COMPUTATIONAL BIOLOGY AND BIOMEDICAL INFORMATICS

100 College St., cbb-registrar@yale.edu
http://cbb.yale.edu
M.S., Ph.D.

Directors of Graduate Studies
Mark Gerstein (Bass 432A, 203.432.6105, cbb-dgs@yale.edu)
Steven Kleinstein (300 George St., Suite 505, 203.785.6685, cbb-dgs@yale.edu)

Professors Frederick Altice (*Internal Medicine*; *Infections Diseases*; *Epidemiology of Microbial Diseases*), Marcus Bosenberg (*Dermatology*; *Pathology*), Cynthia Brandt (*Emergency Medicine*; *Anesthesiology*), Joseph Chang (*Statistics and Data Science*), Kei-Hoi Cheung (*Emergency Medicine*; *Anesthesiology*), Ronald Coifman (*Mathematics*; *Computer Science*), Stephen Dellaporta (*Molecular, Cellular, and Developmental Biology*), Rong Fan (*Biomedical Engineering*; *Pathology*), Richard Flavell (*Immunobiology*), Joel Gelernter (*Psychiatry*; *Genetics*), Mark Gerstein (*Biomedical Informatics*; *Molecular Biophysics and Biochemistry*; *Computer Science*; *Statistics and Data Science*), Antonio Giraldez (*Genetics*), Jeffrey Gruen (*Genetics*; *Investigative Medicine*; *Pediatrics*), Murat Gunel (*Neurosurgery*; *Genetics*), Ira Hall (*Genetics*), Amy Justice (*Internal Medicine*; *Public Health*), Naftali Kaminski (*Internal Medicine*), Steven Kleinstein (*Pathology*; *Immunobiology*), Yuval Kluger (*Pathology*), Harlan Krumholz (*Internal Medicine*; *Investigative Medicine*; *Public Health*), Haifan Lin (*Cell Biology*; *Genetics*), Shuangge (Steven) Ma (*Biostatistics*), Zongming Ma (*Statistics and Data Science*), Andrew Miranker (*Molecular Biophysics and Biochemistry*; *Chemical and Environmental Engineering*), James Noonan (*Genetics*; *Neuroscience*), Corey O'Hern (*Mechanical Engineering and Materials Science*; *Applied Physics*; *Physics*), Xenophon Papademetris (*Biomedical Informatics and Data Science*; *Radiology and Biomedical Imaging*), Lajos Pusztai (*Internal Medicine*), Anna Pyle (*Molecular, Cellular, and Developmental Biology*; *Chemistry*), David Stern (*Pathology*), Hemant Tagare (*Radiology and Biomedical Imaging*; *Biomedical Engineering*), Jeffrey Townsend (*Public Health*; *Ecology and Evolutionary Biology*), John Tsang (*Immunobiology*), Hua Xu (*Biomedical Informatics and Data Science*), Heping Zhang (*Biostatistics*; *Statistics and Data Science*), Hongyu Zhao (*Biostatistics*; *Statistics and Data Science*), Steven Zucker (*Computer Science*; *Electrical Engineering*; *Biomedical Engineering*)

Associate Professors Julien Berro (*Molecular Biophysics and Biochemistry*), Sidi Chen (*Genetics*; *Neurosurgery*), Forrest Crawford (*Biostatistics*; *Ecology and Evolutionary Biology*), Samah Jarad (*Emergency Medicine*; *Biostatistics*), Smita Krishnaswamy (*Genetics*; *Computer Science*), Bluma Lesch (*Genetics*), Jun Lu (*Genetics*), Ted Melnick (*Biostatistics*; *Emergency Medicine*), Kathryn Miller-Jensen (*Engineering and Applied Science*), John Murray (*Psychiatry*; *Neuroscience*; *Physics*), Renato Polimanti (*Psychiatry*), Edward Stites (*Laboratory Medicine*), Andrew Taylor (*Emergency Medicine*), Zuoheng (Anita) Wang (*Biostatistics*), Yize Zhao (*Biostatistics*)

Assistant Professors Arnaud Augert (*Pathology*), David Braun (*Medical Oncology*), Purushottam Dixit (*Biomedical Engineering*), Salil Garg (*Laboratory*

Medicine; *Pathology*), Leying Guan (*Biostatistics*), Mary-Anne Hartley (*Biomedical Informatics and Data Science*), Albert Higgins-Chen (*Psychiatry*; *Pathology*), Jeffrey Ishizuka (*Internal Medicine*; *Medical Oncology*; *Pathology*), Rohan Khera (*Internal Medicine, Cardiovascular Medicine*; *EPH Biostatistics*), Monkol Lek (*Genetics*), Benjamin Machta (*Physics*), Robert McDougal (*Biostatistics*), Jacob Musser (*Molecular, Cellular, and Developmental Biology*), C. Brandon Ogbunu (*Ecology and Evolutionary Biology*), Carlos Oliveira (*Pediatrics*; *Infectious Diseases*), Steven Reilly (*Genetics*), Wade Schulz (*Laboratory Medicine*), Serena Tucci (*Anthropology*), David van Dijk (*Internal Medicine, Cardiology*; *Computer Science*), Rex Ying (*Computer Science*), Jack Zhang (*Molecular Biophysics and Biochemistry*)

FIELDS OF STUDY

Computational biology and biomedical informatics (CB&B) is a rapidly developing multidisciplinary field. The past two decades have witnessed a revolution in the biological and biomedical sciences driven by the development of technologies such as high-dimensional phenotypic profiling, next-generation sequencing, macromolecular structure determination and high-resolution imaging, wearable sensor devices, and large-scale electronic health records. These data-generation technologies demand new computational analysis approaches, which, in turn, have given rise to the field of computational biology and biomedical informatics (CB&B).

The Yale Computational Biology and Biomedical Informatics program combines research training opportunities in a range of different fields within the biological and biomedical sciences, in addition to the computational sciences, applied mathematics, statistics, and data science. The scope and balance of a student's program are highly individualized. Each student in the CB&B program develops, with the assistance of faculty advisers, a specific program of coursework, independent reading, and research that gives a depth of coverage and fits their background, interests, and career goals.

To enter the Ph.D. program, students apply to the CB&B 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 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 biomedical informatics, (2) biological sciences, and (3) informatics (including computer science, applied mathematics, statistics, and data science). 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 740 and CB&B 752. CB&B requires

a minimum of 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. 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. 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 no later than the end of the third year. There is no foreign language requirement. Students will serve as teaching assistants in two terms. 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

Terminal Master's Degree Program Students can be admitted for a terminal M.S. degree in Computational Biology and Biomedical Informatics with the goal of gaining competency in three core areas: (1) computational biology and biomedical informatics, (2) biomedical sciences, (3) informatics (including computer science, applied mathematics, statistics, and data science). This is a two-year program. Students must complete nine courses at Yale, including at least three graduate CB&B courses (including CB&B 740 and CB&B 752), two graduate courses in the biological sciences, two graduate courses in areas of informatics, and two additional courses in any of the three core areas. In addition, M.S. students must take a one-term graduate seminar on research ethics and attend a CB&B seminar series. Finally, students must meet all of the Graduate School's requirements for the two-year terminal M.S. degree.

Terminal M.S. degree students are also expected to complete an M.S. project, write a research paper describing it, and defend the project in a seminar where they present the project and answer questions about the project as well as demonstrate breadth knowledge of their coursework and track of study. The paper is evaluated by the student's research supervisor and a second reader from the CB&B faculty. Students are expected to identify a faculty member to supervise the M.S. project by the end of the first year or early in the second year. Completion of the research paper is facilitated by enrolling in CB&B 650.

M.S. (en route to the Ph.D.) Students enrolled in the Ph.D. program may be awarded an M.S. degree en route as they satisfy the requirements for the Ph.D. degree. To qualify for the awarding of the en route 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 ten required course credits taken at Yale including three successful research rotations; and (3) meet the Graduate School's grade requirements.

CB&B 523a / ENAS 541a / MB&B 523a / PHYS 523a, Biological Physics Yimin Luo
This course has three aims: (1) to introduce students to the physics of biological systems, (2) to introduce students to the basics of scientific computing, and (3) to familiarize students with characterization methods and analysis tools. We focus on studies of a broad range of biophysical phenomena including diffusion, polymer statistics, entropic forces, membranes, and cell motion using computational tools and methods. We provide intensive tutorials for Matlab including basic syntax, arrays, functions, plotting, and importing and exporting data.

CB&B 562b / AMTH 765b / ENAS 561b / INP 562b / MB&B 562b / MCDB 562b / PHYS 562b, Modeling Biological Systems II Thierry Emonet
This course covers advanced topics in computational biology. How do cells compute, how do they count and tell time, how do they oscillate and generate spatial patterns? Topics include time-dependent dynamics in regulatory, signal-transduction, and neuronal networks; fluctuations, growth, and form; mechanics of cell shape and motion; spatially heterogeneous processes; diffusion. This year, the course spends roughly half its time on mechanical systems at the cellular and tissue level, and half on models of neurons and neural systems in computational neuroscience. Prerequisite: a 200-level biology course or permission of the instructor.

CB&B 568b, Applied Artificial Intelligence in Healthcare Andrew Taylor and Wade Schulz
Recent advances in machine learning (ML) offer tremendous promise to improve the care of patients. However, few ML applications are currently deployed within healthcare institutions and even fewer provide real value. This course is designed to empower students to overcome common pitfalls in bringing ML to the bedside and aims to provide a holistic approach to the complexities and nuances of ML in the healthcare space. The class focuses on key steps of model development and implementation centered on real-world applications. Students apply what they learn from the lectures, assignments, and readings to identify salient healthcare problems and tackle their solutions through end-to-end data engineering pipelines.

CB&B 570b, Privacy-Enhancing Technologies in Biomedical Data Science Hoon Cho
Biomedical data science increasingly depends upon access to large and diverse collections of sensitive human subject data. Conventional data sharing frameworks offer limited privacy protection, often resulting in isolated data silos that hinder scientific collaboration. This course explores Privacy-Enhancing Technologies (PETs) as a solution to these challenges. Specific technologies covered include secure multiparty computation, homomorphic encryption, differential privacy, federated learning, and trusted execution environments. We examine the landscape of privacy risks in biomedicine and study the conceptual and mathematical foundations of PETs as well as their applications in a range of biomedical domains, including genomics and health informatics. Additional special topics delve into the latest developments in this field, concerning both technical and social aspects of PETs. Students engage in hands-on experiences throughout the course, including privacy attack demonstration, literature survey, and the implementation of PET algorithms for various biomedical tasks. Prerequisites: We expect students to have some level of mathematical maturity,

including an understanding of probability/statistics and experience writing proofs. We also expect students to be comfortable with Python programming; homeworks include hands-on programming components in Python. A basic understanding of biology and genetics is helpful but not required. Feel free to contact us if you have any questions regarding the requirements.

CB&B 571a, Data Science Grant-Writing Practicum Lucila Ohno-Machado
This is a hands-on course where students review funded and non-funded grant proposals for different types of NIH awards, as well as the critiques provided by the reviewers. Proposals in informatics and data science are different than traditional basic sciences proposals and clinical research proposals, so this course is specific for those proposing data science and informatics innovation that can be applied in biology and/or medicine. Although there is an emphasis on K (mentored) and F awards, we also cover the basics of R (non-mentored) research awards. Instructors and classmates review proposals that students prepare as part of the course.

CB&B 574a or b, Biomedical Natural Language Processing: Methods and Applications Staff
This course examines current natural language processing (NLP) methods and their applications in the biomedical domain. It provides a systematic introduction to basic knowledge on NLP and AI (e.g., linguistics, machine learning, and deep learning algorithms), advanced NLP tasks (e.g., information extraction, information retrieval, question answering), and corresponding approaches including the recent large language models (LLMs) and hands-on experience in developing biomedical NLP systems for different applications, ranging from biomedical literature mining to clinical decision support. Assessment in this course consists of technical exercises, exams, and projects, to demonstrate the applicability of skills learned during the course.

CB&B 575a, Bioinformatics Applications in Biomedicine Jihoon Kim
This course covers the latest advances in bioinformatics in the context of human diseases. Students learn background knowledge and practical skills to analyze omics data for human disease research. By the end of this course, students should be able to: (1) process bioinformatics data with linux-based pipelines and data tools, (2) apply exploratory data analysis techniques in Python and R, (3) perform analysis of DNA, RNA, and protein data, and (4) conduct a biobank-scale analysis using the platform such as the All of Us Research Workbench.

CB&B 576b, Foundations of Real World Data Science: Electronic Health Records Daniella Meeker
The course covers scientific principles, best practices, and limitations of using observational data from administrative records, including hypothesis generation, feasibility assessments, and causal inference. Students learn pragmatic skills required to prepare analytic data from large, complex transactional databases. We cover methods for data quality characterization and profiling for study planning. Coursework includes application of methods for creation and validation of computable phenotypes, electronic clinical quality measures, and derived analytic variables. Skills include preparation of real-world data for visualization and reporting in business intelligence tools commonly used in population health and health administration. Students reproduce results from published literature using existing databases for predictive modeling, public health, and outcomes research. Completion of this course positions students for externships in healthcare analytics and health data science. Prerequisites: BIS 638, Clinical Database

Management Systems and Ontologies (or equivalent); proficiency in SQL and Python, or R; HIPAA and HSR training; YNHHS Research Basic Access; COS550 or equivalent; execution of DUAs for public and proprietary databases; demonstrated use of YU-YNHHS data science platform.

CB&B 601b, Fundamentals of Research: Responsible Conduct of Research Staff
A weekly seminar presented by faculty trainers on topics relating to proper conduct of research. Required of first-year CB&B students, first-year Immunobiology students, and training grant-funded postdocs. Pass/Fail.

CB&B 634a, Computational Methods for Informatics Robert McDougal
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 Kei-Hoi Cheung and George Hauser
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 647a / GENE 645a, 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 / GENE 663b, 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 and CB&B 712b and CB&B 713b, Lab Rotations Steven Kleinstein
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 Andrew Taylor
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 interoperation 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 and Matthew Simon
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.