KA YEE YEUNG, PH.D.

Email: kayee.yeung@gmail.com

I have extensive experience in the design of algorithms for the mining and integration of big data. I am proficient in finding patterns, building predictive models, feature selection, inferring relationships and networks from large complex data. My scientific experience, blending both computer science and statistics, focuses on the development, customization and application of data mining, machine learning and pattern recognition methods.

EDUCATION

1998-2001	Ph.D. Computer Science, University of Washington, Seattle, WA *Advisor:* Professor Walter L. Ruzzo.
1996-1998	M.S. Computer Science, University of Washington, Seattle, WA *Advisor:* Professor Richard Karp.
1995-1996	M. Mathematics in Computer Science, University of Waterloo, Ontario, Canada *Advisor:* Professor Ian Munro.
1992-1995	B. Mathematics in Computer Science and Actuarial Science from University of Waterloo, Ontario, Canada

EMPLOYMENT EXPERIENCE

7/11-present	**Research Associate Professor**, Department of Microbiology, University of Washington, Seattle, WA
6/04-6/11	**Research Assistant Professor**, Department of Microbiology, University of Washington, Seattle, WA
1/02-5/04	**Research Scientist/Senior Fellow**, Department of Microbiology, University of Washington, Seattle, WA *Mentor:* Associate Professor Roger Bumgarner

PROFESSIONAL SKILLS & EXPERIENCE

Scientific Research

- Diverse research expertise spanning multiple disciplines, including computer science, statistics, computational biology, cancer biology and systems biology.
- Extensive experience in the development and assessment of machine learning, data mining and pattern recognition methods, including unsupervised learning (e.g. clustering), supervised learning (e.g. classification), variable selection (feature extraction), model selection, building of predictive models, dimension reduction and network inference.
- Collaboration with biomedical and computational scientists to identify research problems, formulate the scientific framework and engineer practical solutions.
- My current research projects target the design of systems biology approaches and software tools to infer networks and elucidate biological insights by integrating heterogeneous 'omics' data sources. The methods and skills I developed for big data analyses in biology are applicable to other domains.
- Analyses, mining and integration of multiple types of high-dimensional data.
- Experience mentoring graduate students, postdoctoral fellows and software developers in research projects.
- 5 years experience as the Principal Investigator (PI) on a NIH R01 basic research project grant.
- 15 years experience with the design and implementation of algorithms.

- Open-minded approach to problem solving.

Scientific Writing and Presentation Experience

- Proven track record of scientific manuscript publications (24 published peer-reviewed scientific articles).
- Experience with grant writing and progress reports.
- Experience with scientific presentations at international meetings and conferences.
- Excellent reviews from guest lectures, invited lectures and workshops.
- Excellent communication and presentation skills.

Technical Skills

- Proficient in R, C, C++, Java, Perl, Splus, Matlab and Maple.
- Computer science techniques, including data structures, algorithm design, graph theory, optimization, combinatorics, feature selection, computational complexity, dynamic programming.
- Statistics techniques, including probabilities, Bayesian statistics, regression, model selection, variable selection, multivariate analyses, dimension reduction, clustering, classification, simulation, parameter estimation.
- Big data computing in the cloud.
- Development of publicly available open-source software.
- Scientific writing using Latex, Bibtex, Microsoft Word and EndNote.
- Experience using both commercial and academic data analysis tools, including Bioconductor packages, R, cytoscape, Splus, Matlab, Maple, Rosetta Resolver and SpotFire.
- Conversant with various operating systems, including Linux, Unix, Mac OS X, Windows; and standard software applications.

Other Professional Experience

- Experience with managing research projects and budget.
- Reviewer for scientific journals.
- Grant reviewer for the University of Washington Royalty Research Fund and NIH adhoc reviewer.
- Invited lectures and seminars.
- Poster presentations at scientific meetings.
- Served on the Admissions Committee for the Computational Molecular Biology program (2012-2013) at University of Washington and the University of Washington Faculty Senate (2012-2014).

PUBLICATIONS

Original Journal Publications

1. Raftery A.E., Niu X., Hoff P. and **Yeung K.Y.** Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics* 2012, 21(4): 901-919.
2. Lo K., Raftery A.E., Dombek K.M., Zhu J., Schadt E.E., Bumgarner R.E. and **Yeung K.Y**. Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. *BMC Systems Biology* 2012, 6:101.
3. **Yeung K.Y.**, Gooley T.A., Zhang A., Raftery A.E., Radich J.P., and Oehler V.G. Predicting relapse prior to transplantation in chronic myeloid leukemia by integrating expert knowledge and expression data. *Bioinformatics* 2012, 28(6): 823-830.
4. **Yeung K.Y.**, Dombek K.M., Lo K., Mittler J.E., Zhu J., Schadt E.E., Bumgarner R.E. and Raftery A.E. Construction of regulatory networks using time series microarray data in genotyped yeast segregants. PNAS 2011, 108(48): 19436 - 41.

5. Zarbl H., Gallo M.A., Glick J., **Yeung K.Y.** and Vouros P. The vanishing zero revisited: Thresholds in the age of genomics. *Chemico-Biological Interactions 2010*, 184(1-2): 273-8.
6. Oehler V.G.*, **Yeung K.Y.***, Choi E., Bumgarner R.E., Raftery A.E. and Radich J.P. "The derivation of diagnostic markers of chronic myeloid leukemia progression from microarray data". *Blood 2009*, 114: 3292-3298. *Co-first authors.
7. Annest A., Bumgarner R.E., Raftery A.E., **Yeung K.Y**. Iterative Bayesian Model Averaging: a method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinformatics* 2009, 10: 72.
8. Chu V.T., Gottardo R., Raftery A.E., Bumgarner R.E. and **Yeung K.Y**. MeV+R: Using MEV as a GUI for Bioconductor applications. *Genome Biology* 2008, 9:R118.
9. Liu X., Sivaganesan S., **Yeung K.Y.**, Guo J., Bumgarner R.E. and Medvedovic M. Bayesian Context-specific infinite mixture model for clustering of gene expression profiles accross diverse microarray datasets. *Bioinformatics* 2006, 22: 1737-1744
10. Gottardo R., Raftery A.E., **Yeung K.Y.** and Bumgarner R.E. Bayesian robust inference for differential gene expression in cDNA microarrays with multiple samples. *Biometrics* 2006*,* 62: 10-18.
11. Gottardo R., Raftery A.E., **Yeung K.Y.** and Bumgarner R.E. Robust estimation of cDNA microarray intensities with replicates. *Journal of the American Statistical Association* 2006 101: 30-40.
12. Li Q., Fraley C., Bumgarner R.E., **Yeung K.Y.** and Raftery A.E. Donuts, scratches and blanks: Robust model-based segmentation of microarray images. *Bioinformatics* 2005 21: 2875 - 2882.
13. **Yeung K.Y.**, Bumgarner R.E. and Raftery A.E. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* 2005 21: 2394-2402.
14. Vanasse G.J., Winn R.K., Rodov S., Zieske A.W., Li J.T., Tupper J.C., Peters M.A., **Yeung K.Y.**, and Harlan J.M. Bcl-2 overexpression leads to increases in suppressor of cytokine signaling-3 expression in B cells and de novo follicular lymphoma. *Molecular Cancer Research* 2004, 2: 620-631.
15. **Yeung K.Y.**, Medvedovic M., Bumgarner R.E. From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biology* 2004, 5:R48.
16. Medvedovic M., **Yeung K.Y.**, Bumgarner R.E. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics* 2004, 20:1222-1232.
17. **Yeung K.Y.**, Bumgarner R.E. Multi-class classification of microarray data with repeated measurements: application to cancer. *Genome Biology* 2003, 4:R83.
18. **Yeung K.Y.**, Medvedovic M., Bumgarner R.E. Clustering Gene-Expression Data with Repeated Measurements. *Genome Biology* 2003, 4:R34.
19. Barrett M.T., **Yeung K.Y.**, Ruzzo W.L., Hsu L., Blount P.L., Sullivan R., Zarbl H., Delrow J., Rabinovitch P.S., Reid B.J. Transcriptional Analyses of Barrett's Metaplasia and Normal Upper GI Mucosae. *Neoplasia* 2002, 4(2):121-128.
20. **Yeung K.Y.**, Fraley C., Murua A., Raftery A.E., Ruzzo W.L. Model-based Clustering and Data Transformations for Gene Expression Data. *Bioinformatics* 2001, 17:977-987.
21. **Yeung K.Y.**, Ruzzo W.L. Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics* 2001, 17:763-774.
22. **Yeung K.Y.**, Haynor D.R., Ruzzo W.L. Validating Clustering for Gene Expression Data. *Bioinformatics* 2001, 17:309-318.

Book Chapters and Review Articles

23. **Yeung K.Y**. Discovery of expression signatures in chronic myeloid leukemia by Bayesian Model Averaging. *Statistical Diagnostics of Cancer: Genetics and Genomics* 2013, Chapter 3. Wiley-Blackwell Publisher. Edited by Frank Emmert-Streib and Matthias Dehmer.
24. **Yeung K.Y**. Bayesian model averaging for biomarker discovery from genome-wide microarray data. *A Practical Guide to Bioinformatics Analysis* 2010, Chapter 2. Concept Press Ltd. Edited by Gabriel P. C. Fung.
25. Bumgarner R.E. and **Yeung K.Y**. "Methods for the inference of biological pathways and neworks" in *Computational Systems Biology*. Methods in Molecular Biology. 2009; 541:225-45. Edited by Jason McDermott, Ram Samudrala, Roger Bumgarner, Kristina Montgomery and Renee Ireton.
26. **Yeung K.Y.** and Bumgarner R.E. Pattern recognition in expression data. *Recent Developments in Nucleic Acids Research* 2004, 1: 333-354.
27. **Yeung K.Y**. "Clustering or automatic class discovery: non-hierarchical, non-SOM" in *A practical approach to microarray data analysis*. Kluwer Academic Publisher 2003, Chapter 16, pages 274-288. Edited by Daniel Berrar, Werner Dubitzky and Martin Granzow.

Conference Proceeding (peer reviewed)

28. **Yeung K.Y.** Signature Discovery for Personalized Medicine. Proceedings of the 2013 IEEE International Conference on Intelligence and Security Informatics, Part III, workshop papers, pp. 333-338.
29. Karp R.M., Stoughton R., **Yeung K.Y.** Algorithms for Choosing Informative Differential Gene Expression Experiments. *Proceedings of the 3rd annual international conference on computational biology (RECOMB)* 1999, pp. 208-217.

Meeting Abstracts

1. **Yeung K.Y.,** Blau C.A., Oehler V.G., Lee S.I., Miller C., Chien S., Martins T.J., Estey E. and Becker P.S. Personalized Approach to Acute Myeloid Leukemia Using a High-throughput Chemosensitivity Assay. To appear in *Blood (ASH Annual Meeting Abstracts*) 2013.
2. **Yeung K.Y.**, Dombek K.M., Lo K., Mittler J.E., Zhu J., Schadt E.E., Bumgarner R.E. and Raftery A.E. Construction of regulatory networks using time series microarray data in genotyped yeast segregants. *20th Annual International Conference on Intelligent Systems and Molecular Biology (ISMB Highlight Track*), July 2012.
3. Oehler V.G., **Yeung K.Y.**, Zhang A., Gooley T.A., and Radich J. P. Differential Gene Expression Associated with Chronic Myeloid Leukemia (CML) Progression Predicts Relapse and Survival Prior to Allogeneic Transplantation In Chronic Phase CML Patients. *Blood (ASH Annual Meeting Abstracts)*, Nov 2010; 116: 3507.
4. **Yeung K.Y.**, Oehler V.G., Choi E., Bumgarner R.E., Raftery A.E. and Radich J.P. Derivation of Diagnostic Gene Predictors for the Progression of Chronic Myeloid Leukemia from Microarray Data and Independent PCR Validation. *Blood (ASH Annual Meeting Abstracts*), Nov 2008; 112: 3211.

Software Packages

- *Mev+R* is an integration of the java MeV program with Bioconductor packages. This is now part of the official MeV release (http://www.tm4.org/mev.html).
- i*terativeBMA* is a bioconductor package for variable selection using high-dimensional gene expression data (http://www.bioconductor.org/packages/release/bioc/html/iterativeBMA.html).
- i*terativeBMAsurv* is a bioconductor package for survival analyses using high-dimensional gene expression data (http://www.bioconductor.org/packages/release/bioc/html/iterativeBMAsurv.html).
- *networkBMA* is a Bioconductor package for network inference and assessment (http://www.bioconductor.org/packages/release/bioc/html/networkBMA.html).

Data submission

- Time series gene expression data consisting of genotyped yeast segregants under rapamycin perturbation are publicly available from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) with accession number E-MTAB-412. We profiled the time-dependent expression levels over 6 time points of a set of 95 genomically characterized haploid yeast segregants in response to rapamycin (total 582 microarrays).
- Gene expression data of yeast ARO80, DAT1 and RTG3 deletion mutants after rapamycin perturbation are publicly available from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) with accession number E-MTAB-446. We profiled the expression levels of three single deletion mutants (ARO80, DAT1, RTG3) of the wild type strain BY4742 50 minutes after rapamycin perturbation (total 24 microarrays).