UC San Diego Undergraduates Forge New Area of Bioinformatics

July 2, 2008

Daniel Kane

Undergraduate students from the University of California San Diego have forged a new area of bioinformatics that may improve genomic and proteomic annotations and unlock a collection of stubborn biological mysteries. Their work will be published in the July 2008 issue of the journal *Genome Research.*

The new area of bioinformatics is called "comparative proteogenomics." It sits at the intersection of the fields of "comparative genomics" and "proteomics" - which is the study of all of an organism's proteins.

"This could be a powerful way to improve both genome and proteome annotations and to address notoriously difficult biological problems that remain outside the reach of previously proposed bioinformatics approaches," said Pavel Pevzner, the UC San Diego computer science professor who organized the project.

"Our bioinformatics undergraduates have shown that you can simultaneously analyze multiple genomes and proteomes, and use this information for scientific discovery," said Pevzner, who put together the Bioinformatics [Under]graduate Research Consortium in Comparative Proteogenomics at UCSD.

Nature Reviews Genetics recently highlighted this work. "As the efficiency of high-throughput mass spectrometry improves, it is likely that proteomics will be used increasingly in genome annotation. As well as improving the accuracy of annotation, proteomics can provide information that other annotation methods are blind to, such as RNA editing and novel protein modifications," writes Patrick Goymer in Nature Reviews Genetics.

Battling Floods of Genomic Data

Researchers are currently being flooded with genomic and proteomic data, and the volume is only expected to increase as the genomes of more and more organisms are sequenced. This overwhelming volume of information is making the industry-standard manual genomic annotations less and less feasible, the researchers say.

The new area of comparative proteogenomics offers a promising automated solution to the growing gap between the number of sequenced genomes and researchers' ability to manually annotate them.

"We have shown that you can use the proteins in the proteome data sets to correct what people think the DNA says," explained Jesse Rodriguez, one of the UC San Diego undergraduate researchers publishing in *Genome Research*. "You could do a manual check, but that is expensive. We are letting the proteins do much of the work for us...they let us infer how the genome actually should be labeled," Rodriguez explained during a telephone interview from Stanford University, where he is now in the first year of a Ph.D. program in bioinformatics.

In the *Genome Research* paper, the students looked at three species of the aquatic bacterium *Shewanella* which is both a model organism and a useful creature for bioremediation projects. The team combined proteomic data sets generated by mass spectrometry with comparative genomics data. The work yielded better annotations of the *Shewanella* genomes. The student researchers also identified post translational modifications, proteolytic events and even such important and "exotic" biological mechanisms as programmed frameshifts.

Beyond Comparative Genomics

Comparative proteogenomics marks a significant step beyond comparative genomics, which itself is a relatively new field that capitalizes on the fact that evolution conserves the more important parts of the genome (genes, for example) and recycles less important parts. With comparative genomics, researchers find similar strings of nucleic acids - A, T, G, C - in multiple species in order to identify important genes that have been conserved over millions of years of evolution.

"The power of comparative genomics fades, however, when one starts asking questions about proteomes rather than genomes," said Pevzner, who is the director of the UCSD Center for Algorithmic and Systems Biology, located at the UCSD division of Calit2.

Comparative genomics, for instance, does not provide insights into how proteins break into smaller pieces, an important process called proteolysis that is responsible for many life and death decisions that cells make. In fact, there are still no high-throughput technologies for studying proteolysis. This makes it difficult to characterize signal peptides, neuropeptides, and many other important molecules representing "broken pieces" of various proteins. By looking at multiple genomes and their corresponding proteomes simultaneously, the UCSD researchers say you can start answering some of these tough questions.

For example, comparative proteogenomics, can help solve the one-hit-wonder problem which arises when researchers can only identify a single peptide that belongs to a protein. Without a second peptide from a particular protein, researchers can not confirm that a particular gene is actually expressed in a species.

For each of the three *Shewanella* species, at least 20 percent of identified proteins have only one identified peptide and this "leads to a significant reduction in the number of identified proteins," the authors write.

Resolving such "one-hit wonders" is one of the many ways in which digging into both the proteomes and the genomes at the same time can improve genome annotations or provide other biological insights.

Undergrad Research Experiment

"We took a bold and risky approach to undergraduate research. Instead of applying existing approaches to new datasets, which is very common in undergraduate research, we challenged them to actually develop new approaches," said Pevzner, the brainchild of this undergraduate-dominated research project.

With funding from his Howard Hughes Medical Institute Professor Award, Pevzner organized the Bioinformatics [Under]graduate Research Consortium in Comparative Proteogenomics at UCSD, hired undergraduates for summers, sent undergraduates to scientific meetings and supported Nitin Gupta, the UCSD Bioinformatics Ph.D. candidate who managed the many branches of this research project.

The state-of-the-art bioinformatics algorithms that UCSD undergraduates developed required massive mass spectrometry datasets. To provide the consortium with the required dataset, Gupta and Pevzner collaborated with Dick Smith from Pacific Northwest National Laboratory (PNNL).

"We framed the open questions and introduced the undergraduates to the datasets. We then asked the undergrads to come up with their own questions...and challenged them to find new solutions," said Gupta.

Each student developed the algorithmic tools necessary to do his or her research.

"It was a big learning experience for me. I was already planning on going to grad school, but it was great to have this research experience ahead of time to be able to say 'wow this is really fun...I'm looking forward to doing this for the next 5 years,'" said Rodriguez.

Seven undergraduate and two-first year graduate students are authors on the *Genome Research* paper. They worked either by themselves or in pairs and met weekly with both Gupta and Pevzner, both individually and as a group.

One pair of students, Liz Kain and Ian Kerman, worked together to better understand what proteins are actually being detected in the mass spectrometry data.

Kain graduated from UC San Diego's bioinformatics program in June 2008 and has already accepted a job offer at Apple. "The programming and research experience I gained from this project has been really beneficial. It especially helped in my bioinformatics classes and in preparing for a technical career," said Kain.

Kerman, on the other hand, will be at UCSD a little while longer. He is now working on his master's degree in Biology as a part of a BS/MS program.

"Participating in the consortium gave me invaluable 'dry lab' experience which I am now using to drive and design my 'wet lab' experiments," said Kerman. "I also think the bioinformatics research helps out with my part-time job at Biomatrica," - a San Diego biotech that develops technologies for stabilizing biological samples at room temperature.

"We encouraged teamwork and synergy between the students," explained Pevzner. "We are very proud that many of them were accepted to top bioinformatics graduate programs at Stanford, UCSF, UCSC, and other universities."

Jesse Rodriguez, the undergraduate researcher who is now at working on a Ph.D. at Stanford, also published a paper in the Journal of Proteome Research earlier this year based on his work with Gupta and Pevzner at UCSD.

"The students all know the computer science and the biology. They are kind of superheroes!" said Pevzer.

Author contacts: Pavel Pevzner ppevzner AT cs DOT ucsd DOT edu Nitin Gupta ngupta AT ucsd DOT edu

Genome Research paper "Comparative proteogenomics: Combining mass spectrometry and comparative genomics to analyze multiple genomes," by

Nitin Gupta 1 , Jamal Benhamida 1 , Vipul Bhargava 1 , Daniel Goodman 1 , Elisabeth Kain 1 , Ian Kerman 2 , Ngan Nguyen 1 , Noah Ollikainen 1 , Jesse Rodriguez 1 , Jian Wang 1 , Mary S. Lipton 3 , Margaret Romine 3 , Vineet Bafna1 4 , Richard D. Smith 3 , and Pavel A. Pevzner 1 4

1 Bioinformatics Program, University of California San Diego, La Jolla, California 92093, USA; 2 Division of Biology, University of California San Diego, La Jolla, California 92093, USA; 3 Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, USA; 4 Department of Computer Science and Engineering, University of California San Diego, La Jolla, California 92093, USA

Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.074344.107

Media Contact: Daniel Kane, 858-534-3262


CHARGERS
SAN DIEGO


Preview File Edit View Go Tools Bookmarks Window Help
MacBook Pro