COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

300 George Street, Suite 501, 203.737.6029
http://cbb.yale.edu
M.S., Ph.D.

Directors of Graduate Studies
Mark Gerstein (Bass 432A, 203.432.6105, mark.gerstein@yale.edu)
Steven Kleinstein (300 George St., Suite 505, 203.785.6685, steven.kleinstein@yale.edu)

Professors Marcus Bosenberg (Dermatology; Pathology), Cynthia Brandt (Emergency Medicine; Anesthesiology), Kei-Hoi Cheung (Emergency Medicine), Ronald Coifman (Mathematics; Computer Science), Chris Cotsapas (Neurology), Stephen Dellaporta (Molecular, Cellular, & Developmental Biology), Richard Flavell (Immunobiology), Joel Gelernter (Genetics; Neuroscience), Mark Gerstein (Biomedical Informatics; Molecular Biophysics & Biochemistry; Computer Science; Stats & Data Science), Antonio Giraldez (Genetics), Jeffrey Gruen (Genetics; Investigative Medicine; Pediatrics), Murat Gunel (Neurosurgery; Genetics), Ira Hall (Genetics), Amy Justice (Internal Medicine; Public Health), Na'ali Kaminski (Internal Medicine; Pediatrics), Yuval Kluger (Pathology), Harlan Krumholz (Internal Medicine; Investigative Medicine; Public Health), Haifan Lin (Cell Biology; Genetics), Shuangge (Steven) Ma (Public Health), Andrew Miranker (Molecular Biophysics & Biochemistry; Chemical & Environmental Engineering), James Noonan (Genetics), Corey O'Hern (Mechanical Engineering & Materials Science; Applied Physics; Physics), Lajos Pusztai (Internal Medicine), John Murray (Psychiatry; Neuroscience; Physics), Andrew Taylor (Emergency Medicine), Zuoheng (Anita) Wang (Public Health)

Associate Professors Julien Berro (Molecular Biophysics & Biochemistry), Forrest Crawford (Public Health), Smita Krishnaswamy (Genetics), Jun Lu (Genetics), Kathryn Miller-Jensen (Engineering & Applied Science), Anna Pyle (Molecular Biophysics & Biochemistry), David Stern (Pathology), Hemant Tagare (Radiology & Biomedical Imaging; Biomedical Engineering), Jeffrey Townsend (Public Health; Ecology & Evolutionary Biology), Hongyu Zhao (Public Health; Genetics), Steven Zucker (Computer Science; Electrical Engineering; Biomedical Engineering)

Assistant Professors Leying Guan (Biostatistics), Samah Jarad (Emergency Medicine), Monkol Lek (Genetics), Bluma Lesch (Genetics), Morgan Levine (Pathology), Zachary Levine (Pathology), Benjamin Machta (Physics), Robert McDougal (Biostatistics), C. Brandon Ogbunu (Ecology and Evolutionary Biology) Serena Tucci (Anthropology), David vanDijk (Cardiology), Jack Zhang (Molecular Biophysics & Biochemistry)

FIELDS OF STUDY
Computational biology and bioinformatics (CB&B) is a rapidly developing multidisciplinary field. The systematic acquisition of data made possible by genomics and proteomics technologies has created a tremendous gap between available data and their biological interpretation. Given the rate of data generation, it is well recognized
that this gap will not be closed with direct individual experimentation. Computational and theoretical approaches to understanding biological systems provide an essential vehicle to help close this gap. These activities include computational modeling of biological processes, computational management of large-scale projects, database development and data mining, algorithm development, and high-performance computing, as well as statistical and mathematical analyses.

To enter the Ph.D. program, students apply to an interest-based track within the interdepartmental graduate program in Biological and Biomedical Sciences (BBS), https://medicine.yale.edu/bbs.

INTEGRATED GRADUATE PROGRAM IN PHYSICAL AND ENGINEERING BIOLOGY (PEB)

Students applying to one of the interest-based tracks of the Biological and Biomedical Sciences program may simultaneously apply to be part of the PEB program. See the description under Non-Degree-Granting Programs, Councils, and Research Institutes for course requirements, and http://peb.yale.edu for more information about the benefits of this program and application instructions.

SPECIAL REQUIREMENTS FOR THE PH.D. DEGREE

With the help of a faculty advisory committee, each student plans a program that includes courses, seminars, laboratory rotations, and independent reading. Students are expected to gain competence in three core areas: (1) computational biology and bioinformatics, (2) biological sciences, and (3) informatics (including computer science, statistics, and applied mathematics). While the courses taken to satisfy the core areas of competency may vary considerably, all students are required to take the following courses: CB&B 562 or CB&B 750, CB&B 740, and CB&B 752. A typical program will include ten course credits. Completion of the core curriculum will typically take three to four terms, depending in part on the prior training of the student. With approval of the CB&B director of graduate studies (DGS), students may take one or two undergraduate courses to satisfy areas of minimum expected competency. Students will typically take two to three courses each term and three research rotations (CB&B 711, CB&B 712, CB&B 713) during the first year. After the first year, students will start working in the laboratory of their Ph.D. thesis supervisor. Students must pass a qualifying examination normally given at the end of the second year or the beginning of the third year. There is no language requirement. Students will serve as teaching assistants in two term courses. In addition to all other requirements, students must successfully complete CB&B 601, Fundamentals of Research: Responsible Conduct of Research (or another course that covers the material) prior to the end of their first year of study. In their fourth year of study, all students must successfully complete B&BS 503, RCR Refresher for Senior BBS Students.

M.D./PH.D. STUDENTS

Students pursuing the joint M.D./Ph.D. degrees must satisfy the course requirements listed above for Ph.D. students. With approval of the DGS, some courses taken toward the M.D. degree can be counted toward the ten required course credits. Such courses must have a graduate course number, and the student must register for them as
graduate courses (in which grades are received). Laboratory rotations are available but not required. One teaching assistantship is required.

**MASTER’S DEGREE**

**M.S. (en route to the Ph.D.)** To qualify for the awarding of the M.S. degree a student must (1) complete two years (four terms) of study in the Ph.D. program (2) complete the required course work for the Ph.D. program with an average grade of High Pass or higher, with ten required course credits taken at Yale including three successful research rotations and (3) meet the Graduate School’s Honors requirement of at least two Honors grades.

**COURSES**

Additional courses focused on the biological sciences and on areas of informatics are selected by the student in consultation with CB&B faculty.

**CB&B 555a / AMTH 553a / CPSC 553a / GENE 555a, Unsupervised Learning for Big Data**  Staff

This course focuses on machine-learning methods well-suited to tackling problems associated with analyzing high-dimensional, high-throughput noisy data including: manifold learning, graph signal processing, nonlinear dimensionality reduction, clustering, and information theory. Though the class goes over some biomedical applications, such methods can be applied in any field. Prerequisites: knowledge of linear algebra and Python programming.

**CB&B 634a, Computational Methods for Informatics**  Staff

This course introduces the key computational methods and concepts necessary for taking an informatics project from start to finish: using APIs to query online resources, reading and writing common biomedical data formats, choosing appropriate data structures for storing and manipulating data, implementing computationally efficient and parallelizable algorithms for analyzing data, and developing appropriate visualizations for communicating health information. The FAIR data-sharing guidelines are discussed. Current issues in big health data are discussed, including successful applications as well as privacy and bias concerns. This course has a significant programming component, and familiarity with programming is assumed. Prerequisite: CPSC 223 or equivalent, or permission of the instructor.

**CB&B 638a, Clinical Database Management Systems and Ontologies**  Staff

This course introduces database and ontology in the clinical/public health domain. It reviews how data and information are generated in clinical/public health settings. It introduces different approaches to representing, modeling, managing, querying, and integrating clinical/public health data. In terms of database technologies, the course describes two main approaches — SQL database and non-SQL (NoSQL) database — and shows how these technologies can be used to build electronic health records (EHR), data repositories, and data warehouses. In terms of ontologies, it discusses how ontologies are used in connecting and integrating data with machine-readable knowledge. The course reviews the major theories, methods, and tools for the design and development of databases and ontologies. It also includes clinical/public health use cases demonstrating how databases and ontologies are used to support clinical/public health research.
CB&B 647b / GENE 645b, Statistical Methods in Human Genetics  Hongyu Zhao
Probability modeling and statistical methodology for the analysis of human genetics data are presented. Topics include population genetics, single locus and polygenic inheritance, linkage analysis, quantitative trait analysis, association analysis, haplotype analysis, population structure, whole genome genotyping platforms, copy number variation, pathway analysis, and genetic risk prediction models. Offered every other year. Prerequisites: genetics; BIS 505; S&DS 541 or equivalent; or permission of the instructor.

CB&B 663b / AMTH 552b / CPSC 552b, Deep Learning Theory and Applications  Smita Krishnaswamy
Deep neural networks have gained immense popularity within the past decade due to their success in many important machine-learning tasks such as image recognition, speech recognition, and natural language processing. This course provides a principled and hands-on approach to deep learning with neural networks. Students master the principles and practices underlying neural networks, including modern methods of deep learning, and apply deep learning methods to real-world problems including image recognition, natural language processing, and biomedical applications. Course work includes homework, a final exam, and a final project—either group or individual, depending on enrollment—with both a written and oral (i.e., presentation) component. The course assumes basic prior knowledge in linear algebra and probability. Prerequisites: CPSC 202 and knowledge of Python programming.

CB&B 711a, Lab Rotations  Staff
Three 2.5–3-month research rotations in faculty laboratories are required during the first year of graduate study. These rotations are arranged by each student with individual faculty members.

CB&B 740a, Introduction to Health Informatics  Staff
The course provides an introduction to clinical and translational informatics. Topics include (1) overview of biomedical informatics, (2) design, function, and evaluation of clinical information systems, (3) clinical decision-making and practice guidelines, (4) clinical decision support systems, (5) informatics support of clinical research, (6) privacy and confidentiality of clinical data, (7) standards, and (8) topics in translational bioinformatics. Permission of the instructor required.

CB&B 750b, Core Topics in Biomedical Informatics  Samah Jarad
The course focuses on providing an introduction to common unifying themes that serve as the foundation for different areas of biomedical informatics. It is designed for students with programming experience who plan to build databases and computational tools for use in biomedical research. Emphasis is on understanding basic principles underlying informatics approaches to interoperability among biomedical databases and software tools, standardized biomedical vocabularies and ontologies, biomedical natural language processing, predictive analytics, information extraction, deep learning, and other related topics.

CB&B 752b / CPSC 752b / MB&B 752b and MB&B 753b and MB&B 754b / MB&B 753b and MB&B 754b / MB&B 754b / MCDB 752b, Biomedical Data Science: Mining and Modeling  Mark Gerstein
Biomedical data science encompasses the analysis of gene sequences, macromolecular structures, and functional genomics data on a large scale. It represents a major practical
application for modern techniques in data mining and simulation. Specific topics to be covered include sequence alignment, large-scale processing, next-generation sequencing data, comparative genomics, phylogenetics, biological database design, geometric analysis of protein structure, molecular-dynamics simulation, biological networks, normalization of microarray data, mining of functional genomics data sets, and machine-learning approaches to data integration. Prerequisites: biochemistry and calculus, or permission of the instructor.