Michael Paul Stewart Brown

1175 Gaviota Dr, Number 4
Laguna Beach, CA 92651
michaelbrown14142@gmail.com
415-867-5620
http://www.openboundlabs.com/

Executive Summary: Skills and Experience

- 23 years of industry experience in scientific leadership roles

- Healthcare Diagnostics: Discovery, characterization, and development of diagnostics. Applications in HIV transmission, viral pathogen genomic seqeuencing, heart transplant rejection, lung transplant rejection, Lupus flares and drug effects. Clinical trials. Experimental design and statistical analysis. FDA clearances. CLIA laboratories.

- Statistical Pattern Recognition, Machine Learning, Mathematical Modeling: theory and practice

- Computer Science: formal languages, algorithms, high performance computing, compute pipelines

- Bioinformatics: diverse application areas

- Next Generation Sequencing, Gene Expression, Mass Spectrometry

Summary

My goal is to participate in high impact projects concentrating on important real-world problems whose success relies upon the use of scientifically-rigorous statistical analysis, pattern recognition, and mathematical modeling techniques implemented using efficient computer algorithms and flexible data handling techniques.

I have been involved in a variety data analysis and pattern discovery projects, mostly concentrated on biological data: sequencing DNA using new technology to identify new biological discoveries, using gene expression to measure the immune system and make clinically-relevant predictions about disease states, using proteomic mass spectrometry data to predict complex disease states like cancer, predicting new gene products in *Entamoeba histolytica*, using natural language in biomedical literature abstracts to predict the occurrence of names of organisms, using genomic sequence data to predict the best DNA markers to identify biowarfare pathogens, using video stream data to identify the presence of objects like people and vehicles, using customer data to rank customers to receive mail advertisements, using credit card data to predict fraudulent charges, using gene expression data to predict the functional classes of genes, using RNA sequences to predict the secondary structure of the molecule, ...

I have experience in many areas that help to build successful projects. I have done theoretical development of models and algorithms including original work with pattern recognition performance estimates, hidden Markov models (HMMs), stochastic context-free grammars, and support vector machines to the application of biological data such as DNA sequences, microarray data, and proteomic mass spectrometry data. I have implemented theoretical ideas using efficient algorithms on various computer platforms including GNU/Linux compute clusters. I have performed large-scale experiments including the design, statistical analysis, and interpretation of them.

I enjoy pattern recognition and data analysis and am very fortunate to live in an era in which information is electronically available, the volume of information is growing exponentially (especially for biological data), and cheap computing power allows the automated analysis of it. This ultimately yields knowledge that enriches our lives.

Keywords: *statistical pattern recognition, Bayesian statistical analysis, statistics, information theory, support vector machines (SVMs), information retrieval, natural language processing, bioinformatics, genomics, gene expression analysis, proteomics, Hidden Markov Models (HMMs), stochastic context-free grammars, algorithm development, programming, Linux compute clusters, Python, C++, C, Java, Perl, R, matlab*

Education _____

1999   Ph.D Computer Science.
University of California, Santa Cruz (UCSC).
Title: "RNA Modeling Using Stochastic Context-Free Grammars".
Advisor: David Haussler.

At UCSC, I was able to be a part of a spectrum of new work: hidden Markov models (HMMs) applied to sequence comparison, Dirichlet mixtures applied to protein priors, stochastic context-free grammars (SCFGs) applied to RNA secondary structure, support vector machines (SVMs) applied to DNA microarray gene expression data, SVMs on SCFG Fisher score vectors applied to organism identification, and others. All this work was done under a rigorous statistical pattern recognition and machine learning framework.

1992   B.S. Computer Science.
North Carolina State University (NCSU): Raleigh, NC.
Summa Cum Laude.

Work Experience _____

• 2010-2023 _____

Pacific Biosciences, Bioinformatics Sr. Staff Scientist

Pacific Biosciences sequences DNA using Single Molecule Real Time (SMRT(tm)) sequencing. This is a "third-generation" DNA sequencer that uses a light reporting system on a polymerase to read off the identity of a *single* molecule of DNA as it is being replicated in real time. My primary goal was to increase DNA seqeuencing accuracy using machine learning to examine low-level signals from the seqeuencer. I

also researched many application areas including sequencing entire Polio and HIV genomes from single molecules, estimating complex viral subspecies (quasi-species) in samples, identifying recombination events between different strains of Polio, confidently identifying low frequency variants, identifying new splicing patterns in HIV, BCR-ABL fusion mutant drug-resistance analysis, identifying viral species in samples from an outbreak in the Congo, examining ancient mitochondrial sequences, and metagenomic analysis on environmental samples. I did internal research by doing data analysis for the enzyme screening group which performed large-scale screening of different polymerase variations in order to find high performance characteristics. I developed barcoding strategies that included information-theoretic barcode design, ligation strategies, and sensitive identification techniques using HMMs that was validated on an EGFR-MET cancer barcoded panel using a Fluidigm-based ligation protocol. I did theoretical and practical work in the development of Circular Consensus Sequencing in which we try to estimate the most likely sequence from a set of observed sequencing reads. I did software development work on the PacBio high-performance analysis pipeline which performs parallel cluster computations to keep up with the vast amount of sequencing data.

• 2007-2010 ——

XDx Expression Diagnostics, Bioinformatics Staff Scientist

XDx (www.xdx.com) is a company that monitors a patient's immune system status in a clinical setting by examining gene expression. They have a FDA cleared blood test for predicting heart transplant rejection events. My work at XDx has involved pattern recognition in clinically relevant problems including heart transplantation, lung transplantation, and Systemic Lupus Erythematosus. My pattern recognition work has involved very challenging goals of finding robust classifiers with small samples sizes, noisy data (including incorrect labeling of data and outliers), small signal to noise ratios, clinical confounder variables, and a large number of irrelevant features with respect to the number of informative features. I have developed a system that uses simple feature selection and classifier algorithms in a framework that guards against overfitting using various levels of computational cross-validation / bootstrapping as well as theoretical bounds that emphasize possibly small but unlikely events and provide a guide to interpret the empirical computational results. I work with immunologists and molecular biologists to reinforce the data-driven results. I have also developed exploratory data analysis techniques that rely on co-clustering and visualization to help interpret data with unknown (or incorrectly labeled) structure. In addition to this biological pattern discovery, I have done a wide variety of work in clinical study design (power calculations, endpoint significance), clinical study management (sample selection, confounder balancing, accrual estimation and tracking), and even laboratory procedures (statistical acceptance criteria). At XDx, I have been fortunate to have experienced almost every aspect of discovering and developing a biologically meaningful biomarker that has real clinical utility.

• 2006 ——

Stanford University Upi Singh Lab, Independent Collaborator

My work at the Singh lab has concentrated on the genomic analysis of the medically important parasite *Entamoeba histolytica*. This protozoan parasite causes colonic and liver disease that results in 100,000 deaths per year making it the second most common cause of parasitic death in humans. My genomic analysis has concentrated on identifying and characterizing spliced gene transcripts using large-scale alignments of EST libraries against the recently sequenced genome. I developed code to compute these alignments, store the large number of results in a database, identify possible spliced introns, correlate our findings with existing annotations, and explore new hypotheses regarding spliced gene products, alternatively spliced genes, and transposable elements. I also used hidden Markov models and stochastic

context-free grammars to identify spliceosomal complex components in the full genome. I also maintain a webpage that details all the methods used, results and summaries of experiments, and allows interaction with the results database in a way that is easy to use for the biologists in the lab. This webpage has been crucial in making the collaboration with the biologists at Stanford possible as it is the easiest way to make available the large quantity of results that have been generated. The biological expertise of the lab coupled with my bioinformatic analysis has been fruitful.

• 2002-2005 _____

Predicant Biosciences, Sr. Scientist

My work at Predicant Biosciences involved the diagnosis of cancer disease states using mass spectrometry data collected from blood samples. We developed and built our own TOF mass spectrometer, microfluidics CE sample prep chip, and biomarker discovery platform and mainly targeted a prostate cancer application. My primary responsibility was to discover a pattern in protein abundance data that would differentiate healthy from cancer patients. Pattern discovery involved theoretical work as well as algorithm development. I worked daily with highly talented lab scientists to solve critical problems.

The important challenge here is to find significant patterns that will hold up in an independent validation test set. This is challenging because the number of patients in the training set is relatively small (order 100) while the dimension of the raw data collected from each patient is large (order $10^6, 10^8$). The probability of overfitting is very significant. I used rigorous and well-founded techniques from statistical pattern recognition and machine learning. These techniques afford superior performance and an understanding of the data that any ad-hoc analysis technique cannot match. This is especially important in areas of medicine where an incorrect or misapplied analysis can have catastrophic consequences such as misdiagnosing a disease state like cancer.

I developed several pattern recognition techniques and implementations, feature selection wrappers, statistical tests, Quality Control tools, visualization software, and an automated pattern recognition system. Pattern recognition runs involving thousands of parameter sweeps were performed. Given a run specification, data servers were started, thousands of pattern recognition runs were made, the results were collected, and statistics were generated. This was done automatically and any result could be readily brought up for inspection.

In addition to the pattern discovery, I was intimately involved with the shear data processing burden of applying very complex algorithms to datasets whose total size could reach 60 gigabytes of data per day. This involved the use of a Linux compute cluster that ran cluster job control software that I developed. The analysis pipeline was complex that had about 6 steps with dependencies and would take about 4 hours of computation for each experiment. All this computation happened automatically as soon as the raw data was available and results were stored and available to everyone. QC decisions and system performance was gauged every day based on these computations.

Predicant was initially named Biospect.

• 1999-2002 _____

Fair Isaac Corporation (FIC), Sr. Staff Scientist

I did most of my work in the Advanced Technologies Division of HNC. HNC was acquired by FIC in 2002.

I worked on several projects including: structural RNA identification working directly with lab scientists at ISIS Pharmaceuticals (funded through DARPA), a probabilistic model for information extraction from video streams (funded through through ARDA), the bugID system for identifying optimal DNA probes for bacterial pathogens using a spectrum of data sources including biomedical literature abstracts and genomic sequences (funded through DARPA *principal investigator*), and identifying optimal DNA probes for viral pathogens (funded through DARPA and USAF *principal investigator*)

I was also involved with several projects involving natural language processing, information retrieval, fraud detection, and advertising optimization.

• 1996 ————————————————————————————————

Affymetrix, Researcher.

At Affymetrix, I worked on sequencing by hybridization using 10-mer DNA gene chips. I worked directly with lab scientists whose expertise in chip chemistries and characteristics helped solve many problems. The Affymetrix chip had every possible 10-mer on it and the problem was to do DNA *de novo* sequencing by using information about which probes were present in the sample.

Grants ————————————————————————————————————

2003 **This grant was well reviewed but SBIR funding rules prevented us from accepting it.** National Cancer Institute R44 CA105545-01 "Advanced Biological Pattern Discovery System". Principal investigator.

We propose a discovery tool that accurately detects, diagnoses, and predicts outcomes of cancer using molecular signatures that characterize tumors and their interaction with their microenvironments and the body. One of the key components is a data analysis system that is robust and can reliably discover significant patterns from the large datasets produced by the separations-mass spectrometry components of our platform. Such a system will enable the discovery and assay of proteomic patterns that detect and distinguish different forms of cancer, including those that have similar clinical presentations, but differ at the molecular level. Predicant's integrated platform and advanced informatics system will become vital for the informed clinical management of cancer.

This grant received an excellent priority score of 153 (scores range from 100 to 500 with 100 being the best). The grant was scientifically sound but funding rules kept it from being funded.

2001 DARPA N66001-01-C-8010. "The BugID System for Discovering Optimal Nucleotide Probes". Principal investigator.

The bugID system is a probabilistic method for discovering optimal DNA probes for pathogens. The system uses a wide variety of relevant information: genomic sequences from the pathogen, its close neighbors, and all other known organisms, the physical properties of the target molecule, phylogenetic relationships between organisms, biological truths mined from English text in biomedical literature abstracts, and physical and chemical properties of DNA hybridization. It combines all this information in a probabilistic framework to find optimal probes that have very high probability of hitting the target and very low probability of hitting any other organism sequence. Probes designed using this system have much better performance with fewer false positives and false negatives than a more simple system that does not account for all the relevant information. The software computes and keeps track of the large amount of information generated from the variety of sources, a complicated dataflow management problem.

1991 NSF Research Experience for Undergraduates. "Natural Language Processing in Scripted Domains". Dr. R. Rodman.

Publications ————————————————————————————————

1. Lal, Avantika, Michael Brown, Rahul Mohan, Joyjit Daw, James Drake, and Johnny Israeli. (2021) "Improving long-read consensus sequencing accuracy with deep learning." bioRxiv : 2021-06.

2. Le Noun, Cyril, Thomas McCarty, Lijuan Yang, Michael Brown, Eckard Wimmer, Peter L. Collins, and Ursula J. Buchholz. (2021) "Rescue of codon-pair deoptimized respiratory syncytial virus by the emergence of genomes with very large internal deletions that complemented replication." Proceedings of the National Academy of Sciences 118, no. 13 : e2020969118.

3. Dilernia, Dario A., Jung-Ting Chien, Daniela C. Monaco, Michael PS Brown, Zachary Ende, Martin J. Deymier, Ling Yue et al. (2015). "Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing." *Nucleic acids research* 43, no. 20 : e129-e129.

4. Ocwieja K, Sherrill-Mix S, Mukherjee R, Custers-Allen R, David P, Brown M, Wang S, Link D, Olson J, Travers K, Schadt E, Bushman, F. (2012) Dynamic regulation of HIV-1 mRNA populations analyzed by single molecule enrichment and long read sequencing. *Nucleic Acids Res.* 2012 Nov 1;40(20):10345-55. doi: 10.1093/nar/gks75

5. Smith, Catherine C., Michael Brown, Jason Chin, Corynn Kasap, Sara Salerno, Lauren E. Damon, Kevin Travers et al. (2011) "Single molecule real time (SMRT) sequencing sensitively detects polyclonal and compound BCR-ABL in patients who relapse on kinase inhibitor therapy." Blood 118, no. 21 : 3752.

6. de Valpine P, Bitter HM, Brown MPS, Heller J. (2009). A simulation-approximation approach to sample size planning for high–dimensional classification studies. *Biostatistics*, 10(1):. doi:10.1093/biostatistics/kxp001

7. Davis Carrie, Brown M.P.S., Singh U. (2007). Functional Characterization of Spliceosomal Introns and Identification of U2, U4, and U5 snRNAs in the Deep-Branching Eukaryote *Entamoeba histolytica*. *Eukaryotic Cell*, 6(6):940-948. doi:10.1128/EC.00059-07

8. Sassi AP, Andel F 3rd, Bitter HM, Brown MP, Chapman RG, Espiritu J, Greenquist AC, Guyon I, Horchi-Alegre M, Stults KL, Wainright A, Heller JC, Stults JT. (2005). An automated, sheathless capillary electrophoresis-mass spectrometry platform for discovery of biomarkers in human serum. *Electrophoresis*, 26(7-8):1500-12.

9. Ahmed, Zulfikar, Hans Bitter, Michael Brown, Jonathan Heller, David Donoho, Jim Quarato, Arjuna Balasingham, and David De Valpine. (2005) "Methods for accurate component intensity extraction from separations-mass spectrometry data." U.S. Patent Application 10/846,996, filed November 17, 2005.

10. Andel, Frank, Hans Bitter, Mikhail Belov, Michael Brown, Alfred Greenquist, Jonathan Heller, Alexander Sassi, John Stults, and Kathy Stults. (2005) "Biological patterns for diagnosis and treatment of cancer." U.S. Patent Application 10/845,773, filed November 3, 2005.

11. Guyon I., Bitter H.M., Ahmed Z., Brown M., Heller J., (2003). Multivariate Non-Linear Feature Selection with Kernel Multiplicative Updates and Gram-Schmidt Relief. In *Proc. BISC FLINT-CIBI 2003 Workshop, Berkeley.*

12. Brown, M. P. (2000). Small subunit ribosomal rna modeling using stochastic context-free grammars. In *ISMB 2000*, pages 57–66.

13. Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Furey, T. S., M.Ares, J., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data using support vector machines. *PNAS*, 97(1):262–267.

14. Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I., and Haussler, D. (1996). Dirichlet mixtures. *CABIOS*, 12(4):327–345.

15. Brown, M. P. S. and Wilson, C. (1995). Rna pseudoknot modeling using intersections of stochastic context free grammars with applications to database search. In Hunter, L. and Klein, T., editors, *Pacific Symposium on Biocomputing*, pages 109–125.

16. Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjölander, K., Underwood, R. C., and Haussler, D. (1994a). Stochastic context-free grammars for tRNA modeling. *NAR*, 22:5112–5120.

17. Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *JMB*, 235:1501–1531.

18. Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjölander, K., Underwood, R. C., and Haussler, D. (1994b). Recent methods for RNA modeling using stochastic context-free grammars. In *Proc. Asilomar Conf. on Combinatorial Pattern Matching*, New York, NY. Springer-Verlag.

19. Brown, M. P., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K., and Haussler, D. (1993). Using Dirichlet mixture priors to derive hidden Markov models for protein families. In Hunter, L., Searls, D., and Shavlik, J., editors, *First International Conference on Intelligent Systems for Molecular Biology*, pages 47–55, Menlo Park, CA. AAAI/MIT Press.