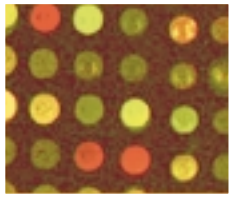
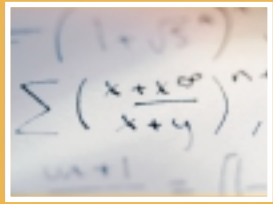
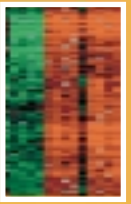
GENOMICS AND COMPUTATIONAL BIOLOGY

UNIVERSITY OF CALIFORNIA, BERKELEY

BERKELEY LAB

ERNEST ORLANDO LAWRENCE BERKELEY NATIONAL LABORATORY



GENOMICS AND COMPUTATIONAL BIOLOGY

UNIVERSITY OF CALIFORNIA, BERKELEY

ERNEST ORLANDO LAWRENCE BERKELEY NATIONAL LABORATORY

Welcome Message i

Applying to Berkeley 1

Courses 2

Faculty Research 6

DISCLAIMER: This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California. Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.

LBNL/PUB-861 TEID-4872

WELCOME MESSAGE

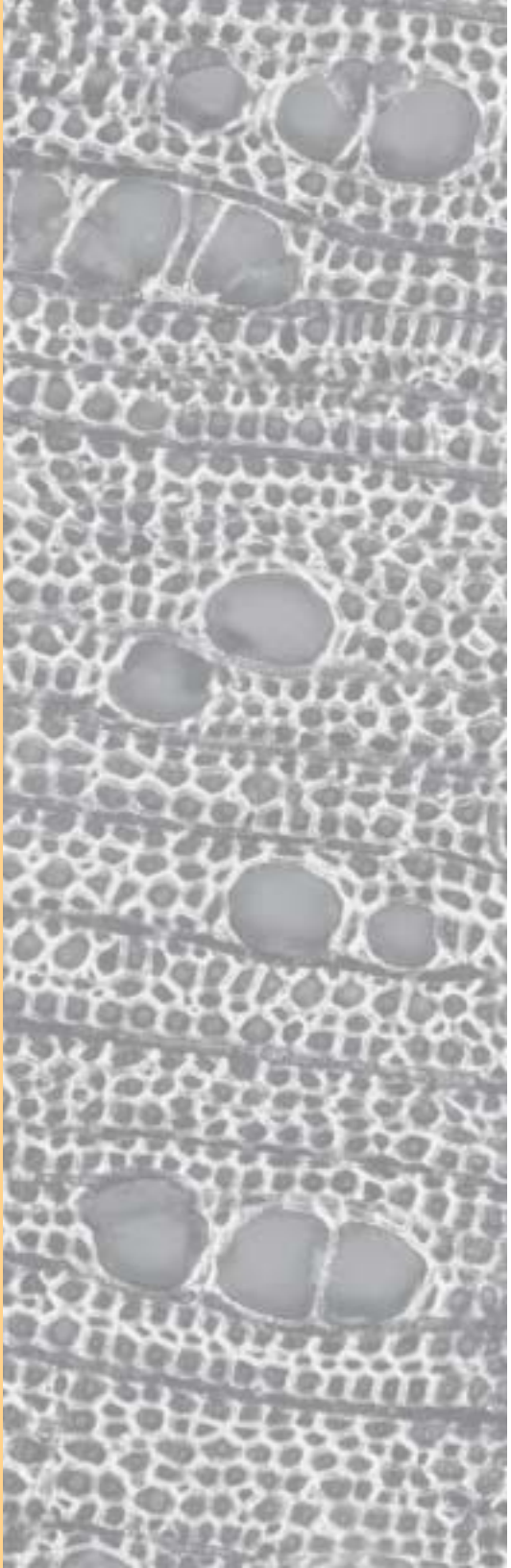
As you enter graduate school, biology is in the midst of an information revolution. Not only do we have to figure out how to make sense of all the raw data that's still pouring out of the sequencing centers, but with our appetites whet by the genome-scale perspectives that this data provides, we're hungry for more and more system-wide information, and impatient with one-gene-at-a-time approaches. What kinds of protein structures are encoded in a complete genome? How do these molecules interact and regulate one another? How are batteries of genes coordinately regulated? How do these networks generate complex tissues, organisms, and disease? How have organisms and their genomes been shaped by evolution? These challenging questions bring with them the feelings of excitement that come with the discovery of new concepts and approaches.

Sometime during your graduate career (perhaps many times!) you can realistically hope to have an "aha!" moment of insight into a basic principle or relationship underlying the organization, function, and evolution of living systems. A striking feature of these new genomic approaches is the way they seamlessly combine biological data and questions with concepts and tools more traditionally associated with computer science, engineering, physics, and chemistry, to not only process and interpret the data, but to generate it in clever ways as well. One of the challenges you'll face in your graduate career—whether you are coming from a biological or a physical/computational background—is finding a way to integrate information and methodologies from these fields with core training in biology.

The goal of our program at Berkeley is to help you combine diverse courses and "book learning" with research opportunities at the forefront of this information revolution in biology. The University of California at Berkeley is one of the world's premier research universities, with top-rated programs in the biological, physical, computational, and engineering sciences that provide an extensive array of opportunities for multidisciplinary training and research. New campus-wide initiatives in health sciences, structural biology, neuroscience, bioengineering, and computational biology promise to bring researchers from these diverse fields together in search of new synergies. Graduate students in our program learn and contribute to innovative multidisciplinary perspectives by their participation in courses, seminars, and research.

Berkeley is also fortunate to be home to the Lawrence Berkeley National Laboratory, a Department of Energy facility located adjacent to the campus. The Lab provides access to large scale facilities that no university can provide, and is home to world class researchers that are leading the way in genome sequencing, structural genomics, and developing novel approaches for functional genomics. (And don't forget that the San Francisco Bay area combines great natural beauty with a thriving urban center!)

As you read through this brochure, or browse our website, you'll get a glimpse of the richness of graduate and post-graduate training and research opportunities in genomics and computational biology that are growing here at Berkeley. Supported by a genomics training grant from the National Human Genome Research Institute, and an interdisciplinary training grant in physical bioscience from the National Science Foundation, our program in genomics and computational biology aims to train a select group of students from diverse academic disciplines. Our goal is simple: to develop tomorrow's leaders in an exciting, multidisciplinary environment. We welcome your application to our program, and look forward to learning more about you.



APPLYING TO BERKELEY

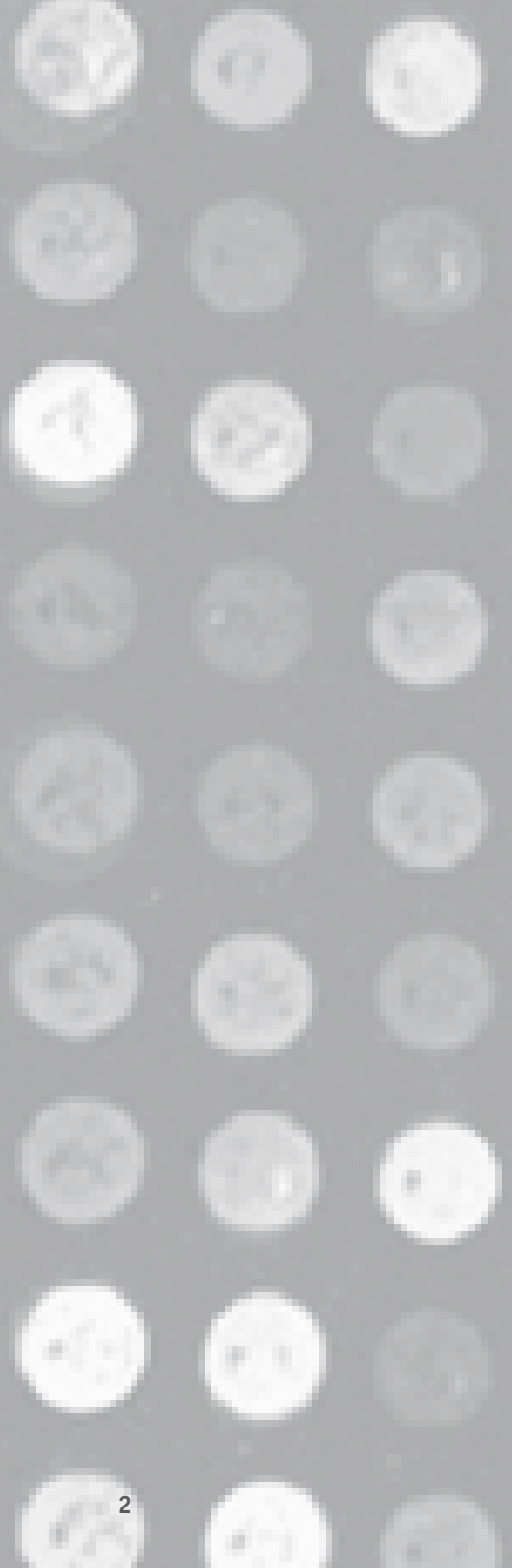
The Berkeley Training Program in Genomics provides graduate and postdoctoral training in all areas of genomics. Our goal is to provide multidisciplinary training through a wide range of course offerings, laboratory rotations, seminars, group meetings, and an annual retreat.

Graduate students in the Genomics Training Program may belong to any of the thirteen graduate programs at UC Berkeley listed below. Once admitted to a specific graduate program, students are responsible for fulfilling its core coursework and teaching requirements. By participating in the training program, students contribute to the Genomics and Computational Biology seminar series, enroll in additional courses outside their home program, and may form interdisciplinary group meetings. Students are also encouraged to discuss research with a secondary advisor in a complementary (computational or experimental) field.

As a prospective student, you should request and complete an application form from a home program. You should choose and apply to a single home program based on several factors. Define your primary emphasis: would your research predominantly use a computational approach or an experimental approach? You also want to choose a subject area whose core coursework meets your educational objectives. In addition, you should consider the departmental affiliations of faculty members whose work interests you.

Note that each program has its own application prerequisites, as well as its own policy on coursework and rotation projects conducted outside of the program. These policies can be investigated in the graduate bulletins and websites of the programs outlined below:

UC Berkeley Graduate Online Application	**http://www.grad.berkeley.edu/grad/admis/**
SPECIFIC PROGRAMS	
Bioengineering	http://www.coe.berkeley.edu/bioengineering/prospective_applicants.html
Biophysics	http://www.coe.berkeley.edu/biophysics/
Biostatistics (School of Public Health)	http://www.stat.berkeley.edu/biostat/
Chemistry	http://www.cchem.berkeley.edu/~chemgrad/grad_program/index.html
Computer Science	http://www.eecs.berkeley.edu/Prospective/grad.shtml
Integrative Biology	http://ib.berkeley.edu/grad/index.html
Mathematics	http://www.math.berkeley.edu/graduate/graduate.html
Molecular & Cell Biology	http://mcb.berkeley.edu/grad/
Neuroscience	http://neuroscience.berkeley.edu/PhDProgram.htm
Nutritional Sciences	http://nature.berkeley.edu/departments/nut/graduate.html
Physics	http://physics.berkeley.edu/studentlife/grad/application.shtml
Plant & Microbial Biology	http://mollie.berkeley.edu/graduate/GradDescr.html
Statistics	http://oz.berkeley.edu/~gyc/dept/



COURSES

Genomics and computational biology are inherently interdisciplinary endeavors that combine genome-wide biological experiments with the development of new computational methods to analyze these data. Researchers in these fields should become fluent with the core principles of statistical analysis and computational algorithms that can be applied to specific biological problems.

Several departments now offer courses that are integrated with case studies from genomics and computational biology research. In addition to fulfilling their home graduate program requirements, students may be interested in the following courses that have been offered in recent years:
(* denotes an advanced course)

STATISTICAL ANALYSIS

BIOENGINEERING/STATISTICS C141
Introduction to Statistics for Bioinformatics
Instructor: Bin Yu

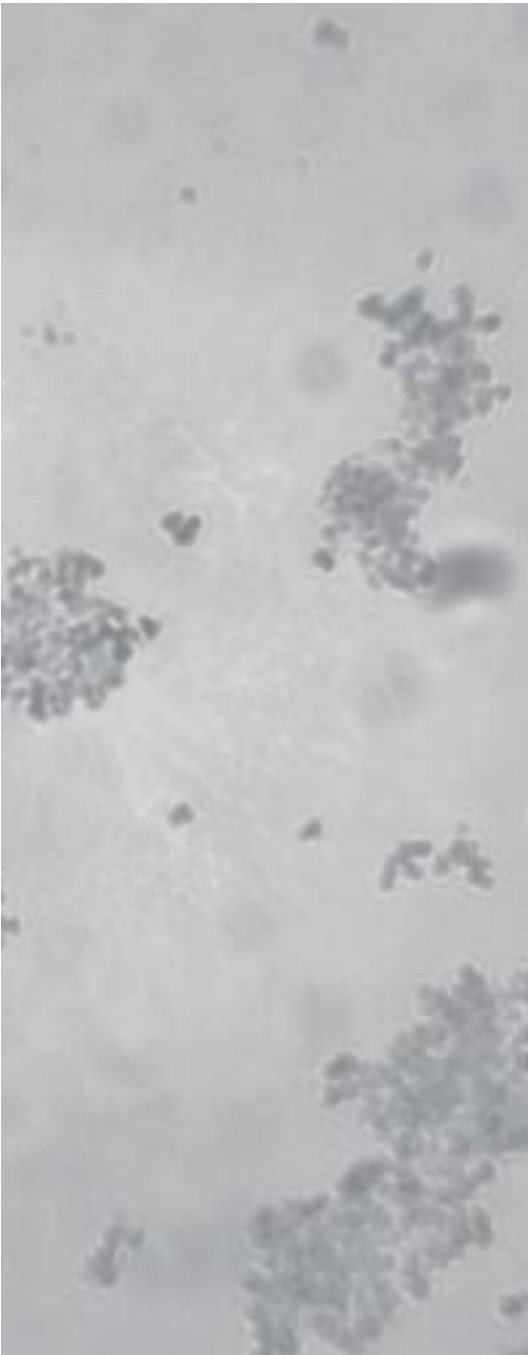
The purpose of this course is to give students an understanding of the principles of statistics and probability theory that underlie data analysis. A second objective is to introduce the statistical underpinnings of bioinformatics relevant to modern molecular biomedicine and bioengineering. It will deal primarily with the study of bioinformatics problems such as DNA pattern finding, gene expression data analysis, molecular evolution models, and biomolecular sequence database searching. The necessary probability and statistics foundations will be introduced. Topics in probability include: events; conditional and unconditional probability; random variables and their distributions; joint and conditional distributions; the central limit theorem. Introduction of necessary statistical inference techniques include: descriptive statistics; estimation, including maximum likelihood and least squares; correlation and regression; hypothesis testing, including likelihood ratios and permutation tests to test for group differences and for association.

PUBLIC HEALTH 243A*
Statistical Techniques in Computational Biology
Instructor: Mark van der Laan

This course will teach statistical techniques which have been useful in the analysis of data involving gene expression profiles and, possibly, the non-coding sequences of the genes. With the new biotechnologies, a typical longitudinal study following up a sample of subjects (or organisms, in general) can involve collecting of multi-thousand dimensional gene expression profiles (or other biomarkers) at one or more points in time, change of medical treatments over time, and missing/censored data on clinical outcomes such as survival. To deal with the large volumes of resulting data, subset selection and advanced cluster analysis with visualization is particularly helpful and the variability of such procedures needs to be estimated. In addition, selection of predictive genes is an important problem. Particular techniques include: model (and thus also covariate) selection, bootstrap, simultaneous confidence levels and sample size formulas, clustering and (causal) regression with censored data. Each technique is discussed in the context of a real data set.

STATISTICS 260*
Statistical Genetics
Instructor: Terry Speed

Statistical methods play an important role in a number of aspects of modern genetics, such as linkage and other types



of mapping, biomolecular sequence analysis, and the analysis of gene expression data. There is a need for statisticians with knowledge of the use of statistics in this field. Further, there will be students and researchers in the field desiring a better understanding of this area of application of statistics.

Material will be selected from the following topics: (1) Genetic linkage analysis: modeling meiosis, linkage mapping, pedigree analysis, and genetic epidemiology; (2) Genome sequence mapping: clone libraries, physical mapping of chromosomes, and radiation hybrid mapping; (3) DNA and protein sequence analysis, molecular evolution, sequence alignment, and database searching; (4) Analysis of microarray expression data: precision, reliability, discriminant and cluster analysis.

Other Courses of Interest:

PUBLIC HEALTH 296*
Applications of Statistics to Genetics and Molecular Biology: Seminar and Reading Group

STATISTICS 134
Concepts of Probability

STATISTICS 135
Concepts of Statistics

STATISTICS 150
Stochastic Processes

STATISTICS 153
Introduction to Time Series

STATISTICS 200*
Introduction to Probability and Statistics at an Advanced Level

STATISTICS 206*
Stochastic Processes

STATISTICS 230*
Linear Models

STATISTICS 238*
Bayesian Statistics

STATISTICS 242*
Analysis of Multidimensional Data

COMPUTATIONAL METHODS

BIOENGINEERING 142
Software Development for Bioinformatics (Course in development)

COMPUTER SCIENCE C281A/STATISTICS C241A*
Statistical Learning Theory
Instructor: Michael Jordan

This course will provide a thorough grounding in probabilistic and computational methods for the statistical modeling of complex, multivariate data. The emphasis will be on the unifying framework provided by graphical models, a formalism that merges aspects of graph theory and probability theory. Topics: classification, regression, clustering, dimensionality, reduction, and density estimation. Mixture models, hierarchical models, factorial models, hidden Markov and state space models, Markov properties, and recursive algorithms for general probabilistic inference, nonparametric methods including decision trees, kernel methods, neural networks, and wavelets. Ensemble methods.

COMPUTER SCIENCE 294*
Algorithms in Molecular Biology
Instructor: Richard Karp

Molecular life science seeks to understand life at the level of genes, proteins, and cells. Specific goals include sequencing and comparing the genomes of different organisms, identifying the genes and determining their functions, understanding the genetic basis of disease, understanding the evolutionary relationships among organisms, understanding how genes and proteins work in concert to control cellular processes, and predicting the three-dimensional structure of proteins. All of these endeavors require algorithms and databases for the analysis of complex data. The new discipline of computational molecular biology, or bioinformatics, has arisen in response to these challenges. This course will survey the fundamental algorithms of computational molecular biology.

Material will be selected from the following topics: (1) Genome sequencing and physical mapping: STS mapping, radiation-hybrid mapping, restriction mapping, linkage analysis, sequence assembly algorithms; (2) Sequence analysis: dynamic programming, pairwise and multiple alignment, motif discovery, hidden Markov models, expectation-maximization algorithm, gene finding; (3) Phylogeny: construction of evolutionary trees, parsimony, distance-based methods, Jukes-Cantor model, maximum-likelihood methods, genome rearrangements; (4) Gene expression analysis: DNA microarrays, clustering, partitioning, supervised learning, modeling of regulatory networks.

MATHEMATICS 195
Mathematical and Computational Methods in Molecular Biology
Instructor: Lior Pachter

The focus of the class will be the recent publication of the human genome. In particular, we will survey the relevant mathematical and computational techniques that have contributed to the sequencing and analysis of the genome. Topics include: a review of molecular biology; DNA structure and topology; genome sequencing and assembly; measures of sequence similarity; gene finding; comparative genomics; the human genome sequence.

Other Courses of Interest:

COMPUTER SCIENCE 61A
Structure and Interpretation of Computer Programs

COMPUTER SCIENCE 61B
Data Structures

COMPUTER SCIENCE 170*
Efficient Algorithms and Intractable Problems

COMPUTER SCIENCE 186*
Introduction to Database Systems

COMPUTER SCIENCE 281*
Machine Learning

BIOLOGICAL APPLICATIONS

MOLECULAR & CELL BIOLOGY 137
Computer Simulation in Biology
Instructors: George Oster & Robert Macey

This course explores the modeling and computer simulation of dynamic biological processes using special graphical interfaces requiring very little mathematical or computer experience. The first half focuses on realistic models from current literature to teach concepts and technique. The second half of the course is a workshop for student-selected individual projects.

PLANT & MICROBIAL BIOLOGY/MOLECULAR & CELL BIOLOGY/BIOENGINEERING C246
Topics in Genomics and Computational Biology
Instructors: Steven Brenner & Michael Eisen

Students successfully completing this course will have a general understanding of topics in computational biology and genomics, will be able to critically analyze and understand experiments and research articles in this field, and will be able to perform standard computational genomics analysis. Current topics include: genome sequencing; secondary structure prediction; protein folding problem; structural classification; phylogeny; DNA arrays and applications; correlation methods; genetic mapping.

Other Courses of Interest:

BIOENGINEERING 131
Computational Methodologies in Biomaterials Science

MOLECULAR & CELL BIOLOGY 200*
Advanced Biochemistry & Molecular Biology

MOLECULAR & CELL BIOLOGY 230*
Advanced Cell Biology

MOLECULAR & CELL BIOLOGY 240*
Advanced Genetic Analysis

PLANT & MICROBIAL BIOLOGY/MOLECULAR & CELL BIOLOGY 148
Microbial Genetics & Genomics

FACULTY

Berkeley faculty pursue an extremely broad range of topics at the cutting edge of genomics research. This diversity of experimental and computational expertise offers students the best possible opportunities for pursuing interdisciplinary training. Since collaborations are encouraged, it is possible to combine complementary aspects of research from the individual labs to design a truly interdisciplinary research project.

Research interests for a sampling of faculty in genomics and computational biology are described on the following pages. For additional faculty listings, consult the individual program Web sites listed on page 1. Recruitment efforts for additional faculty are also underway.

	RESEARCH INCLUDES COMPUTATIONAL ANALYSIS	RESEARCH INCLUDES EXPERIMENTAL ANALYSIS	
Adam Arkin	●	●	Quantitative cell biology and network informatics
Mark Biggin*		●	Animal transcriptional regulatory networks
Jeff Boore*	●		Comparative and organelle genomics
Steven Brenner	●		Computational structural & functional genomics
Sandrine Dudoit	●		Applications of statistics to problems in genetics and molecular biology
Michael Eisen	●	●	Experimental genomics and gene expression analysis
Tracy Handel	●	●	Protein structure, function, and computational design of proteins
Teresa Head-Gordon	●	●	Computational and experimental approaches to protein folding and structure prediction
Richard Karp	●		Algorithms for computational molecular biology
Richard Mathies		●	Microfabricated analysis systems for high-throughput genotyping and sequencing
John Ngai		●	Analysis of gene expression in the central nervous system using DNA microarrays
George Oster	●		Theoretical models of molecular biological processes
Lior Pachter	●		Gene finding, alignment, assembly, whole genome analysis
Jasper Rine		●	Genomic exploration of yeast
Dan Rokhsar	●	●	Decoding genomes and their regulatory networks
Eddy Rubin*	●	●	Biological annotation of mammalian genomes
Terry Speed	●		Microarray data analysis; Genetic mapping
Mark van der Laan	●		Computational biology and causality
Chris Vulpe		●	Comparative genomic approaches to copper and iron metabolism

*Researcher exclusively affiliated with Lawrence Berkeley National Lab. Students working with these investigators must find a University faculty sponsor from their home department.

FACULTY RESEARCH INTERESTS

FACULTY MEMBER	DEPARTMENT AFFILIATION*	GENOMIC DATASETS (▲ = ACTIVE EXPERIMENTAL WET LAB)								COMPUTATIONAL APPROACHES							BIOLOGICAL & MEDICAL INTERESTS							
		Biomolecular Sequence	Microarrays	Dynamic Biochemical & Biophysical Data	Crystallography & NMR	Mass Spectrometry	Imaging	Populations, Polymorphism, & Patients	BioMEMS & Microfabrication	Databases (Development & Data Mining)	Statistical Methods & Probabilistic Models	3D Molecular Models	Dynamic Chemical & Physical Models, Systems Biology	Combinatorics, Graph Theory, and Other Discrete Algorithms	Image Processing	Visualization Tools	Cell Biology	Biochemistry & Molecular Biology	Developmental Biology	Model Organisms & Genetics	Structural Biology	Neurobiology	Evolution & Comparative Genomics	Disease
Adam Arkin	Chem, BE, BP, LBNL		▲	▲			▲		▲	■	■		■	■	■	■	■	■	■	■				
Mark Biggin	LBNL		▲				▲											■	■	■			■	
Jeff Boore	IB, LBNL	■									■			■									■	
Steven Brenner	PMB, MCB, BE, BP, LBNL	■	■		■	■				■	■	■		■				■		■	■		■	
Sandrine Dudoit	BS	■	■				■	■		■	■			■	■	■				■				■
Michael Eisen	MCB, BP, LBNL	▲	▲					▲		■	■				■	■	■	■	■	■			■	■
Tracy Handel	MCB, BP, LBNL			▲	▲					■		■	■					■			■			■
Teresa Head-Gordon	BE, LBNL				▲							■	■					■			■			
Richard Karp	CS, BE, Math	■	■								■			■										
Richard Mathies	Chem, LBNL								▲														■	■
John Ngai	MCB, LBNL		▲														■	■	■			■		
George Oster	MCB			■	■							■	■				■	■			■			
Lior Pachter	Math, LBNL	■								■	■			■		■				■			■	
Jasper Rine	MCB	▲	▲															■		■				
Dan Rokhsar	Physics, LBNL	▲									■			■					■	■			■	
Eddy Rubin	LBNL	▲	▲													■							■	■
Terry Speed	Statistics, LBNL		■					■		■	■				■	■	■	■	■	■		■	■	■
Mark van der Laan	BS	■	■					■		■	■					■		■		■				■
Chris Vulpe	Nutri. Sci, LBNL		▲			▲		▲		■	■		■					■		■			■	

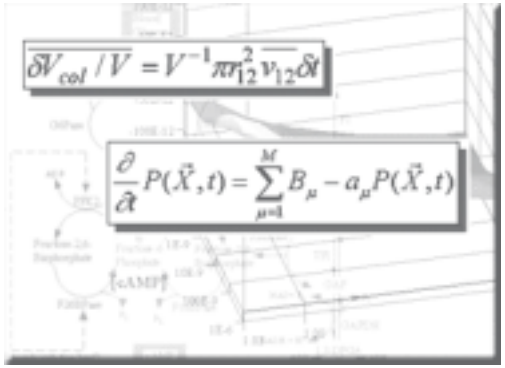
* BE = Bioengineering, BP = Biophysics, BS = Biostatistics, CS = Computer Science, IB = Integrative Biology, LBNL = Lawrence Berkeley National Laboratory (DOE), MCB = Molecular and Cell Biology, Nutri. Sci = Nutritional Science & Toxicology, PMB = Plant & Microbial Biology

ADAM ARKIN

CHEMISTRY/BIOENGINEERING/BIOPHYSICS
http://www.genomics.berkeley.edu
aparkin@lbl.gov

Quantitative cell biology and network informatics

We combine data fusion, statistics, dynamical modeling and other engineering tools to understand cellular regulatory networks. To understand, here, means the ability to predict, control and design. To this end we have been developing a computational framework for storing, relating and analyzing biological data in a "network" context. These tools include data analytical tools to predict biomolecular interactions from data and to score significant behaviors under perturbation, as well as model building and analysis tools for simulation of cellular networks. These networks are highly nonlinear, sometimes stochastic and are usually only partially known. So we must combine experimental, theoretical and computational approaches. We have been building a quantitative cell biology laboratory including high-end microscopes, microarrays and other molecular profiling techniques to collect enough data to make the computational analyses possible. Problems under study in our laboratory include: multiple signal detection and resolution in chemotaxing neutrophils, development of asymmetry and sporulation in *Bacillus subtilis*, comparative bacterial chemotaxis of *Escherichia coli* and *B. subtilis*, and a couple of viral infection processes. In addition there are a number of pure theory projects underway. The goal of studying such diverse sets of systems is both to develop general methodologies for cellular systems analysis and to discover common regulatory motifs in cellular signal processing.



Selected publications

Signal processing by biochemical reaction networks. [A. P. Arkin (1999), In: Biodynamics, J. Walleczek, ed., Cambridge University Press]

Stochastic kinetic analysis of a developmental pathway bifurcation in phage-λ *Escherichia coli*. [A. P. Arkin, J. Ross, and H. H. McAdams (1998) *Genetics* **149**, 1633-1648]

A test case of correlation metric construction of a reaction pathway from measurements. [A. P. Arkin, P.-D. Shen, and J. Ross (1997) *Science* **277**, 1275-1279]

MARK BIGGIN

LAWRENCE BERKELEY NATIONAL LAB
http://www-gsd.lbl.gov/biggin/index.html
mdbiggin@lbl.gov

Animal transcriptional regulatory networks

Animal development is controlled by highly complex transcriptional networks. The enormous complexity of these networks presents a challenge if we are to understand how regulatory factors control transcription, pattern formation, and morphogenesis. We are developing approaches to dissect networks on a genome wide scale, using the *Drosophila* embryo as a model.

Part of our research is conducted in a traditional investigator led laboratory, part is within an interdisciplinary collaboration involving groups at LBNL and UC Berkeley. The Biggin laboratory focuses on understanding how homeoproteins each generate distinct developmental fates despite the fact that they each have similar and promiscuous DNA binding specificities *in vitro*. Our data indicate that, contrary to a widely held belief, these proteins bind and regulate most genes in living embryos. Their different specificities appear to result from differences in the way each homeoprotein regulates common target genes. We are now using biochemical, transgenic, evolutionary, and quantitative genetic approaches to explore the relationship between homeoprotein specificity and morphogenesis.

The interdisciplinary collaboration is pioneering approaches to study complete transcription networks in animals. We are developing methods to measure key parameters governing the activities of transcription factors, including microarray based *in vitro* and *in vivo* DNA binding assays and advanced imaging methods to measure three dimensional patterns of gene expression. Our Bioinformatics group, led by Michael Eisen, will then integrate the resulting data, together with the genome sequence, to model all aspects of the transcriptional network. This long term project seeks to determine how the information that dictates transcriptional patterns in the embryo is encoded in the genome.

Selected publications

Accessibility of transcriptionally inactive genes is specifically reduced at homeoprotein-DNA binding sites. [A. Carr and M. D. Biggin (2000) *Nucleic Acids Res.* **28**, 2839-2846]

A comparison of in vivo and in vitro DNA binding specificities suggests a new model for homeoprotein DNA binding in Drosophila. [A. Carr and M. D. Biggin (1999) *EMBO J.* **18**, 1598-1608]

Eve and Ftz regulate a wide array of genes in blastoderm embryos: The selector homeoproteins directly or indirectly regulate most genes in Drosophila. [Z. Liang and M. D. Biggin (1998) *Development* **125**, 4471-4482]

JEFF BOORE

JOINT GENOME INSTITUTE
http://www.jgi.doe.gov/programs/comparative/ComparativeGenomics.html
jlboore@lbl.gov

Comparative and organelle genomics

The Joint Genome Institute determines the sequence of about 600 million nucleotides per month. Current genome targets include those of the pufferfish *Fugu*, the tunicate *Ciona*, the fungus *Phanerochaete*, and about 40 prokaryotes. Our group provides phylogenetic analysis as part of interpreting these genome sequences, to include: 1) assessing the constitution and organization of gene families; 2) evaluating the role of genomic factors (e.g., gene movements, duplications, and loss) in shaping organismal change; 3) correlating novel genomic features with novel traits (morphological, physiological, behavioral, etc.) to identify candidate genes; 4) identifying *cis* regulatory elements by their phylogenetic footprint; 5) partitioning the genome into subsets that show evidence of positive selection, gene conversion, or horizontal transfer; and 6) reconstructing phylogeny by using genome level markers and large scale sequence comparisons. We also lead projects in targeted gene family evolution (e.g., *Hox* cluster evolution) and in comparing organelle (both mitochondrial and chloroplast) genomes for phylogenetic inference, modeling genome evolution, and addressing questions of biogeography, conservation biology, and population structure. Finally we are also developing novel methods for rapidly isolating, cloning, and sequencing organelle genomes and are developing software for automating the annotation, manipulation, and presentation of this comparative data.

Selected publications

Sequence and structure of the mitochondrial genome of the tapeworm *Hymenolepis diminuta*: Gene arrangement indicates that platyhelminths are derived eutrochozoans. [M. von Nickisch-Rosenegk, W. M. Brown, and J. L. Boore (2001) *Mol. Biol. Evol.* **18**, 721-730]

Animal mitochondrial genomes. [J. L. Boore (1999) *Nucl. Acids Res.* **27**, 1767-1780]

Gene translocation links insects and crustaceans. [J. L. Boore, D. Lavrov, and W. M. Brown (1998) *Nature* **392**, 667-668]

STEVEN BRENNER

PLANT & MICROBIAL BIOLOGY/ MOLECULAR & CELL BIOLOGY/ BIOENGINEERING/BIOPHYSICS
http://compbio.berkeley.edu
brenner@compbio.berkeley.edu

Computational structural & functional genomics

By enumerating complete gene repertoires, genomes provide an unprecedented unbiased view of biology. Our group probes sequenced genomes with computational methods to discover previously overlooked molecular and cellular biological activities. In doing so, we aim both to deepen our understanding of recognized facets of molecular biology and to broaden the realm of known functions.

Structural genomics projects attempt to provide an experimental structure or a good theoretical model for every tractable protein in all completed genomes. Our work aims to use structural genomics to decipher the function of uncharacterized proteins in sequenced genomes. This research involves organizing proteins into families according to homology, and classifying proteins and RNA according to structure. In addition, as members of the Berkeley Structural Genomics Center (http://www.strgen.org) we select proteins for experimental characterization and analyze solved structures to detect homology and functional information.

To explore the function of proteins without structures, our functional genomics efforts involve creating algorithms using molecular sequence, structure, phylogeny, and expression information. This work includes the use of gene genealogies to trace gene histories and functional divergences; reverse-genomics comparison of multiple complete genomes to locate genes associated with characterized cellular or biochemical functions; and continued refinement of sequence comparison methods. We also combine sequence comparison with expression and other experimental data to improve molecular and cellular functional characterization.

Selected publications

Expectations from structural genomics. [S. E. Brenner and M. Levitt (2000) *Protein Sci.* **9**, 197-200]

Errors in genome annotation. [S. E. Brenner (1999) *Trends Genet.* **15**, 132-133]

Assessing sequence comparison methods with reliable structurally-identified distant evolutionary relationships. [S. E. Brenner, C. Chothia, and T. J. P. Hubbard (1998) *Proc. Natl. Acad. Sci. USA* **95**, 6073-6078]

SANDRINE DUDOIT

BIOSTATISTICS
http://www.stat.berkeley.edu/~sandrine
sandrine@stat.berkeley.edu

Applications of statistics to gene expression

Current research projects concern the development and implementation of statistical methods for the design and analysis of gene expression experiments using DNA microarrays. We are implementing these methods in a freely available library of R functions, SMA (Statistics for Microarray Analysis).

Image analysis. We have developed new addressing, segmentation, and background correction methods for extracting information from microarray scanned images.

Normalization. This step identifies and removes systematic sources of variation (other than differential expression) in the measured fluorescence intensities and is required before any analysis which involves comparing expression levels within or between slides. We have developed location and scale normalization methods which take into account intensity, spatial, and other effects using robust local regression methods.

Multiple testing. An important and common question in microarray experiments is the identification of differentially expressed genes. The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels and the responses or covariates of interest. Special problems arising from the large multiplicity problem include defining an appropriate false positive error rate and devising powerful multiple testing procedures which control this error rate and incorporate the joint distribution of the expression levels across genes.

Cluster analysis and discriminant analysis. DNA microarrays are being applied increasingly in cancer research with the aim of deriving a finer and more reliable classification of tumors. Statistical problems associated with tumor classification include: the identification of new tumor classes using gene expression profiles (cluster analysis) and the classification of malignancies into known classes (discriminant analysis). We have developed prediction-based resampling methods for estimating the number of tumor subclasses, increasing clustering accuracy, and assessing the confidence of cluster assignments for individual tumors.

Selected publications

Comparison of discrimination methods for the classification of tumors using gene expression data. [S. Dudoit, J. Fridlyand, and T. P. Speed (2001) *J. Am. Statist. Ass.* (in press). (Tech report #576, Department of Statistics, UC Berkeley)]

Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. [S. Dudoit, Y. H. Yang, T. P. Speed, and M. J. Callow, *submitted* (Tech report #578, Department of Statistics, UC Berkeley)]

Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. [J. Fridlyand and S. Dudoit, *submitted* (Tech report #600, Department of Statistics, UC Berkeley)]

MICHAEL EISEN

MOLECULAR & CELL BIOLOGY/BIOPHYSICS

http://rana.lbl.gov

mbeisen@lbl.gov

Experimental genomics & gene expression analysis

Our lab is interested in gene regulation. Specifically we want to understand how the information that governs complex temporal and conditional patterns of gene expression is encoded in genome sequences and is read out by the cell, and how the variation and evolution of gene regulation is related to cellular and organismal phenotypes. We address these questions by integrating experimental genomics methods with sophisticated computational analysis of experimental data, genome sequences and other relevant information. Current projects include:

Understanding cis-regulation in yeast: Labs across the world (including ours) have used DNA microarrays to monitor gene expression in the budding yeast *Saccharomyces cerevisiae*. We develop and apply computational techniques that relate expression data to the yeast genome sequence with the goal of understanding how the observed expression patterns are encoded in the genome.

Understanding cis-regulation in Drosophila: In collaboration with other labs at Berkeley and LBNL we have initiated a comprehensive project to characterize and understand cis-regulation in the developing *Drosophila* embryo. Our collaborators are characterizing the *in vitro* and *in vivo* DNA binding activities of all relevant transcription factors, determining sequences of regulatory regions from four additional *Drosophila* species, and developing a spatial and temporal atlas of gene expression in the developing embryo. Our lab is developing concepts and tools to transform this data into a systematic understanding of cis-regulation in Drosophila that we believe will be the basis of a similar understanding of gene regulation in humans.

Classification and characterization of human tumors: A dominant confounding factor in the treatment of human tumors is the remarkable clinical heterogeneity of the disease. Tumors of the same "type" (e.g. breast cancer) vary wildly in their prognosis and response to therapy. Since this clinical heterogeneity results primarily from underlying molecular heterogeneity, we use DNA microarrays to characterize tumors at the molecular level, and develop and apply analytical tools to use this data to define more accurate and clinically relevant molecular taxonomies of human tumors.

Selected publications

Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. [D. Y. Chiang, P. O. Brown, and M. B. Eisen (2001) *Bioinformatics* **17**, S49-S55]

Distinct types of diffuse large B-cell lymphoma iden-

tified by gene expression profiling. [A. A. Alizadeh, M. B. Eisen, *et al.* (2000) *Nature* **403**, 503-511]

Cluster analysis and display of genome-wide expression patterns. [M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein (1998) *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868]

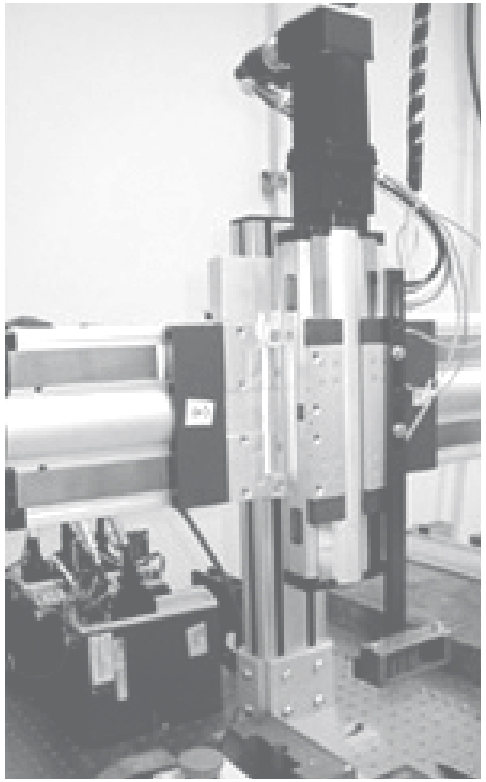
TRACY HANDEL

MOLECULAR & CELL BIOLOGY/BIOPHYSICS
http://mcb.berkeley.edu/faculty/BMB/handelt.html
handel@paradise1.berkeley.edu

Protein structure, function, and design

Our interests are in understanding the relationship between the primary sequence and 3-dimensional structure and function of proteins using two converging approaches.

The first approach involves investigating the structure and function of proteins using NMR or X-ray crystallography as well as biochemical and computational methods. We have focused primarily on chemokines and chemokine receptors, proteins that have a critical role in the immune response by virtue of their ability to control leukocyte migration and homing. These proteins are also of considerable medical interest due to their involvement in a large number of diseases and processes like transplant rejection. A second group of proteins fall under the category of viral immunomodulatory proteins, proteins that viruses have evolved to suppress the host immune response. As an example, one structural project is aimed at determining structures of complexes between soluble viral proteins and chemokines. The viral proteins have no homology to host chemokine receptors, but bind many different chemokines with high affinity, and thereby block their ability to control leukocyte migration. A second computational project



is aimed at gleaning structural and functional data about host chemokine receptors by pattern recognition and functional clustering; these proteins are seven transmembrane receptors and therefore challenging to study by conventional structural approaches.

The second strategy involves the development of computational algorithms to design completely novel proteins, and biophysical characterization of the designs to validate and optimize the algorithms. In the simplest case, we design new sequences that adopt a predefined fold. More recently we have begun to design new protein-protein interactions with the aim of swapping specificity and controlling molecular self-assembly. We have also begun to design proteins that bind novel ligands as a first step in the rational creation of enzymes that catalyze reactions not found in nature. While there is much fundamental knowledge to be gained from the successes and failures of design, our philosophy is that the sequence content of genomes is just the beginning, and that protein design will ultimately have practical applications in medicine and biotechnology. Examples include the generation of proteins with improved properties, new functions, or altered ligand specificity.

Selected publications

Solution structure and dynamics of the Melanoma Inhibitory Activity protein (MIA). [J. C. Lougheed, P. J. Domaille, and T. M. Handel. *Biochemistry* (submitted)]

Protein design: where we were, where we are, where we're going. [N. Pokala and T. M. Handel (2001) *J. Struct. Biol.* (in press online)]

De novo design of the hydrophobic cores of proteins. [J. R. Desjarlais and T. M. Handel (1995) *Protein Sci.* **4**, 2006-2018]

TERESA HEAD-GORDON

BIOENGINEERING
http://www.lbl.gov/~thg
tlhead-gordon@lbl.gov

Computational and experimental approaches to protein folding and structure prediction

While the experimental effort in structural genomics is partly focused on providing new fold classifications, computation and theory should play a complementary role of completing structural, kinetic, and thermodynamic information across whole genomes. In this case, reduction in computational complexity of the model will be necessary but retaining physical-

chemical connections to experiment will be vital. Our recent work seeks to design protein models with stronger quantitative connections to experiments. We have now completed multiple studies of a minimalist protein folding model addressing issues of protein sequence design, solvation and interaction complexity in protein folding models, the ability to design and validate the folding of complex topologies, longer protein chains, and most recently in regards to protein engineering (phi-value analysis) studies. To further the quantitative robustness of these models we also study aqueous hydration by experiment and simulation. Even though it is appreciated that water environment is a vital determinant of protein tertiary structure, the experimental tools available to characterize aqueous hydration is comparatively minimal at present. By combining solution scattering experiments and molecular dynamics simulations, we have determined the solvation potentials of mean force of amino acid association in water, as well as conducting high-quality experiments on neat, ambient water under various conditions. This work provides an important benchmark for simulation force fields and emerging simulation methodologies such as *ab initio* molecular dynamics, connections to the IBM Blue Gene effort, and extensions to molten globule intermediates. We also develop local optimization and global optimization algorithms to predict protein structure directly from sequence. The "implicit" hydration potentials between amino acids in solution derived from experiment/simulation have been used to define new energy functions for structure prediction. We have obtained blind prediction results with our method and energy function in the 4th Critical Assessment of Techniques for Protein Structure Prediction (CASP4) competition, and show that our approach is more effective on targets for which less information from known proteins is available, and produces the best prediction for one of the most difficult targets of the competition.

Selected publications

Computational challenges in structural and functional genomics. [T. Head-Gordon and J. Wooley (2001) *IBM Research Journal on Deep Computing in the Life Sciences*, B. Robson, J. Coffin, W. Swope, editors, in press]

A hierarchical approach for parallelization of large tree searches. [S. Crivelli and T. Head-Gordon (2001) *J. Parallel & Distributed Computing* (submitted)]

Matching simulation and experiment: a new simplified model for simulating protein folding. [J. M. Sorenson and T. Head-Gordon (2000) *J. Comput. Bio*. **7**, 469-481]

RICHARD KARP

COMPUTER SCIENCE/BIOENGINEERING/MATHEMATICS

http://www.cs.berkeley.edu/~karp

karp@icsi.berkeley.edu

Algorithms for computational molecular biology

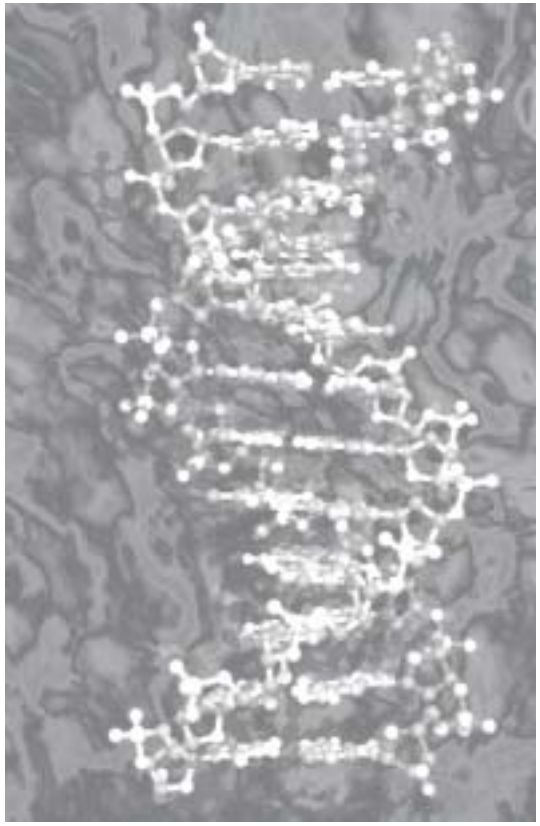
Richard Karp is a computer scientist working on the algorithmic aspects of computational molecular biology and genomics. He has developed algorithms for constructing various kinds of physical maps of DNA targets, including radiation hybrid maps, STS maps and restriction maps. He has also worked on the problem of probe selection for DNA microarrays. With his student Eric Xing he has developed methods for classifying biological samples on the basis of gene expression data. His current interests are in the application of combinatorial and probabilistic methods to finding hidden patterns in gene expression data and discovering the structure of genetic regulatory networks. He serves on the Scientific Advisory Board of Affymetrix Corp., an NIH Review Panel in genomics, and the Steering Committee of the annual RECOMB conference in computational biology.

Selected publications

CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. [E. P. Xing and R. M. Karp (2001) *Bioinformatics* **17**, S306-S315]

Universal DNA tag systems: a combinatorial design scheme. [A. Ben-Dor, R. Karp, B. Schwikowski, and Z. Yakhini (2000) *J. Comput. Biol.* **7**, 503-519]

An algorithm combining discrete and continuous methods for optical mapping. [R. M. Karp, I. Pe'er, and R. Shamir (2000) *J. Comput. Biol.* **7**, 745-760]



RICHARD MATHIES

CHEMISTRY
http://www.cchem.berkeley.edu/~ramgrp/LabWebPage/index.html
rich@zinc.cchem.berkeley.edu

Microfabricated analysis systems for high-throughput genotyping and sequencing

We work on the development of novel high-throughput genotyping and sequencing methods and technologies and their application. Microfabricated genetic analysis systems provide the ability to perform ultrafast electrophoretic analyses coupled with integrated nanoliter sample preparation. Capillary array electrophoretic microplates with from 96-384 lanes microfabricated in a single wafer are used in these studies. Current projects include the development of a fully integrated BAC-chip to sequence up to 100 kb fragments in one integrated operation. A second project involves the development of a new method for SNP identification called Polymorphism Ratio Sequencing and its application for the identification of mitochondrial variations associated with cancer. A third project will develop new SNP genotyping methods and apply them to deep SNP typing of matings in the Child Health and Development Studies cohort repository at the School of Public Health. These data will be used for the identification of disease-associated genetic markers.

High-throughput DNA sequencing in a 96-lane microfabricated electrophoresis device. [B. M. Paegel, C. A. Emrich, G. J. Wedemayer, J. R. Scherer, and R. A. Mathies, *submitted*]

High performance multiplex SNP analysis of 3 haemochromatosis related mutations with capillary array electrophoresis microplates. [I. Medintz, W. W. Wong, L. Berti, L. Shiow, J. Tom, J. R. Scherer, G. Sensabaugh, and R. A. Mathies (2001) *Genome Research* **11**, 413-421]

Radial capillary array electrophoresis microplate and scanner for high-performance nucleic acid analysis. [Y. Shi, P. Simpson, J. R. Scherer, D. Wexler, C. Skibola, M. T. Smith, and R. A. Mathies. (1999) *Analytical Chemistry* **71**, 5354-5361]

JOHN NGAI

MOLECULAR & CELL BIOLOGY
http://mcb.berkeley.edu/labs/ngai/pages/homepage.html
jngai@socrates.berkeley.edu

Analysis of gene expression in the central nervous system using DNA microarrays

Concomitant with genome sequencing efforts, investigators are attempting to determine which sets of genes, out of the entire repertoire in an organism, are expressed under what conditions, in which tissues, at what stage in development, and in response to what internal or external cues. This latter information—the "expressed genome" of an organism—has opened up entirely new ways of addressing and understanding basic biological processes. The methodology for assessing the expressed genes in an organism or cell is based on using glass slides onto which DNA corresponding to all of the genes in a organism are arrayed in an addressable pattern. These DNA microarrays—also referred to as "gene chips"—are then probed with mRNA isolated from the cells or tissues of interest. In this way, a global picture of which genes are expressed, and which are not, under any circumstance can be obtained. With currently available technology, it is now possible to monitor simultaneously the expression of all of the recognizable genes in yeast (6,307), fruit flies (~14,200), worms (~19,100) and even mice and humans (~35-50,000). The ability to perform such analyses en masse provides the ability to characterize biological and pathological processes at unprecedented levels of detail. At the same time, this field is still in its infancy with regard to optimizing and refining both the technology for producing and querying microarrays and the computer-based methods for analyzing and extracting biological meaning from such massive amounts of data.

We have established in our laboratory the full capabilities for carrying out DNA microarray analysis of gene expression. These techniques allow the analysis of mRNA expression from tens of thousands of genes at a time. To date, we have created high density cDNA microarrays from the mouse and the zebrafish. We are using these microarrays as tool to investigate patterns of developmentally-regulated and spatially-restricted patterns of gene expression in the vertebrate central nervous system, as well as more generally during development. We have also generated DNA microarrays containing all identified open reading frames in the worm and are using these microarrays as tools for monitoring gene expression in this model genetic organism. In collaboration with Terry Speed's group (Statistics), we are also working to develop improved methods for DNA microarray experimental design and statistical analysis.

Selected publications

Normalization for cDNA microarray data. [Y. H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed. *submitted*]

Three-dimensional patterns of gene expression in the olfactory bulb as revealed by DNA microarray analysis. [D. M. Lin, Y.H. Yang, J. Scolnick, L. Brunet, V. Peng, T.P. Speed, and J. Ngai. *In preparation*]

An improved method for mRNA amplification and expression profile analysis. [D. M. Lin, P. Luu, T. Serafini, and J. Ngai. *In preparation*]

GEORGE OSTER

MOLECULAR & CELL BIOLOGY
http://www.cnr.berkeley.edu/~goster/home.html
goster@nature.berkeley.edu

Theoretical models of molecular biological processes

My research involves construction and testing of theoretical models of molecular, cellular and developmental processes. Current projects include investigations into the basic physics and chemistry of protein motors, prokaryotic and eukaryotic cell motility and membrane organization.

Selected publications

The physics of molecular motors. [C. Bustamante, D. Keller, and G. Oster (2001) *Acc. Chem. Res.* **34**:412-420]

Regulation of organelle acidity. [M. Grabe and G. Oster (2001) *J. Gen. Physiol.* **117**:329-344]

Reverse engineering a protein: the mechanochemistry of ATP synthase. [H. Wang and G. Oster (2000) *Biochem. Biophys. Acta* **1458**: 482-510]

LIOR PACHTER

MATHEMATICS
http://www.math.berkeley.edu/~lpachter/
lpachter@math.berkeley.edu

Gene finding, alignment, assembly, whole genome analysis

The primary goal of my research effort is to understand the structure, organization and function of the genome. My approach to genomics is based on the observation that comparison of sequences is particularly useful in forming and validating biological hypotheses. The comparative approach to genomics is powerful because of the relationship between functional regions and conserved regions in genomes. Some of my current work and collaboration is based

on the following projects:

Comparative assembly. The problem is how to improve assemblies of whole genomes by using comparisons to sequenced and assembled genomes. The primary application is the assembly of the chimp genome by hanging it off the assembled human genome.

Large alignment problems. The problems of aligning large genomic regions are best tackled with a variety of algorithms and tools from computer science; this is mostly theoretical work.

Cross-species based gene finding. The idea is to improve gene predictions by simultaneously aligning and finding genes in two related organisms rather than one. The theoretical side of the work is based on probability, statistics and combinatorics. Implementations of programs are used to predict genes in newly sequenced genomes.

Whole genome analysis. I'm involved in a collaboration with scientists at LBNL whose aim is to generate biological hypotheses for testing. For example, one project consists of performing whole genome human and mouse comparisons for predicting transcription factor binding sites that can then be tested experimentally.

Selected publications

Applications of generalized pair hidden Markov models to alignment and gene finding problems. [L. Pachter, M. Alexandersson, S Cawley (2001) *RECOMB 2001: Proceedings of the Fifth Annual International Conference on Computational Biology*, 241-248]

From first base: the sequence of the tip of the X Chromosome of *Drosophila melanogaster*, a comparison of two sequencing strategies. [Benos *et al.* (2001) *Genome Res.* **11**: 710-730]

Human and mouse gene structure: comparative analysis and application to exon prediction. [S. Batzoglou, L. Pachter, J. P. Mesirov, B. Berger, and E. S. Lander (2000) *Genome Res.* **10**: 950-958]

JASPER RINE

MOLECULAR & CELL BIOLOGY
http://mcb.berkeley.edu/faculty/GEN/rinej.html
jrine@uclink4.berkeley.edu

Genomic exploration of yeast

The research in my lab is focused on the yeast *Saccharomyces cerevisiae* in which we use genetic analysis to explore issues of gene regulation and cell biology. In addition, we are developing genomic-based approaches to the study of cells by parallel analysis of the expression of all genes simultaneously. Our goal is to create a new kind of genetic analysis in which the genome is used as the unit of function, rather than individual genes or proteins. The gene regulation work focuses on the coupling between certain origins of DNA replication and the establishment of domains with different transcriptional states. The cell biology work focuses on the regulation of the cholesterol biosynthetic pathway and the roles intermediates of this pathway play in the covalent modification of numerous proteins including the Ras oncoprotein and in the trafficking of prenylated proteins.

Upc2p and Ecm22p, dual regulators of sterol biosynthesis in *Saccharomyces cerevisiae* [A. Vik and J.

Rine (2001) *Mol. Biol. Cell* **21**, 6395-6405]

DNA replication-independent silencing in S. cerevisiae. [A.L. Kirchmaier and J. Rine (2001) *Science* **291**, 646-650]

A role for the replication proteins PCNA, RF-C, polymerase epsilon and Cdc45 in transcriptional silencing in *Saccharomyces cerevisiae*. [A. E. Ehrenhofer-Murray, R. T. Kamakaka, and J. Rine (1999) *Genetics* **153**, 1171-82]

DAN ROKHSAR

PHYSICS
http://marichal.berkeley.edu
dsrokhsar@lbl.gov

Decoding genomes and their regulatory networks

My group is broadly interested in the use of both computational and high-throughput experimental methods to decode genomes and their gene regulatory networks. Our ultimate goal is to learn how the genome gives rise to the chordate body plan and especially the nervous system. Working in close collaboration with the Department of Energy's nearby Joint Genome Institute (JGI)—the second largest public sequencing center in the US—we have developed a whole genome assembler for use with large genomes. Our first application was to the 40 MB genome of a white rot fungus, which produces a host of enzymes that degrade lignin and other aromatic compounds. This fungus is the first basidiomycete sequenced to date, also providing insight into the origins of fungal multicellularity. In the coming year, we will be assembling the 400 MB genome of the pufferfish *Fugu rubripes* (the smallest known vertebrate genome), which will be useful for identifying and characterizing conserved genomic elements in human and mouse. A second genome to be analyzed is that of *Ciona intestinalis*, a primitive chordate whose 2,500 cell tadpole has a 340 cell central nervous system that mediates swimming modulated by both light and gravity sensing organs. In collaboration with Michael Levine's lab, we are setting up a functional screen for gene-regulatory elements at the JGI, and developing analytical and computational methods for mining this rich source of data.

Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. [P. Dehal *et al.* (2001) *Science* **293**, 104-111]

The information content of spontaneous retinal waves. [D. Butts and D. Rokhsar (2001) *J. Neurosci.* **21**, 961-73]

Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G. [V. Pande and D. Rokhsar (1999) *Proc. Natl. Acad. Sci. USA* **96**, 9062-9067]

EDDY RUBIN

LAWRENCE BERKELEY NATIONAL LAB
http://www-gsd.lbl.gov/rubin/index.html
emrubin@lbl.gov

Biological annotation of mammalian genomes

The Rubin lab works in two interrelated areas: 1) assigning biological function to high throughput genomic data and 2) addressing hypotheses relevant to complex human disorders using mouse genetics. Recent work has emphasized the use of cross-species sequence analyses and expression profiling to identify gene regulatory sequences in mammals.

Selected publications

Regulation and activity of the human ABCA1 gene in transgenic mice. [L. B. Cavelier, Y. Qiu, J. K. Bielicki, V. Afzal, J. F. Cheng, and E. M. Rubin (2001) *J. Biol. Chem.* **276**, 18046-18051]

Perspectives for vascular genomics. [E. M. Rubin and A. Tall (2000) *Nature*, **407**, 267-269]

Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. [G. G. Loots, R. M. Locksley, C. M. Blankespoor, Z-E. Wang, W. Miller, E. M. Rubin, and K. A. Frazer (2000) *Science* **288**, 136-140]

TERRY SPEED

STATISTICS
http://www.stat.berkeley.edu/users/terry
terry@stat.berkeley.edu

Microarray data analysis and genetic mapping

My research interests all involve the application of statistics to problems in genetics and molecular biology, more generally to genomics. They include mapping disease genes and quantitative trait loci using data from planned crosses or pedigree data, the analysis of DNA and protein sequence data (e.g. for finding genes in genomic sequence), and the design and analysis of experiments concerning gene expression, especially using microarrays. Where possible we collaborate with the biologists who carry out the experiments to produce the data, and where necessary we develop novel methods for the analysis of the data.

Selected publications

Phat—a gene finding program for *Plasmodium falciparum*. [S. Cawley, A. Wirth, T. P. Speed (2001). *Molecular and Biochemical Parasitology* (in press)]

Microarray expression profiling identifies genes with altered expression in HDL deficient mice. [M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin (2000). *Genome Research* **10**, 2022-9]

The limits of random fingerprinting. [D. O. Nelson, T. P. Speed, and B. Yu (1997). *Genomics* **40**, 1-12]

MARK VAN DER LAAN

BIOSTATISTICS
http://www.stat.berkeley.edu/~laan/
laan@stat.berkeley.edu

Computational biology and causality

In collaboration with Michael Eisen, we are concerned with the analysis of databases containing gene expression data on an organism, the complete genome sequence of the organism, and biological knowledge on each of the genes. One of the main goals is to find new binding sites and their corresponding transcription factors.

We also work on the development of statistical methods and algorithms (such as new clustering algorithms) to analyze cancer databases with gene expression data. We consider data sets generated in longitudinal studies that follow up cancer patients over time. These data sets consist of right-censored data on survival times, gene-expression profiles measured at surgery, possibly informative treatment assignments and many biomarkers of interest. We are interested in developing statistical methods that quantify the development of gene expression profiles over time and that link this development to clinical outcomes such as survival time. In addition, we study the application of causal inference methods to these data structures.

Gene expression analysis with the parametric bootstrap. [M. J. van der Laan and J. Bryan (2001) *Biostatistics* (in press)]

Hybrid clustering of gene expression data with visualization and the bootstrap. [M. J. van der Laan and K. S. Pollard (2001). *Journal of Biological Systems* (in press)]

Paired and unpaired comparison and clustering with gene expression data. [J. Bryan, K. Pollard and M.J. van der Laan (2001). *Statistica Sinica* (in press)]

CHRIS VULPE

NUTRITIONAL SCIENCES & TOXICOLOGY
http://nutrition.berkeley.edu/directory/professors_vulpe.html
vulpe@uclink.berkeley.edu

Comparative genomic approaches to copper and iron metabolism

Copper and iron play fundamental role in eukaryotic metabolism. We are using genomic approaches to understand copper and iron metabolism in yeast, plants and mammals. Our approach is to use a systematic analysis of mutants in known copper and iron metabolism genes to identify unknown components of copper and iron metabolism in these species. The metabolic effects of the mutations are determined by: 1) identifying the genes differentially regulated in mutants using cDNA microarrays; 2) quantifying changes in protein production in the mutants using a targeted mass-spectrometry proteomics approach; 3) phenotype analysis of both known mutants and novel components; 4) integrating the effects on gene expression, protein production and biochemical function with informatics to identify metabolic pathways. Comparison of the similarities in response to copper or iron stress in a variety of organisms (yeast, plants, and mammals) will identify the fundamental components of copper and iron metabolism.

Selected publications

Discriminant analysis to evaluate clustering of gene expression data. [C. Hödar, M. Méndez, V. Cambiazo, P. Rámirez, F. Avalos, S. Talbi, A. Armendáris, C. Vulpe, and M. González (submitted)]

Applying robust and resistant regression analysis (MM-estimator) to find significantly expressed genes in microarray data. [A. V. Loguinov, R. Y. Yukhananov, S. I. Mian, and C. Vulpe (2001) *Pacific Symposium on Biocomputing* Poster Abstracts, p. 102]

Using robust and resistant regression analysis (MM-estimator) to find differentially expressed genes in microarray data. [A. V. Loguinov, R. Y. Yukhananov, C. Vulpe and S. Mian (2001) *MSRI Workshop on Nonlinear Estimation and Classification* (submitted).