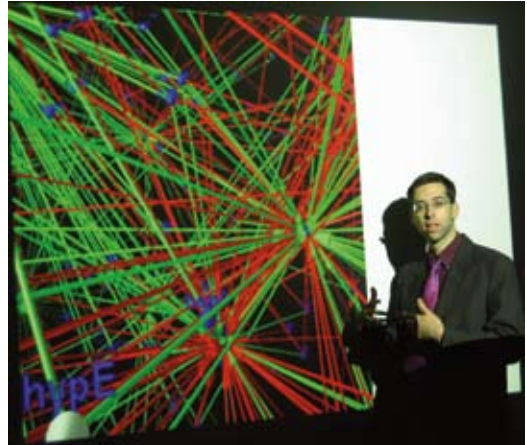


# Massachusetts Institute of Technology Integrates Cancer Research in the Lab and Classroom with MathWorks Tools

Diagnosing cancer in its earliest stages can greatly improve a patient's chances of survival. Ovarian cancer, for example, is often identified only after it has progressed to stage three or four. For patients who are diagnosed with the disease in stage one or two, the odds of surviving in five years are increased from less than 50% to about 95%.

Researchers and students at the Massachusetts Institute of Technology (MIT) are exploring methods to diagnose cancer in earlier stages by examining blood proteins. Using MathWorks tools, these researchers are identifying concentrations of proteins and protein interactions present only in cancer patients to enable early cancer detection. Students use MathWorks tools to learn from and contribute to the research group's efforts while gaining the knowledge and experience to drive future biomedical advances.

"In bioinformatics, research conducted two years ago is considered old. With MathWorks tools, we can engage students in leading-edge research that our group is doing today," says Dr. Gil Alterovitz, an NIH Biomedical Informatics Fellow in the MIT/Harvard Division of Health Sciences and Technology. "MathWorks tools enable the research group and the students—including biology majors and engineers—to focus on research and spend less time programming."



Dr. Alterovitz presenting on biomolecular networks.

## THE CHALLENGE

To improve diagnostic techniques for cancer by identifying proteins and analyzing their interactions

## THE SOLUTION

Use MathWorks tools to enable students and researchers to analyze mass spectrometry data, model complex protein interactions, and visualize results

## THE RESULTS

- Education integrated with research
- Computation time shortened by an order of magnitude
- Research grant won

## THE CHALLENGE

To better identify proteins that may signal the presence of cancer, researchers at MIT and Harvard Medical School, including Alterovitz, Marco F. Ramoni, and Isaac S. Kohane, sought to combine mass spectrometry (MS) results with knowledge of how proteins interact. MS data includes characteristic peaks and valleys that can be analyzed to distinguish molecular compounds in a sample. The researchers needed tools to process this data and to build a sophisticated model to represent protein interactions.

"We had to analyze mass spectrometry data that included millions of data points," explains Alterovitz. "We also needed to model a network of interacting biological molecules, perform statistical calculations, as well as other analysis on the properties of this network, and combine these with the mass spectrometry results."

In parallel with this research, Alterovitz initiated and directed a new course called Bioinformatics and Proteomics: an Engineering Problem-Solving Based Approach. Upper-level undergraduate students as well as first- and second-year graduate students attended the class.

Alterovitz wanted to standardize the course on a set of tools that enabled the students to benefit from ongoing research, yet would be easy to learn.

“Since we had schedule constraints, we did not want to waste time teaching the students a new language,” Alterovitz explains. “We needed a tool that the majority of students were already familiar with, and one that could be learned easily by both biologists and engineers.”

## THE SOLUTION

Researchers at MIT are using MathWorks tools to advance bioinformatics and proteomics. MIT students are using the same tools to gain hands-on experience in these fields.

### In the Lab

Alterovitz and his research group used MATLAB® to develop algorithms for analyzing the MS data and to model the protein interactivity network, which consisted of more than 20,000 nodes and 100,000 edges. Each network node represented a mass associated with a protein, and each edge represented an interaction between nodes.

The researchers also used MATLAB to visualize data, plot results, and access databases shared with other biomedical researchers.



New methods for transforming complex biomolecular networks (left) into abstract representations (right) are facilitating the discovery and characterization of their inherent components.

Because MS data resembles the series of peaks and valleys in sound or voice data, researchers can apply signal processing techniques to process the data. MIT researchers used Signal Processing Toolbox™ to process this MS data and applied filters to eliminate noise and irrelevant data, enabling them to concentrate on a more manageable data set.

Bioinformatics Toolbox™ enabled the team to quickly obtain information about proteins from a variety of Internet resources. The team used Bioinformatics Toolbox to calculate molecular weights, obtain amino acid sequences as well as other properties of specific proteins, and to download as well as parse information into data structures accessible by MATLAB.

MIT researchers used Statistics Toolbox™ to calculate network properties, including connectivity and power law distributions. They used models for calculating the number of proteins in a sample using Statistics Toolbox to simplify curve fitting and generate negative binomial, gamma, and exponential distributions.

The group's research involved millions of MS data points from hundreds of patients. However, because each patient's data was independent, the task of processing the information was ideal for parallelization. Using Parallel Computing Toolbox™ and MATLAB Distributed Computing Server™, the group executed their MATLAB algorithms concurrently on a large cluster of computers.



Using MATLAB® to analyze and generate biological networks for interactive research via 3-D stereoscopic glasses, voice recognition, and head tracking.

The group analyzed each patient's MS data independently on a different processor. Alterovitz explains, "In addition to significantly reducing computation time, Parallel Computing Toolbox enabled us to program this approach quickly. Instead of learning distributed programming, we used our existing MATLAB code, and made it parallel using Parallel Computing Toolbox."

The team also used a distributed approach to speed the calculation of network properties and statistics by dividing the network into chunks and running the tasks in parallel.

### **In the Classroom**

For the bioinformatics and proteomics course, Alterovitz and his fellow course instructors chose MATLAB for its ease of use, interoperability with other tools, and ability to present concepts at increasing levels of abstraction.

“About 90 percent of the class had already used MATLAB,” says Alterovitz. “Everyone began using MATLAB immediately—even those with no prior experience—because you do not need to know how to program in order to use it.”

In addition, MATLAB provided the students with an easy way to access and learn from leading research conducted at MIT and Harvard.

The course’s teaching approach was based on elaboration theory. It involved using a limited set of concepts and examples, and gradually adding complexity. Alterovitz explains, “MATLAB intrinsically supports different levels of complexity, through various levels of abstraction. In the beginning, students run the code and visualize results. Later, they can explore, update, and even integrate the code with other programming languages to add more detail.”

The coursework also mirrored this approach across biological levels. The students first used MathWorks tools to analyze fundamental DNA sequence information. They then progressed to more complex expression data, proteins, and eventually interactions between proteins and other molecules using a network model.

## APPLICATION AREAS

- Academia
- Algorithm development
- Biotechnology, pharmaceutical, and medical
- Data analysis
- Distributed computing

## PRODUCTS USED

- MATLAB®
- Signal Processing Toolbox™
- Bioinformatics Toolbox™
- Statistics Toolbox™
- Parallel Computing Toolbox™
- MATLAB Distributed Computing Server™

“

Researchers are typically interested in results, not programming. MATLAB enables us to think at a higher level of abstraction and spend less time developing, debugging, testing, and creating graphs. As a result, we get research results much faster.”

”

**Dr. Gil Alterovitz, Massachusetts Institute of Technology and Harvard University**

## THE RESULTS

### ▪ Education integrated with research.

“With MATLAB, we can provide the students with the latest code and research results from my group and other groups,” notes Alterovitz. “With the experience they gain, students can assist the research group and contribute to our efforts.”

### ▪ Computation time shortened by an order of magnitude.

“Using a distributed approach with MATLAB code, we ran our analysis on a computer cluster and reduced computation time by an order of magnitude—from about a week to much less than a day. That was crucial because we were facing a conference deadline, and the results played a critical role in our work getting accepted,” says Alterovitz.

### ▪ Research grant won.

“After completing the course, a biology student worked in my research group the following semester and won an MIT Undergraduate Research Opportunities Program grant,” says Alterovitz. “With MATLAB, he became very productive quickly and got results in time to apply for that grant; this would not have been possible otherwise.”

**To learn more about the Massachusetts Institute of Technology, visit [www.mit.edu](http://www.mit.edu)**