An Interdepartmental Ph.D. Program in Computational Biology and Bioinformatics: The Yale Perspective

Mark Gerstein, Ph.D. [1,2], Dov Greenbaum[1], Kei Cheung, Ph.D. [3,4,5], Perry L. Miller, M.D., Ph.D. [3,4,6]

[1]Department of Molecular Biophysics and Biochemistry, [2]Department of Computer Science,

[3]Center for Medical Informatics, [4]Department of Anesthesiology, [5]Department of Genetics,

[5]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT

Running Head: Yale's Interdepartmental PhD Program in CBB

Keywords: Bioinformatics, Computational Biology, Training

To whom correspondence should be addressed:

Prof. Mark Gerstein, MB&B Department, Bass 432A, 266 Whitney Avenue, Yale University, New Haven, CT 06520; mark.gerstein@yale.edu; 203-432-8189

1. Introduction

Computational Biology and Bioinformatics (CBB), the terms often used interchangeably, represent a rapidly evolving biological discipline. With the clear potential for discovery and innovation, and the need to deal with the deluge of biological data, many academic institutions are committing significant resources to develop CBB research and training programs. Yale formally established an interdepartmental Ph.D. program in CBB in May, 2003. As of September, 2004, the program has four students entering year 3, three students entering year 2, and four students entering year 1. This paper describes Yale's program, discussing the scope of the field, the program's goals and curriculum, as well as a number of issues that arose in implementing the program.

2. What is Computational Biology and Bioinformatics?

The Oxford English Dictionary traces the term Bioinformatics back to 1978, with the term Computational Biology having an even earlier usage. Much early work focused on the interpretation and simulation of molecular structures. Other early work in evolutionary biology involved with computing of phylogenies. The difference between the two terms is subtle and for the most part, historical. There is significant overlap and increasing amount of blurring as to exactly what each term implies. Bioinformatics often refers more to the application of the principles of information sciences and other technologies to make biomedical data more understandable and to add value. Computational biology often refers more to the use of mathematical and computational approaches to deal with theoretical and experimental problems in the biomedical sciences.

CBB is a broad discipline that includes 1) the management, analysis, and integration of diverse types of biological data, 2) the modeling of biological systems and biological structures, and

3) the use computational techniques to support and enable virtually all areas of bioscience research. The field is undergoing rapid evolution and growth, and will continue to expand in its scope in the years to come.

Within the past decade, the growth of experimental technologies such as microarrays, protein arrays, yeast two hybrid, high throughput DNA sequencing, and protein mass spectrometry, and the simultaneous advances in computational aspects such as the Internet, databases, and algorithms, has created a robust and dynamic environment for the development of CBB methodologies. Additional impetus has stemmed from the sequencing of an increasing number of genomes, most prominently, the human genome. The unprecedented amount of data contained in the these genomic sequences, coupled with the desire to understand and interpret it, has fundamentally changed the nature of molecular biology. As a result, while many components of CBB have been in existence for a while, the field itself has not been widely established as an independent discipline in the biological sciences until fairly recently.

One way in which we define CBB is as a field conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied math, computer science, and statistics) to understand and organize the information associated with these molecules on a large-scale. In general, the aims of CBB are three-fold. First, CBB organizes data in a way that allows researchers to access existing information and to submit new entries as they are produced: for example, using the Protein Data Bank for 3D macromolecular structures. While data curation is an essential task, the information stored in these databases is essentially useless until analyzed. Thus the purpose of CBB extends much further. A second aim is to develop tools and resources that aid in the analysis of data. For example, having sequenced a particular protein, it is of interest to compare it with previously characterized sequences.

This needs more than just simple text-based searching, and programs such as FASTA and PSI-BLAST must consider what constitutes a biologically significant match. Development of such resources dictates expertise in computational theory, as well as a thorough understanding of biology.

A third aim is to use these tools to analyze the data and interpret the results in a biologically meaningful manner. Traditionally, biological studies examined individual systems in detail, and frequently compared them with a few other related ones. In CBB, we can now conduct global analyses of large amounts of available data with the aim of uncovering common principles that apply across many systems and highlight novel features.

Another major focus relates to genomics and proteomics. 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. The DNA sequence has been determined for a number of organisms and a draft of the genome of humans has been completed. Future biological research will determine 1) the function of the tens of thousands of genes identified by the genome analysis of these different organisms, 2) how the different genes are regulated, and 3) how they work together to mediate complex biological processes. Such information is expected to help elucidate the processes that are critical for the development of organisms as well as the genetic basis of disease. It is also expected that this

information will be useful for developing genetic strategies to manipulate plant and animal genomes for a variety of agricultural and medical applications.

3. CBB: A National Perspective

CBB programs have, as a major goal, the idea of bringing together many different types of scientists, including: 1) computational scientists who can develop algorithms, computer languages, software, data structures, knowledge frameworks, and hardware, 2) bioinformaticists who have the capability to then adjust and use the resources developed by the computational scientists to tackle biological issues, and 3) experimental and clinical biomedical and behavioral researchers, who generate the underlying data that is then mined and modeled using the methodologies of the other two groups.

The national need for research and training in this general area was highlighted in early 2000 by the "BISTI" (Biomedical Information Science and Technology Initiative) report written by an NIH Working Group formed at the request of Dr. Harold Varmus, former NIH Director. (See: http://grants2.nih.gov/grants/bistic/bistic.cfm.) The BISTI report responds to the recognition that massive amounts of computer-related infrastructure and research will be required to enable future research discoveries in the biosciences. A particular focus of the report is on the need for training new scientists in these areas. Indeed, there is currently great demand for researchers trained in CBB both at academic institutions and in industry. National and international interest in this area is growing very rapidly. BISTI was a spur to develop our program and other programs across the country.

In the process of implementing Yale's CBB program, we explored the status of CBB-related programs at other educational institutions. According to a 2002 survey of American universities by

Bioinform (www.bioinform.com), some 31 universities were identified as offering a PhD in Bioinformatics or CBB. (The exact name used for the field varies somewhat from institution to institution.) At these universities, the degree was often offered as a track/concentration in a department such as biology, biostatistics, computer science, or biomedical informatics. The fact that only 6 of these programs existed prior to 1999 and that 12 had been established in 2001 or 2002 attests to the rapidity with which the academic community has recognized and responded to the need to train researchers in this field.

There are a variety of ways in which these PhD programs might be categorized. We found that it was helpful in explaining the field to others to use the following three categories.

- Multi-disciplinary "biology-centered" programs These are based in biologically-oriented departments or programs, and have curriculum similar to other biological disciplines. These programs tend to require between 6 to 9 semester courses (or an equivalent load of trimester or quarter courses). Students usually do three intensive research rotations during their first year, and start work in a chosen PhD supervisor's laboratory after the first year. As described below, Yale's program is based on this biology-centered model.
- Engineering-centered programs These tend to start with two years, and sometimes three years, of coursework. Concentrated work on the PhD dissertation project does not typically start until this coursework is largely completed.
- Programs based in Biomedical Informatics These programs may combine CBB with some exposure to clinical concepts and clinical informatics, and may require two years or so of coursework before concentrated work on the PhD dissertation project commences.

4. Why Develop an Interdepartmental Program?

One question that we confronted is whether a PhD program in CBB should be offered as a specialization track within one or more existing departments, or whether it should be offered as an interdepartmental program. Yale has many departments and programs that have overlapping interests with CBB. No existing program, however, covers the highly interdisciplinary nature of CBB. We therefore felt that it was important to establish an interdepartmental program that built on all of Yale's relevant strengths. No individual department at Yale would have been able to host adequately the type of program we envisioned for several reasons.

- CBB faculty are spread across diverse departments at Yale. These include many biological and biomedical departments, as well as departments such as Computer Science, Statistics, Biostatistics, Applied Mathematics, Biomedical Engineering, and other engineering departments. Faculty from any of these departments might supervise CBB dissertations.
- The spectrum of students who will be strong CBB candidates, as well as the range of dissertation projects that they will undertake, do not fit well within any existing department at Yale. The CBB program provides an academic home for this highly collaborative research that builds on many disciplines.
- CBB is not merely the result of the incorporation of a hodgepodge of disciplines. Rather it represents a new way of thinking and tackling biological problems. For example: neither the computer scientist nor the biologist necessarily has the optimal mindset to work on CBB. Rather what is necessary is the synthesis of both into a new field that is somewhat different from its component disciplines. We believe that this fusion of many fields and the subsequent training of

new scientists with this mindset can best be accomplished through the introduction of a new, innovative program.

One major advantage of developing an interdepartmental PhD program is that students can complete our CBB curriculum (described below), but work for any faculty member at Yale (e.g., in a biological department, computer science, statistics, applied mathematics, etc.) without needing to satisfy the curriculum requirements of the advisor's department. This allows the students to take a flexible set of courses that we help them define, but work on PhD dissertation projects that may have a very diverse character depending on the home department of the faculty advisor.

5. General Concepts within CBB that Form a Foundation for the Field

Given the broad nature of the discipline, it is impossible for students to master the full complexities of all the component fields. It is therefore important to extract and describe the most significant features, relative to CBB, for each disciplines associated with CBB.

"Organism-level" Biology A student should have some understanding of basic language and terms of biology such as cells, organs and species. An important concept in bioinformatics, given the stress placed on comparative analysis among genes, proteins, metabolic systems, cells, and organisms, is the idea of biology as a broad continuum of diverse species, and the unifying idea of evolution. As a result, it is worthwhile to have some understanding of biological classification and some of the main types of structures found in organisms and in diverse cell types.

Molecular Biology & Genetics The students should have a thorough comprehension of the main molecules in biology: proteins, DNA, RNA and metabolites. This includes an understanding of their organizational structure (e.g., open reading frames, promoters, introns, and exons), their physical structure (primary, secondary, tertiary, and quaternary), how the cell creates these molecules (e.g., transcription translation, modifications, splicing and mutations), and how they interact and function with each other and the other molecular populations in the cell. It is also important to understand the basic chemistry governing the behavior of molecules such as the forces and interaction between molecules, the concepts of chemical reactions, and the rates at which the reactions occur. Finally students should have some understanding of the major techniques used to glean information about molecules such as how the genome is sequenced through a sequencing reaction, how the structure of molecules is determined, and how the quantitative levels of molecules are determined in cells. These basic ideas and concepts in molecular biology and genetics provide an important platform for an intuitive understanding of the data that will be analyzed. It is important that students, especially those coming from a more computational or physics background, understand the background and specific conditions that result in the data they are analyzing. Biological organisms do not exist in a vacuum, and as such, there are many biological factors that can influence the results of a bioinformatics analysis that have to be taken into account.

Chemistry and Physics The importance of internalizing the concept that all molecules are governed by the laws of chemistry and physics cannot be understated. Thus, a basic knowledge of many of the underlying ideas in these sciences is important. Central problems in bioinformatics such as protein folding and structure determination require more than a rudimentary understanding of the physics and chemistry involved in protein development. Many of the basic ideas are covered in

advanced biochemistry courses where molecular interactions, protein synthesis and degradation and cellular metabolism are learned. Additionally, physics, and to some extent, chemistry, have long been considered quantitative sciences, whereas biology is a relative newcomer as a quantitative discipline. As such there are many techniques and methodologies that have been developed in physics and chemistry that can be applied to biology.

Computational Science The ability to program in a modern computer language is a basic and integral tool in most bioinformatics projects. Additionally, a student needs to have basic understanding of how information is stored on the computer in the form of databases. Knowledge of more advanced computational topics is also valuable. Students should be familiar with various organizational schemes for storing data such as hashes and trees, and how data can be rapidly searched and sorted with a variety of algorithms. The basic concept of temporal complexity in knowledge representation and analysis is very important. Furthermore, the student needs to understand various issues in representing different types of information on the computer such as string information or 3 dimensional coordinates. Presently much of the focus on computation in biology textbooks has been in relation to algorithms. Some basic bioinformatics courses may focus primarily on teaching students a recipe book of individual algorithms without informing the student of the underlying logic and ideas behind these powerful tools, an understanding that is integral to applying algorithms to future novel and unique problems, and to the development of new algorithms and approaches. It is important that CBB students be able to create new tools and new approaches for analyzing data.

Statistics and Applied Math In statistics, important concepts that the student would need to know, include concepts such as distributions, significance, P-values, and how one practically deals with and analyzes data using variety of standard statistical techniques such as regression, clustering and tree construction. Inherent in many bioinformatics analyses are the concepts of probability, data presentation, sampling and hypothesis testing. Also particularly useful for the CBB student is understanding of some of the issues regarding machine learning and how statistics and computer science can be combined with biology within the framework of large data sets so as to identify new inferences that are not immediately obvious in the data. An understanding of applied math relative to bioinformatics includes concepts such as network theory and simulations. The future of bioinformatics will involve whole cell and possibly organism simulations resulting from a broad understanding of how cellular populations interact with each other.

In developing a curriculum for CBB, one confronts the dichotomy and tension between learning in the biological sciences and learning in the physical sciences and engineering. The tradition in the former is more oriented towards understanding basic facts. Students are then expected to use this knowledge to conduct research, independent study, and laboratory experimentation. Much emphasis is placed on learning and using laboratory techniques. There is considerably less focus on graduate coursework and theory. Conversely, the disciplines of math, physics, chemistry and engineering are more oriented towards a completion of a set body of courses followed by an examination on them. There is more emphasis on understanding the underlying theoretical equations and their use in a conceptually integrated fashion to solve problems.

One can debate the relative merits of these two different approaches, but the reality is that one has to synthesize them in creating a CBB curriculum. The curriculum we chose to develop at

Yale reflects balance between these extremes, somewhat tilted towards the biological end of the spectrum.

6. Yale's CBB Curriculum

This section outlines the curriculum of Yale's interdepartmental CBB program. Because of the interdisciplinary nature of the field, we anticipate that CBB students will be extremely heterogeneous in their background and training. As a result, a welcoming/advisory committee helps students individually tailor the curriculum to their background and interests. The emphasis is on gaining competency in three broad core areas of competency:

- computational biology and bioinformatics,
- biological sciences,
- informatics (including computer science, statistics, and applied mathematics).

Specifically, we expect that all CBB Ph.D. students will do the following.

- Take at least nine (9) courses as follows:
  - three (3) core graduate courses in CBB,
  - two (2) graduate courses in the biological sciences,
  - two (2) graduate courses in areas of informatics,
  - two (2) additional courses in any of the three core areas (which may be undergraduate courses taken to satisfy areas of minimum expected competency, as described below),
  - any additional courses required to satisfy areas of minimum expected competency.
- Take a one-semester graduate seminar on research ethics.

- Participate in intensive research rotations.
- Attend a CBB seminar series.
- Serve as a teaching assistant in two semester courses.

Students typically take 2-3 courses each semester and 3 research rotations during the first year. After the first year, students start working in the laboratory of their chosen PhD thesis supervisor. Completion of the core curriculum typically takes about 4 semesters, depending in part on the prior training of the student. Since students may have very different prior training in biology and computing, the courses taken to satisfy the core areas of competency may vary considerably.

As we gain experience tailoring the curriculum to the diverse backgrounds of students who enroll in the program, the CBB Executive Committee will periodically discuss whether the formal requirements should be changed. The overall goal will be to assure that the students leave Yale with a solid foundation upon which to build careers as leaders in this field.

Areas of Minimum Expected Competency This section describes our approach to defining areas of minimum expected competency. Some students may have satisfied all of these areas prior to entering our program. Other students may need to take undergraduate or graduate courses at Yale to satisfy one or more of these areas. We consider it generally desirable for students to have training in the following areas.

- Biology Introductory biology and biochemistry.
- Computer Science Introduction to programming. Introduction to the concepts, techniques, and applications of computer science. Data structures (arrays, stacks, queues, lists, trees, heaps, graphs), sorting and searching, storage allocation and management, data abstraction.

- Math and Statistics Multivariate Calculus, Linear Algebra and Introductory Statistics.

Each student's background is examined from the perspective of these areas, and in the context of the student's interests and likely research directions. The goal is to use these areas flexibly as guidelines to help identify specific areas in which coursework would be desirable. It is not essential, however, that every student take courses that cover every topic outlined above.

CBB Qualifying Exam Each CBB student is expected to take an oral qualifying exam, typically toward the end of the second year or at the beginning of the third year of study. The student first has a pre-qualifying meeting with the qualifying committee where a short summary of the proposed dissertation topic is discussed, and a set of 3-4 additional topics are identified as areas for questions during the exam. The student then submits a written prospectus describing the proposed thesis research (15-20 pages double-spaced). At the qualifying examination, the student answers questions about the proposed research as well as the additional topic areas.

7. Summary

Our program is designed to produce students competent in a broad range of CBB approaches and techniques. Depending on their focus during their studies, we expect our students to be competent in a varied set of areas. Graduating students should have the ability to independently analyze data, store and integrate heterogeneous data, and propose methodologies for mining the data. They should have a clear understanding of molecular structure, the interrelationship between species and other central dogmas of bioinformatics – being able to apply these concepts and ideas to varied problems and issues.

CBB graduates can look forward to rewarding career prospects in all sectors of biomedical sciences. Given the pivotal role of bioinformatics in various large projects (such as genome-scale projects) and the continued growth and development of emerging fields such as functional and structural genomics, proteomics, and systems biology, we believe that the students should be attractive candidates for a range of different positions and careers in academia and industry, both in and outside biomedical research.

Appendix I: CBB Core Courses

As described previously, our curriculum requires that each student take three (3) graduate CBB courses, two (2) graduate courses in the biological sciences, two (2) graduate courses in informatics, and two (2) elective courses in any of the three core areas. To help make this requirement more concrete, we list below the four current CBB courses, from which three CBB courses must be chosen. The other courses can be chosen from a very wide variety of possible courses given by many departments at Yale. For more detail see: http://ycmi.med.yale.edu/CBB/CBBProg.html, and http://info.med.yale.edu/bbs/main.html.

MBB 752a, Genomics and Bioinformatics Genomics describes the determination of the nucleotide sequence as well as many further analyses used to discover functional and structural gene information about all the genes of an organism. Topics include the methods and results of analysis on a genome-wide scale as well as a discussion of the implications of this research. Bioinformatics describes the computational analysis of gene sequences and protein structures on a large scale. Topics include sequence alignment, biological database design, geometric analysis of protein structure, and macromolecular simulation.

MCDB 750b, Core Topics in Biomedical Informatics Introduction to common unifying themes that serve as the foundation for different areas of biomedical informatics, including clinical, neuro-, and genome informatics. Emphasis is on understanding basic principles underlying informatics approaches to biomedical data modeling, interoperation among biomedical databases and software tools, standardized biomedical vocabularies and ontologies, modeling of biological systems, and other topics of interest.

STAT 645, Statistical Methods in Genetics and Bioinformatics Stochastic modeling and statistical methods applied to problems such as mapping quantitative trait loci, analyzing gene expression data, sequence alignment, and reconstructing evolutionary trees. Statistical methods include maximum likelihood, Bayesian inference, Monte Carlo Markov chains, and some methods of classification and clustering. Models introduced include variance components, hidden Markov models, and Bayesian networks.

CHEM 526a, Computational Chemistry and Biochemistry An introduction to modern computational methods employed for the study of chemistry and biochemistry, including molecular mechanics, quantum mechanics, statistical mechanics, and molecular dynamics. Special emphasis on the hands-on use of computational packages for current applications ranging from organic reactions to protein-ligand binding and dynamics.

Appendix II: Example CBB Student Programs

This section makes our discussion of Yale's CBB curriculum more concrete by outlining the courses that might be taken by students with three different types of previous training.

A Computer Science Major with Several Biology Courses The first example is a student who majored in computer science and took several biology and science courses as an undergraduate. This student might take the following courses:

CBB	MBB 752a, Genomics and Bioinformatics
	MCDB 750b, Core Topics in Biomedical Informatics
	STAT 645b, Statistical Methods in Genetics and Bioinformatics
	CHEM 526a, Computational Chemistry and Biochemistry
Biological Sciences	MBB 600a, Biochemistry I
	MBB 601b, Biochemistry II
	MCDB 570b, Biotechnology
Informatics	CPSC 545b, Introduction to Data Mining
	CPCS 577b, Neural Networks for Computing

A Biology Major with some Courses in Computer Science and/or other Computing Experience The second example is a student who majored in biology and has some computer background as an undergraduate. This student might take the following courses:

CBB	MBB 752a, Genomics and Bioinformatics
	MCDB 750b, Core Topics in Biomedical Informatics
	STAT 645b, Statistical Methods in Genetics and Bioinformatics

Biological Sciences	MBB 741a, Structure and Chemistry of Proteins and Nucleic Acids
	MCDB 570b, Biotechnology
Informatics	CPSC 223b, Data Structures and Programming Techniques (an undergraduate course)
	BIS 560b, Database Management in Biomedicine and Epidemiology
	STAT 610a, Statistical Inference
	CPSC 545b, Introduction to Data Mining

A Biology Major with a Masters Degree in Computer Science The third example is a student who majored in biology as an undergraduate and also obtained a Masters degree in computer science. Graduate courses taken elsewhere (as a graduate student) can be used to satisfy our curriculum in the biological sciences and informatics (but not for the core CBB course requirement). This student might take the following courses:

CBB	MBB 752a Genomics and Bioinformatics
	MCDB 750b Core Topics in Biomedical Informatics
	STAT 645b Statistical Methods in Genetics and Bioinformatics
Biological Sciences	MBB 741a Structure and Chemistry of Proteins and Nucleic Acids
	MCDB 570b Biotechnology
Informatics	Requirements satisfied by four graduate Computer Science courses taken elsewhere.