seeing science

SeeingScience

BY KATHARINE MILLER, MANAGING EDITOR

Art That's A BLAST

A group of artists and scientists has created an interactive artwork using BLAST (Basic Local Alignment Search Tool), one of the foundational algorithms for comparative genomics. Normally, the BLAST process of looking for homologous genes is invisible to researchers: They enter a gene sequences into a computer and wait for BLAST to kick out the matches. But the collaborative art work opens that black box by making the BLAST search visible. The result, *Ecce Homology* was shown in early August at SIGGRAPH 2005 in Los Angeles.

"It's artwork that's deeply grounded in science without being didactic," says Ruth West, director of visual analytics and interactive technologies at the University of California, San Diego, National Center for Microscopy and Imaging Research and research associate with the UCSD Center for Research in Computing and the Arts. She heads the collaboration which includes 11 biologists, artists and computer scientists from UCSD, UCLA and the University of Southern California.

Visitors interacting with the Ecce Homology *installation at the Fowler Museum.*

ing set (overfitting) by monitoring the progress of the FES/classifier. The independent test set is used for external cross-validation, but only after completion of the FES and identification of the final classifier. With small datasets, even partitioning into training and test sets is statistically suspect, and k-fold cross-validation is used: the dataset is split into k equal parts (approximately 5 to 10), trained on k to 1 parts and tested on the remaining portion. One then cycles through k times and averages the test results. For small sample sizes, the variance of the averaged test accuracies tends to be unacceptably large, while overtraining is still a threat.

For highly imbalanced classes (e.g., rare disease vs. healthy), overall classification accuracy can be misleading. For example, consider 90 samples in the healthy class, but only 10 in the disease class. Misclassifying all 10 still gives 90 percent overall accuracy. Hence, balanced sensitivity and specificity (i.e., comparable accuracies for both classes) is more appropriate, and can be achieved by undersampling, oversampling or by penalizing misclassifications differently for different classes. (Differing misclassification costs for the classes is an example.)

For each sample, we compute class probabilities. This is relevant clinically (e.g., additional tests would be suggested if a classifier assigned a patient to the disease class with 55 percent probability, immediate treatment would commence if this probability were 90 percent.)

In the biomedical field, the twin curses are generally active. They both must be dealt with in concert, otherwise overly optimistic and frequently wrong conclusions will result. ☐

SeeingScience continued

With a name inspired by Friedrich Nietzsche's *Ecce Homo*, a meditation on how one becomes what one is, the project explores human evolution by examining similarities between genes from human beings and a target organism, in this case the rice plant. *Ecce*

Ecce Homology is a physically interactive new-media work that visualizes genetic data as calligraphic forms.

Homology is a physically interactive new-media work that visualizes genetic data as calligraphic forms. A novel computer-vision based interface allows multiple participants, through their movement in the installation space, to select genes from the human genome for visualization using the Basic Local Alignment Search Tool (BLAST). Five projectors present these changes in *Ecce Homology*'s calligraphic forms across a 40-foot wide wall.

"If we worked on the genomic calligraphy visualization further, it could be useful to scientists," she says, "but the installation is not a tool; it's art. And it's specifically ambiguous and a bit mysterious—by intention."

Ecce Homology, which was first displayed two years ago at the Fowler Museum in Los Angeles, works on many levels both scientifically and artistically. "People assume that there's value in the vast amounts of genomic data we are generating," says West, "but data is not knowledge, and in order for us to derive knowledge from it, we need to interpret it. The more complex it is, the harder it is for human beings to do that and, consequently, the greater our need to find new approaches." So, says West, "we've produced an artwork that both speaks to this need and lets viewers interact fluid-ly with the data in a visceral way."

Ultimately, West says, the exhibit poses the question, "If you were to do work that's truly hybrid art/science, what would that process be like? And would there be any outcome that would point to how art might nurture scientific discovery?"

For more information about *Ecce Homology*, visit www.insilicov1.org. ☐
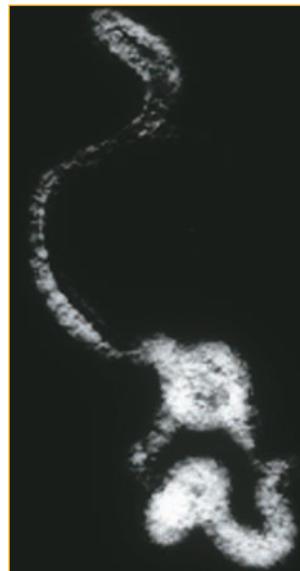
Ecce Homology's *custom software transforms strings of genetic code into luminous, scientifically accurate visualizations that incorporate multiple biological features. For protein sequences, the stroke placement, shape and brush quality are determined by physical and chemical properties, such as the proportion of mass to volume, hydrophobicity, or ionization of the amino acids. The visualization is created from amino-acid sequence chunks that are segmented by a "turn prediction" algorithm. Each segment's corresponding calligraphic stroke is connected to its neighbor by a connection whose shape is based on a secondary structure property of the segment. The result resembles calligraphy. Courtesy: Ruth West*