1NF0RMAT10N 0VERL0AD

By Alexander Gelfand

Data is piling up like snowflakes in a blizzard. A new Ph.D. program at Geisel is training students how to figure out what to do with it all.

ILLUSTRATION BY MARIO ZUCCA

The ability to harness big data drives search engines such as Google. It makes Facebook valuable. And it even helped Barack Obama win reelection.

Big data has come to biomedical science as well. High-throughput sequencing, electronic medical records, and other technological developments mean that more information is available today than ever before. This torrent of data promises to transform medicine by allowing researchers to find new patterns and relationships among the endless bytes accumulating in their databases—if they can make sense of it all.

Cue Geisel's new Ph.D. program in quantitative biomedical sciences (QBS). Launched in 2011 as a track within the Department of Genetics, the QBS program is now housed at the Institute for Quantitative Biomedical Sciences (iQBS), an interdisciplinary entity that supports education and research in three core disciplines: bioinformatics, biostatistics, and epidemiology. Epidemiologist Margaret Karagas, Ph.D., biostatistician Tor Tosteson, Sc.D., and computational geneticist Jason Moore, Ph.D. (the director of iQBS), played key roles in getting the program off the ground. Dartmouth's Board of Trustees formally approved the program in 2012, making it the newest of Dartmouth's doctoral programs.

A growing number of schools expose students to some combination of these fields, but the QBS graduate program is unique in offering in-depth cross-training in all of them. Students do three research rotations in their first year—one in each discipline—and the curriculum examines key concepts from each field's distinct perspective.

"It's the only program that I know of where they're really combining epidemiology, biostatistics, and bioinformatics in an equal and integrative way," says bioinformaticist Casey Greene, Ph.D., one of several recently hired faculty for whom the program was a major draw. A 2009 graduate of the Molecular and Cellular Biology program, Greene returned to Hanover in August as an assistant professor of genetics and director of the iQBS Integrative Genomics Laboratory.

That tripartite structure matters, since each of the three legs on which the program rests plays a crucial role in the world of data-driven biomedical research.

As Jason Moore explains, bioinformatics covers everything from data management to data mining, or the art of identifying patterns in large data sets—often by presenting information in more user-friendly forms. Moore, for example, has pioneered techniques for displaying complex data sets as 3-D visualizations, rather than as plain old tables. "Imagine getting an Excel spreadsheet with one million rows, and trying to figure out the relationships between them," says Amar Das, M.D., Ph.D., another recently recruited faculty member with expertise in biomedical informatics.

Epidemiologists and biostatisticians, meanwhile, work hand-in-hand to develop the tools that are needed to run population-based studies of human disease—precisely the kind of studies that further contribute to the rising tide of biomedical data. "Epidemiologists conduct multidisciplinary, hypothesis-driven research, and we generate a lot of data," says Karagas. Biostatisticians, for their part, determine sample sizes, predict measurement error, and quantify the many other factors that can make or break a large-scale study.

Each of these fields is undergoing its own internal revolution, in part because of the sheer volume and variety of information available, and in part because of the complexity of the biomedical problems that researchers now face. Standard statistical methods aren't always up to the task of tracing relationships between more than two variables in a massive data set. Bioinformaticists must figure out how to present reams of data and sprawling webs of relationships to researchers in forms that they can actually use. And epidemiologists must find ways to incorporate data from wildly disparate sources. For example, Karagas recently led a study of risk factors for bladder cancer that used blood samples to examine genetic variation and toenail clippings to measure levels of arsenic in about 2,000 participants. To top it all off, in an era when interdisciplinary, team-based research is becoming the norm, practitioners of each discipline must learn to speak one another's highly technical language.

Producing scientists who can handle those challenges, says Greene, requires "a different kind of student, and a different kind of training."

Greene ought to know; Moore has described him as "the perfect example of the ideal student for the QBS program, had it been in existence when he was here." (Greene worked in Moore's lab as a graduate student, and the two published a number of papers together.)

Greene likes to make computational predictions that can be tested experimentally. Lately he has been investigating the tissue-specific roles that genes play in various organisms, a task that requires pulling together large quantities of data from different sources. A typical analysis might combine gene expression patterns, for instance, with details about the molecules that regulate those patterns, and use computer programs to predict which genes are most likely to drive a particular biological process.

LEFT: Qian Yang and James Rudd are two of the first-year students in the iQBS Ph.D. program.

RIGHT: Amy Gladfelter, Ph.D., a Dartmouth associate professor of biological sciences, is one of a number of Dartmouth and Geisel faculty members who collaborate on teaching first-year students in the iQBS program. Students James Rudd (left) and Craig MacKenzie look on.

PHOTOGRAPHS: Jon Gilbert Fox

MAK1NG sense of these large, diverse data sets requires advanced computational techniques. One such example is machine learning, which involves training computer programs to classify data and make predictions about biological function using statistical methods. "You take lots of different data sets, and using machine-learning methods, you predict for things we don't currently know," Greene says.

Greene and his collaborators have used this approach to accurately predict some of the genes involved in heart development among zebrafish, which are commonly used as a model for human development. They have also developed a method of predicting gene expression in specific cell lineages in human tissue that outperforms the traditional technique of performing tests on lab animals.

Building better computational tools is one thing; getting them into the hands of bench scientists is another. So Greene also builds web portals that give research biologists access to his cutting-edge algorithms. Greene himself relied on one of those portals—IMP, for Integrative Multi-species Prediction—in his cell-lineage work. Though he had identified a gene of interest in human tissue, Greene still wasn't sure of its precise function. To find out, he searched for an analogous gene with the same expression pattern among the various species (rat, mouse, yeast) represented in IMP to get a sense of what it might actually do. Now Greene would like to allow researchers to add their own knowledge and intuition about what's biologically significant to the automated tools and nearly 2,500 data sets that are already bundled into the platform. It's the sort of problem that begs for the interdisciplinary approach espoused by the new graduate program.

"I could see a QBS student working on this," says Greene, adding that labs that mix QBS candidates with students from more traditional molecular biology programs like Geisel's Molecular and Cellular Biology (MCB) Program or its Program in Experimental and Molecular Medicine (PEMM) "are going to be very successful."

Statistical geneticist Christopher Amos, Ph.D., another new faculty member who participates in the program, agrees. And while Amos anticipates profiting from the presence of QBS students in his own lab, he thinks that advantages will accrue on both sides, with candidates benefiting from the varied perspectives offered by investigators from different disciplines. "If you're in an institution without that kind of training program, you tend to stay in your own little corner," he says.

Amos uses statistical methods to analyze how genes influence our susceptibility to disease. One recent lung cancer study involved 15,000 patients and 30,000 controls and required the analysis of more than 300,000 genetic markers. He also heads the new Center for Genomic Medicine, which supports research into the role that genetic variation plays in disease risk and treatment response. And he is confident that QBS students will be poised to take advantage of the tremendous wealth of information that is being generated not only at Dartmouth, but around the world. "You need a bioinformatics background to pull the data out, identify what kinds of questions make sense, and complete the analysis," he says.

That's especially important, says Amar Das, when you're dealing with data from multiple sources and you have no control over how any of it was generated.

Das specializes in developing software tools to support clinical research. In addition to running the Clinical Epi-Informatics Lab, he has been charged with starting up the new Collaboratory for Healthcare and Biomedical Informatics, and he is especially interested in the treasure trove of clinical data that lies buried in electronic medical records.

In the past, Das has used such data to explore how giving patients with HIV different antiretroviral drugs can lead to mutations in the HIV-1 virus. He has also worked to understand how particular sequences of breast cancer treatment affect clinical outcomes. Now he is examining provider data to better understand the overall quality of care: which doctors are patients seeing, for example, and where delays occur.

The applications vary, but all of them depend on people who are "capable of looking at data and making statistically valid statements," says Das. And all of them use quantitative methods to address real-world biomedical problems.

ROB FROST a second-year graduate student, enrolled in the QBS program after working in related fields. Trained as a mechanical engineer at Stanford, Frost took an early interest in computer science. He spent time as an engineer at a Silicon Valley software company and then started a medical informatics consulting company, VectorC, with a Stanford classmate. While with VectorC, Frost took on several projects for the Center for Biomedical Informatics at Harvard Medical School, an experience that convinced him to explore bioinformatics research further. He left VectorC, took a research position at Harvard, and started applying to Ph.D. programs in bioinformatics.

Frost was drawn to the QBS program at Dartmouth because of its rigorous, multidisciplinary approach to computation and statistics. Though he arrived with little exposure to molecular biology, he has already sat in on the core curriculum courses for the MCB program, and he welcomed the opportunity to spend one of his three first-year research rotations working with molecular epidemiologist Carmen Marsit, Ph.D. "Anyone who is going to be trained in biostatistics and bioinformatics will collaborate with epidemiologists and work with their data sets," Frost says. "You need to be able to speak their language."

Frost's own language can be dauntingly abstruse, larded as it is with references to mathematical concepts like sparse generalized eigenvalue methods. In essence, however, he is trying to develop statistical and computational techniques that will allow researchers to use their prior knowledge about gene function to interpret the enormous amount of data now being generated on genetic variation and gene expression.

One tool that's commonly used to help interpret those very large data sets is known as enrichment, or pathway, analysis, a process in which researchers use what they already know about the functions of individual genes to place them in groups that may be associated with particular clinical outcomes. Frost wants to use machine learning to find better ways of leveraging that prior knowledge about gene function to make sense of the vast quantity of new data about genetic variation and expression—even when some information is missing. As he points out, the skills involved in bioinformatics are highly portable. "It's very easy to take that toolkit and apply it anywhere," he says.

The point has not been lost on the people behind the QBS program. According to both Moore and iQBS associate director Caroline Cannon, M.B.A., the program is designed to produce not just scientists, but leaders—people who can act as catalysts in team-based academic research settings and beyond. "If we're building leaders, then we also need to train them to function outside of academic situations," says Cannon, who points out that QBS graduates will have the expertise to contribute to a range of fields, such as forming start-ups, licensing new technologies, informing public policy decisions, and improving basic clinical protocols. Towards that end, Cannon, who received her M.B.A. from Dartmouth's Tuck School of Business, is exploring entrepreneurial possibilities for the program and plans to create a course on the business of running a lab and bringing ideas "from bench to bedside."

Giving students a firm grounding in not one but three demanding disciplines while also preparing them to lead in the lab and the wider world is no small undertaking. But it seems necessary to create what Moore calls "the scientist of the future." And as Amos points out, it's hardly out of character for either the College or its Medical School. "Dartmouth is known for collaboration," he says. "And this is just another great example of how we'll all work together to make a complex world easier to understand."

"ANYONE WHO IS GOING TO BE TRAINED IN BIOSTATISTICS AND BIOINFORMATICS WILL COLLABORATE WITH EPIDEMIOLOGISTS…YOU NEED TO BE ABLE TO SPEAK THEIR LANGUAGE."