

JASON R. MILLER, Department of Computer Science, Engineering, and Mathematics, Shepherd University, Shepherdstown WV 25443. Bioinformatics in the Sequencing Era.

Transcriptomics, or the analysis of gene expression, is enabled by DNA sequencing machines that decode the nucleotide sequences of RNA molecules from living cells. The large volumes of data require analysis by computer. For organisms with a known genome sequence, computers align each RNA sequence to the genome sequence to identify the gene responsible. The aggregate counts per gene are used to rank the genes by expression. The computational challenge becomes harder when any of the genes under study are similar to each other. The challenge is hardest when the study is focused solely on pairs of genes that are nearly identical: the maternal and paternal copies of the *same gene*. Nevertheless, there is evidence from plant science that flowering plant genes (maternal) compete with pollen genes (paternal) to reduce each other's gene expression in their seeds. As computer scientists, we built a computational pipeline capable of detecting this phenomenon. We benefited from plant science collaborators who provided insight, validation, and three billion RNA sequences. We exploited bioinformatics technology including noise reduction algorithms, sequence alignment software, parallel computing environments, and statistical analysis software. We believe our pipeline is less sensitive to machine error, and more tolerant of biological variation, than processes employed to date. Interestingly, our project is focused on a question that was not even posed until the era of the Human Genome Project. Thus, as DNA sequencing technology becomes ubiquitous, and more questions become addressable through sequencing, students of computer science and biology can expect to have growing opportunities to collaborate for scientific discovery.