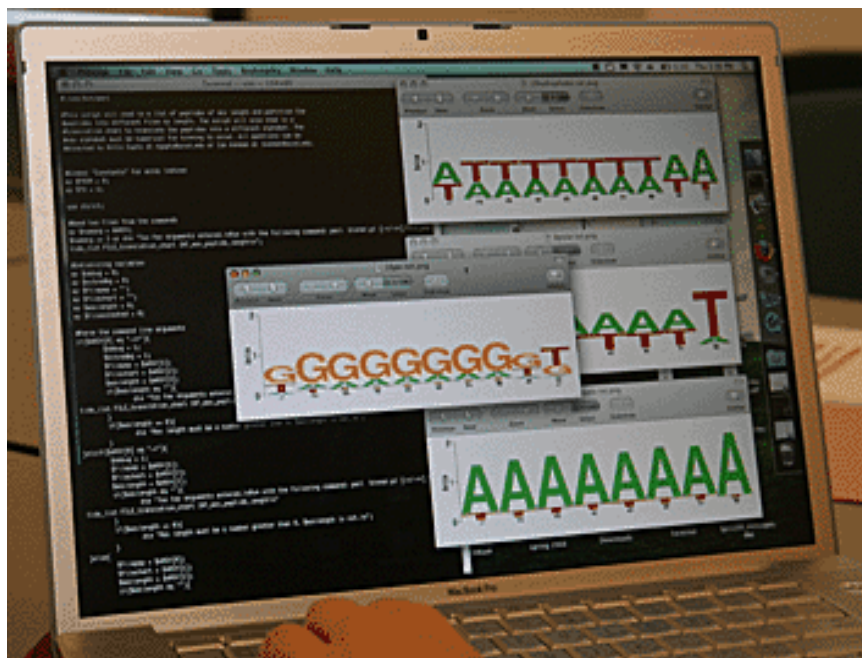


JULY 02, 2008

## Professor Pevzner's Do-It-Yourself Proteomics Class



**Image Title:** Budding bioinformaticists in Pavel Pevzner's undergraduate research experiences class looked at the proteins found in a bacterium and worked backward to identify the genes that created them. - UCSD

The course was not for the faint of heart. Undergraduates in the bioinformatics program at the University of California, San Diego (UCSD) were used to working alongside faculty on “real” projects. But this class, the innocently-named “Research Experience for Undergraduates,” was going to be something very different.

“We told them that it wouldn't be a walk in the park,” recalls Pavel Pevzner, a HHMI professor, a professor of computer science at UCSD, and a tireless advocate for pushing undergraduates into the deep end of bioinformatics research.

---

"We told them it wouldn't be the usual undergraduate research project where you are tightly guarded and you essentially work on a problem with a known solution. We told them that we were starting a new branch of bioinformatics."

- Pavel A. Pevzner

---

Pevzner's idea was to turn undergraduates into experts on genome annotation, which has resulted in a paper published in July in *Genome Research*. It is a first step toward his larger goal of deepening the science research pool. Without more bioinformatics specialists, Pevzner says, we'll either drown in data or, worse, miss the key connections that undergird human health or disease. "There's currently a revolution in biology related to next generation sequencing technologies," he says. "It's becoming cheaper and cheaper to generate genomic and proteomic data." And even more is coming soon. The DNA data of more and more species are now available, and the advent of the "\$1,000 genome" will unleash a flood of individual human genomes too. Bioinformatics has to get smarter to keep up.

With its power to make sense of unimaginably huge data sets, bioinformatics has exploded in all directions since the first genome sequences were analyzed by hand in the late 1970s. Current bioinformatics tools are called on for many tasks: they are used to analyze sets of similar proteins, to compare genomes between species and individuals, to predict molecular structures, to find patterns in drug responses, and to model complex biological systems.

Yet even as genome drafts come spilling out of sequencing robots, they still need a personal touch, careful annotation to weed out errors and spot alternative readings. "In reality, careful, good annotations are still being generated manually by experts," Pevzner explains. "We will quickly run out of experts to annotate these thousands of new genomes."

Pevzner's ambitious class would make undergraduates those experts. They would look at all of the proteins found in a bacterium—the bacterial proteome—and work backward to the genes that created them. What's more, the students would get their bacterial proteomes from a new, relatively unexplored source for this kind of data, high-throughput mass spectrometry. "Mass spec" weighs and sorts molecules into their atomic isotopes. The raw

data would be a long way from a neat protein and peptide list.

Pevzner pitched his course idea to students during a 2007 informational meeting at UCSD. “We told them it wouldn't be the usual undergraduate research project where you are tightly guarded and you essentially work on a problem with a known solution. We told them that we were starting a new branch of bioinformatics.” Prospective students would have to invent this new branch themselves. “So we had some very brave undergraduates.”

“Brave? Definitely not brave. Maybe a little ambitious,” says Jamal Benhamida, a senior at the time. He survived Pevzner's experiment in invent-it-yourself bioinformatics and became one of seven undergraduate co-authors with Pevzner of the equivalent of a class final, a research paper in the July 2008 print issue of *Genome Research*. The paper unveils a new branch of bioinformatics that Pevzner calls “comparative proteogenomics.” Beyond the problem solving skills, Benhamida says the course gave him new confidence. “In research, there's not really a right or a wrong answer. I never thought I would fail because whether or not you find something, you learn something.”

The students started with raw spectrometry data about proteins from three species of an aquatic bacterium called *Shewanella*. The genus is on the scientific agenda at the Department of Energy's Pacific Northwest National Laboratory in Richland, Wash., as a potential “bioremedial” organism. “It essentially breathes metal,” Pevzner explains.

In a contaminated water source, *Shewanella* can reduce heavy metals, such as uranium or chromium, during respiration by absorbing the inert residues, and then, *post mortem*, carrying them along for safe burial in the sediment. Pevzner's longtime collaborator, Dick Smith and his colleagues, ran specimens from three species with sequenced genomes through his mass spectrometry system. Mass spectrometry provides indirect information about proteins by breaking them into electrically charged ions, weighing them, and using these signatures to identify proteins. The process yielded a staggering amount of cryptic data about the smaller peptides that make up the proteins found in *Shewanella*. The students' job was to sort it all out.

Pevzner and Nitin Gupta, the UCSD graduate student supported by Pevzner's HHMI professorship, laid out the problems and the available resources but left the students to define their experiments, select their computer languages, and write their own algorithms. The students then compared their results to the already completed *Shewanella* genomes and contributed to the annotation that explains what genes are where in the genome when they could. In the

process, the students found new proteins, new genes, and new boundaries between genes. They also corrected missing or inaccurate “start sites” for gene translations.

This project in comparative proteogenomics also cleared up many “one-hit-wonders,” single peptides discovered by previous researchers but not linked to a complete protein—essentially abandoned in an experimental twilight zone. That is the beauty of comparative proteogenomics, Pevzner explains. “When you work with a single bacterium, sometimes there are some amazing pieces of evidence but you can't prove that this is new biology -- or just a fluke or an ‘artifact’ of your experiment. However, if you have three different bacteria and you see the same ‘artifact’ in all three, it is probably not an artifact. You're seeing something real.”

The experiment in comparative proteogenomics also offered a more detailed view of the proteomic world itself, Pevzner says. Unlike our relatively stable genes, proteins are in constant flux from the moment they are created. In any given cell, protein numbers and variety change from minute to minute.

Proteins also change purpose through a process called proteolysis that cuts them up into peptide pieces. They become the cell's off-on switches, its throttles and brakes, and its communications system to neighboring cells. For example, signal peptides tag other proteins for delivery or for recycling. In the human brain, neuropeptides are the powerful chemical messengers of memory, emotion, and even social behavior. “They tell us when to go to sleep and when to wake up and whether we are having fun or not,” Pevzner says. His students were able to offer one of the first demonstrations that mass spectrometry coupled with bioinformatics could track proteolysis of proteins into working peptides.

The project was Ngan Nguyen's first research experience, one that she remembers as both exciting and worrying. “As Pavel said, we had to figure out how to do it ourselves. To me, that was exciting part. But I worried a lot too. I didn't know if what I did was right or good enough. Was there some hidden error that I didn't see? Did I cover all the bases?” says Nguyen, who was a junior during the class.

Nguyen credits her survival to constant feedback and advice from Gupta. The class itself served as a group collaboration, with its members offering criticism and practical advice to stay on course. “The excitement was that the project was real and no one had ever done it before,” Nguyen says. “Besides, I always wanted to finish what I was doing to see the results.”

Pevzner admits that the UCSD students who took “Research Experience for Undergraduates” in 2007 were exceptional. Most were seniors in the Jacobs School of Engineering where Pevzner directs the Center for Algorithmic and Systems Biology. And most have gone on to prestigious graduate programs, international fellowships, and top-flight medical schools, he proudly reports. For example, Jamal is now at the University of Chicago Medical School, and Nguyen is part of UCSD's own doctoral program in bioinformatics. But Pevzner regards the class only as a proof-of-principle. He believes that with the right resources, limited mentoring, and a forum for peer discussion, most undergraduates can pull their weight in new bioinformatics research.

In the year since the class ended, Pevzner and Gupta have concentrated on getting their HHMI-supported outreach project ready for a wider test. As an HHMI professor—a program that encourages recognized research scientists to bring the excitement of scientific discovery into the classroom—Pevzner gets \$1 million to develop innovative programs like this one for undergraduates. This summer, they will take their experiment on experiments to the whole world. Called UBER-GRID—the Undergraduate Bioinformatics E-Research Grid—it will be a platform for worldwide, distributed bioinformatics research projects, Pevzner says. “We will put all our projects on the web and invite every student in the world to collaborate.”

The Pevzner lab's projects are not for rank beginners, even if they are intended for undergraduates. Potential projects include annotating bacterial genomes using new mass spectrometry data and improving signal peptide prediction tools. “Instead of meeting with the students in a room, we will meet them on the web. Students in India and in China and in whatever place in the world can collaborate with each other,” he says. The lab will post links to required data sets, genome repositories, downloadable software tools, prediction programs, and literature references.

Students must bring their own curiosity and bravery.