of the number of variables to the number of samples, large number of sets of highly correlated variables (e.g., co-regulated genes), and high between-patient heterogeneity, the question arises as to whether classic statistical methods for discrimination can be used for this new type of data. Together with my Ph. D. adviser Dr. Terry Speed and collaborator Dr. Sandrine Dudoit, we have conducted a thorough comparison study with several publicly available gene expression datasets each containing known cancer subtypes. We were able to perform a comparison of a number of traditional discrimination methods such as K-Nearest-Neighbors (k-NN) and Linear Discriminant Analysis (LDA) with the state-of-the-art machine learning approaches including application of bagging and boosting. We have demonstrated that in a typical gene expression dataset, the otherwise successful machine learning classifiers do not have an advantage over traditional statistical methods noted for their high bias and low variance, e.g. k-NN or LDA with the assumption of uncorrelated variables (Dudoit, Fridlyand et al, 2002). It is very likely that as the number of samples in a typical microarray dataset increases, the machine learning methods capable of exploring the space of interaction will gain an edge over standard discrimination approaches. In particular, I plan to evaluate the performance of highly publicized classification method, Random Forests, in terms of its relative prediction accuracy and variable selection ability. Together with Drs. Spellman and Gray, I am currently involved in the cancer genome profiling project which, if funded, is going to produce data on many thousands of tumors which would allow us to explore these issues.

- Development of the methods for novel class discovery.

Cluster analysis involves the search through data for observations that are similar enough to each other to be grouped together. When a clustering algorithm is applied to a set of observations, a partition of the data is obtained whether or not the data exhibit a true or "natural" grouping structure. This fact causes no problems if clustering is done for obtaining a practical grouping of the given set of objects, for instance for organizational purposes. However, if interest lies more in the recognition of an unknown classification of the data, an artificial clustering is not acceptable, and therefore clusters resulting from the algorithm must be investigated for their relevance. Apart from descriptive, graphical or exploratory methods, this task can be performed by using probabilistic models and suitable statistical significance tests. Discovery of novel tumor classes using gene expression data is one example where the need to reliably estimate the number of clusters and accurately allocate observations arises. With my collaborator Dr. Sandrine Dudoit, we proposed to apply resampling methods to (i) estimate the number of clusters in a dataset and (ii) improve accuracy of the cluster assignment. The approach to (i) uses ideas from discriminant analysis. Since the clusters obtained from cluster analysis are eventually used for prediction purposes, it is natural to apply discrimination techniques in clustering. For (ii), bootstrap aggregation is used to improve cluster accuracy and to assign confidence to the labels of the individual observations. We have successfully demonstrated the utility of both approaches on simulated data and real microarray datasets by conducting careful comparison studies of our methods with the available methods. This work is presented in Dudoit and Fridlyand (2002, 2003). I am interested in continuing to build upon the ideas proposed in these manuscripts and applying them to the tumor datasets of my collaborators at the UCSF

Cancer Center (Drs. Gray, Bastian and Albertson) where one of the objectives is to identify homogeneous patient subsets to improve targeted drug development for breast, ovarian and melanoma cancers. In particular, some of these ideas find an interesting application to the Magnetic Resonance Spectroscopy (MRS) data where objectives include characterizing within-patient tumor heterogeneity based on the metabolite measurements as well as refining tumor margins for better targeted treatment. This work is joint with Drs. Lu, T. McKnight and J. Hwang.

- Development of the methods for the analysis of the array CGH data.

The development of solid tumors is associated with acquisition of complex genetic alterations, indicating that failures in the mechanisms that maintain the integrity of the genome contribute to tumor evolution. Thus, one expects that the particular types of genomic alterations seen in tumors reflect underlying failures in maintenance of genetic stability, as well as selection for changes that provide growth advantage. Microarray-based comparative genomic hybridization (array CGH) can be used to investigate genomic alterations. The computational task is to map and characterize the number and types of copy number alterations present in the tumors, and so define copy number phenotypes as well as to associate them with known biological markers. To utilize the spatial coherence between nearby clones, I proposed to use an unsupervised Hidden Markov Models approach. The clones are partitioned into states which represent underlying copy number of the group of clones. The structural changes in a tumor genome may be recorded and characterized computationally using the above methodology. The method is described in Fridlyand et al, 2004 and has been successfully applied to a number of cell line and primary tumor datasets. This research has greatly benefited from the continuing input of my biological collaborators Drs. Albertson, Pinkel and Snijders. My current ongoing effort is focused on refining the methodology and incorporating it in the aCGH package of R/BioConductor that I have developed with P. Dimitrov, a PhD student at UC Berkeley working with me. Additionally, I am very interested in using the HMM approach for data reduction, so that the resulting lower-dimensional dataset is used as an input to classification and variable selection procedures as well as procedures for combining different data types. Together with the visiting Ph.D student H. Willenbrock, we have also investigated approaches for identifying discrete levels of the copy number across the entire genome and their use in the downstream analysis (Willenbrock and Fridlyand, 2005) We have demonstrated that this leads to increase in power and better detection of copy number alteration events. We also hope to utilize this approach for identification of early versus late events in tumor development. With my biological collaborators, we have successfully applied the above methodology to efficiently identify regions of homozygosity and heterozygosity in backcross mice using array CGH data (Snijders et al, 2005). Currently, together with P. Dimitrov, we are extending my previous work by developing a variable duration HMM for copy number profile segmentation (HsMM). Written as a graphical model, this flexible framework allows for flexible incorporation of the focal aberrations, subclonal events and natural extensions to the allele-specific copy number analyses. One of the issues is computational and we are working on efficient approximations to the full solution to make algorithm applicable to the high density arrays.

- Development of the methods for combining clinical, copy number and expression data for identification of driver genes and discovery of novel pathways.

More cancer datasets are becoming available containing copy number, gene expression and methylation measurement as well as clinical information on the tumor samples. The approaches discussed in the current literature for combining different types of data focus almost exclusively on identifying dosage effect of individual gene via computing correlation coefficient between copy number and mRNA levels of that gene. Clinical information has been largely ignored in this context. We have used our methodology for quantifying genomic instability together with available transcriptional and clinical information to identify the groups of high risk patients and show the functional groups of genes which transcriptional activity is associated with increase in instability. (Fridlyand et al 2006, Chin et al, 2006 (accepted), Neve et al, 2006 (accepted)). This work is continuing to be done in collaboration with Drs. Albertson, Gray and Waldman using breast and ovarian cancer datasets.

Another interesting question is whether one can use different types of genomic data to improve our understanding of the pathway activation, identify novel gene connections in the known pathways and group samples according to whether they are involved in a given pathway and their specific mechanism of involvement. Initially we propose to focus on the known KEGG pathways and with the view of being able to augment them with new genes. Our approach involves building a prediction model for the mRNA level of each pathway genes using all the available information on that gene as well as transcriptional data on the remaining pathway genes. Currently we are investigating the question of model selection for these type of data since existing approaches tend to overfit in the context of equal number of samples and variables, i.e. identify many false connections. The number of connections among genes serves as a measure for the pathway activation status. We also build the shortest path among pairs of genes and compare it to the known pathway structure. The same concept allows us to test whether the gene not known to be in the pathway actually belongs there. Using regression model we can ask which samples are potential outliers and thus may not be involved in a given pathway and whether these samples have particular clinical characteristics. Finally, all the approaches are implemented with transcriptional data only thus allowing us to assess the contribution of the DNA level data (copy number and methylation) when RNA level data are available. This work is done in collaboration with Drs. Yeh, Costello, Gray, Albertson and Pinkel.