

# MATLAB<sup>®</sup> in Computational Biology

BY KRISTEN AMUZZINI

Two years ago, researchers for the Human Genome Project unveiled a component “glossary” of the human body that will ultimately yield a database of the 3,000,000,000 chemical base pairs (the DNA sequence) in the human genome.

High-throughput sequencing technologies such as those used on the Human Genome Project are producing data at a rate that outstrips the ability of scientists to analyze its therapeutic potential. In the relatively new discipline of computational biology, researchers are addressing this challenge by adopting mathematical and statistical software, computer modeling, and other computational and engineering methods. As a result, computational biology has become the latest engineering discipline.

In addition to screening and analyzing huge data sets, computational biologists, like their counterparts in other engineering disciplines, must integrate diverse teams, tools, and specialties. Computational biologists typically come from computer science, mathematics, or engineering. They must work closely with life scientists who understand molecular biology or chemistry but are not programming or math experts.

## Selecting Software

To address these challenges, computational biologists require software that is flexible, supports diverse applications, is scalable, can handle increasingly large data sets, and provides deployment capabilities. It is therefore not surprising that many have adopted MATLAB for a spectrum of tasks, from statistical analysis to pharmacokinetic modeling and application deployment.

## Analyzing and Visualizing Data

Biotechnical and pharmacological researchers use MATLAB and add-on toolboxes to perform data analysis and image

processing, analyze statistics, fit curves to data, and create neural networks. With these products, they perform sequence analysis and alignment, microarray data analysis and normalization, and mass spectrometry data analysis, among other computations, without manually integrating multiple tools.

## Developing Models

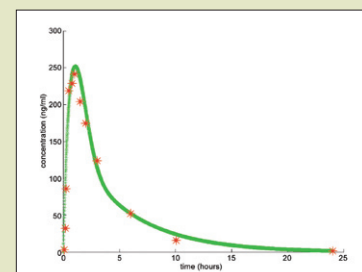
Working in Simulink<sup>®</sup>, researchers build physiologically based pharmacokinetic (PBPK) models to predict the efficacy of potential new drugs and determine the most promising dosing regimens for humans before conducting clinical trials. The models typically include a set of known parameters, such as tissue volume and blood flow rate, and an unknown set, such as rate of absorption and partition coefficients. The Simulink graphical block-diagram environment enables them to organize the organs of the body physiologically, making it easier to explain research findings to specialists in other fields.

## Solving Large-Scale Problems

Even with efficient numeric math and algorithms, many computational tasks in the life sciences are too large and complex to be handled by a single computer. Batch processing of mass spectrometry profiles, for example, may involve terabytes of data. Another very large-scale computing problem is an iterative test, commonly referred to as “bootstrapping,” that delves into evolutionary trees to determine the significance of each branch of the tree.

MATLAB tools for distributed computing help researchers tackle large data sets by dividing algorithms or models into independent tasks that run simultaneously across a cluster of computers. Deployment tools enable them to share their applications as Microsoft Excel plug-ins or as stand-alone executables.

To guide clinical trials, **Novartis** uses MATLAB and Simulink to create PBPK models that describe the uptake and distribution of a drug given to animal species or humans over time. They create these PBPK models using a system of ordinary differential equations, with an associated set of variables and parameters. “We can quickly customize our models to a particular compound or biological system by changing parameters,” says Brian Stoll, senior scientist at Novartis. “The wide use of MATLAB in our industry also facilitates collaboration with other researchers.”

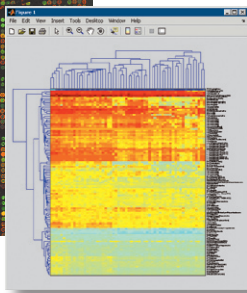
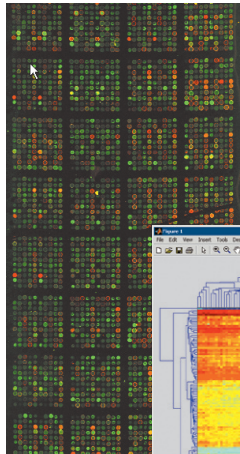


Comparison of experimental and model-predicted tissue concentration of Novartis drug as a function of time following oral administration.

## Research Methods

Biological data is highly heterogeneous and requires a range of analysis methods, including microarrays, mass spectrometry, flow cytometry, microscopy, and two-dimensional gel electrophoresis.

A **microarray** enables scientists to analyze the vast amount of information contained within a genome, such as how genes behave under specific conditions. Before microarrays, scientists studied one gene at a time, and experiments could take days or weeks. Using microarrays, they can study thousands of genes at once.



Microarrays (left) provide a high-throughput mechanism to study large amounts of gene expression. A two-dimensional clustergram (right) indicates relationships between groups of genes.

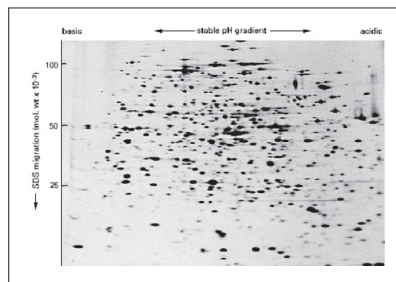
To create a microarray, scientists apply a set of cloned DNA molecules to a solid matrix, such as a microscope slide. Using a scanning microscope, they then measure how much of a specific DNA fragment is present. Analysis of microarray data involves normalizing the data to compare samples; clustering the data to identify groups of genes that behave in a similar way; and visualizing the data to view patterns, trends, or other characteristics.

**Mass spectrometry** is a powerful analytical technique used to identify and measure biological and chemical compounds. It involves introducing enough electrical energy into a target molecule to cause its ionization and disintegration. Researchers analyze the resulting fragments, based on the mass-charge ratio, to produce a molecular “fingerprint” that is used to determine which proteins are present. Analysis involves preprocessing and normalizing the data, aligning the spectra, and searching for statistically significant peaks between samples.

In **flow cytometry**, a laser-powered instrument is used to measure the amount of DNA in cells by monitoring fluorescent reporter proteins placed in the cells, helping researchers to evaluate a person’s risk of developing some cancers.

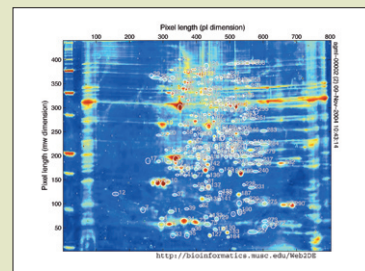
Like flow cytometry, **microscopy** helps quantify protein activity at a cellular level by capturing images of multiple cells over time.

**Two-dimensional gel electrophoresis** enables researchers to separate and identify molecules based on their speed of movement through an electrically charged field. This technique is performed for analysis or to partially purify molecules before applying other methods, such as mass spectrometry, PCR, cloning, DNA sequencing, or immunoblotting.



Two-dimensional gel analysis enables researchers to identify the proteins in a given sample. Gel image courtesy of Alan W. Partin, M.D., Ph.D., The Johns Hopkins University School of Medicine.

Researchers in the Department of Biostatistics, Bioinformatics, and Epidemiology at the **Medical University of South Carolina (MUSC)** use MATLAB to develop applications for genomic and proteomic analysis, such as biomarker identification, two-dimensional gel analysis, and artificial neural networks. They make these packages available to other groups and the scientific community via the Web using the MATLAB Web Server.



Two-dimensional gel analysis.

## The Future

Computational biologists are currently working to integrate analysis methods, such as microarray and mass spectrometry analysis, to give them multiple views into genomic and proteomic data sets and improve their understanding of diseases and medical conditions. They are also inventing ways to sequence the human genome that will eventually enable doctors to develop treatments precisely calibrated to individual patient genomes. Tasks such as these would be impossible without the latest advances in hardware and software. ◀

## RESOURCES

- ▶ **Computational Biology User Stories**  
[www.mathworks.com/res/compbiostories](http://www.mathworks.com/res/compbiostories)
- ▶ **Webinar: MATLAB for Bioinformatics**  
[www.mathworks.com/res/bioinfowebinar](http://www.mathworks.com/res/bioinfowebinar)
- ▶ **Book: Mathematical Models in Biology: An Introduction**  
[www.mathworks.com/res/book5482](http://www.mathworks.com/res/book5482)