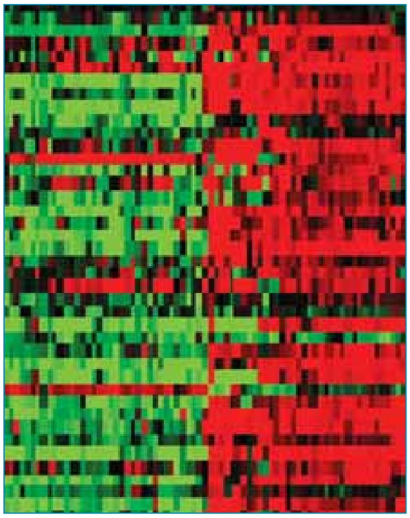
"Figuring out which proteins are secreted into the blood is like searching for a needle in a big, big haystack," says Ying Xu, PhD. "This [algorithm] sorts through all that hay."

mation. Now, scientists have developed an algorithm that sorts through the multitude, expediting the search for blood-based cancer biomarkers.

"Figuring out which proteins are secreted into the blood is like searching for a needle in a big, big haystack," says **Ying Xu**, **PhD**, professor of bioinformatics and computational biology at the University of Georgia. "This [algorithm] sorts through all that hay."

To develop their algorithm, Xu and his colleagues began by scouring the literature for all proteins known to be secreted into the blood, regardless of their origins. They then analyzed the amino-acid sequences of these proteins to identify common features, such as signal peptides, transmembrane domains, solubility, and secondary structure. They discovered 18 features that were powerful predictors of blood secretion, and used them to train a computerized classifier.



This microarray shows genes that differ in regulation between cancerous and non-cancerous lung tissue. Ying Xu's classifier can predict which of the proteins made by these genes may be useful as blood-based biomarkers. Courtesy of Ying Xu.

When the researchers applied the classifier to other data sets, it could distinguish proteins secreted into the blood from all other proteins in the blood with more than 80 percent accuracy. The results appear in the October 2008 issue of *Bioinformatics*.

Xu and his colleagues are now using microarrays to identify differences in gene expression levels between cancerous and non-cancerous stomach tissue. Using their classifier, they can then sift through the data to zero in on genes that produce proteins that are most likely to be secreted into the blood, followed by validation with mass spectrometry.

"We've already identified proteins that are elevated during different stages of stomach cancer," Xu says. "Typically, in order to find out what stage it's in, you'd have to actually cut the patients open and do a biopsy. Our markers could be the first markers to provide information about cancer stage."

By applying his biomarker discovery pipeline to a range of cancers, Xu ultimately hopes to identify general biomarkers that apply to any cancer. He envisions doctors detecting various cancers at early stages with a simple blood test.

Bo Huang, **PhD**, a post-doctoral fellow at Vanderbilt University, hopes to use Xu's classifier to find biomarkers for breast cancer. "These results provide a powerful method to discover potential biomarkers, not only for cancers but also for many other diseases," Huang says.

—By Lizzie Buchen

Blurring Data for Privacy and Usefulness

Hospitals with research agendas share a common problem: how to use medical records for research while protecting patient privacy. One approach—the data-protection equivalent of blurring the face of an anonymous source on television—has now been tested using real-world data. The results, which show promise for protecting privacy without rendering the data set useless, appear in the September/October 2008 issue of the *Journal of the American Medical Informatics Association*.

"It's not a theoretical problem," says **Khaled El Emam**, **PhD**, associate professor at the University of Ottawa and Canada Research Chair in electronic health information, who collaborated with **Fida Kamal Dankar**, **PhD**, on the paper. "We're trying to protect privacy, but we need the tools."

Just as the nightly news renders the faces of anonymous sources unrecognizable, the approach known as k-anonymity blurs distinctive variables to reduce the risk that someone could trace patients with distinctive characteristics. For example, the approach might cut birthdates down to birth years. And easily identifiable outliers—the octogenarian in a college town, the teenager in a retirement community—are omitted. The remaining information contains at least k data points that look identical, where $1/k$ is deemed an acceptable level of risk.

> “I’ve given Little b the power to reason about biological objects,” Aneil Mallavarapu says.

That works in theory, but the actual risk depends on the type of data set and what an intruder wants from it. A prosecutor digging up dirt on a defendant would try to re-identify a specific person in the database. A journalist trying to discredit an organization’s data-security procedures would also only need to re-identify one person, but it wouldn’t matter who. El Emam set out to test whether k-anonymity works in both circumstances. His findings: k-anonymity correctly predicts the risk of re-identifying one specific individual with minimal harm to the value of the database (the prosecutor example). But using k-anonymity to protect against re-identifying an arbitrary person (the journalism example) is unnecessarily strict and compromises the research quality of the data.

Since researchers choose *k* based on statistical theory, El Emam suggests data custodians run test cases to verify if the *k* is sufficient, or if it’s overprotective, as in the journalism example, before making the data available to researchers. If needed, the number of groupings of *k* identical data points could then be adjusted to ensure that the actual risk approximates the theoretical risk of 1/*k* and, in this way, keep the risk acceptably low while preserving data.

“What is needed are the steps to turn this article into a practical tool that custodians can use in conjunction with researchers,” says **Joan Roch**, chief privacy officer for Canada Health Infoway in Montreal, Quebec.

El Emam says he plans to continue exploring actual risks in various data-security scenarios: “It’s a big problem, and we’ve solved part of it.”

—By Stephanie Pappas

Modular Modeling

Biological models can quickly become as complex as the systems they represent. And minor changes can necessitate a complete rewrite of the model. But researchers may soon snap their models together like LEGOs, using a new programming language called Little b, which uses modularity to simplify biological modeling. Eventually, the authors hope to turn Little b into an easy-to-use tool for biology labs.

“I think that as an everyday tool, it [Little b] is going to be kind of like the microscope,” says **Aneil Mallavarapu, PhD**, lead developer of Little b and a senior research scientist in systems biology at Harvard Medical School. “We’re essentially building a new kind of gel, a new type of microscope for the lab.” The work appears in the June 2008 issue of the *Journal of the Royal Society Interface*.

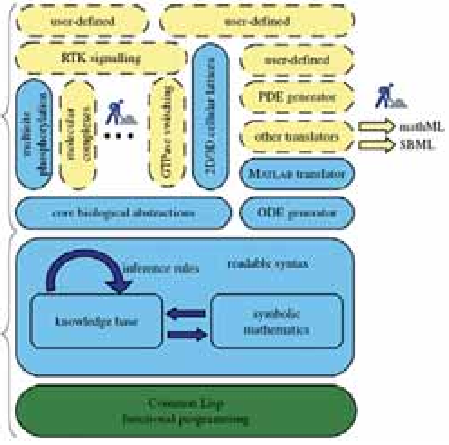
Biologists traditionally create models to describe unique systems, such as the development of fruit fly embryos or the actions of a phosphorylation cascade on gene transcription. Such computational models are usually based on lists of the system’s properties, which detail every molecular interaction in the system. This allows researchers to tailor models to the precise questions being asked, but it also constrains the model’s usefulness, because it can only probe into one area.

Little b strives to break down biological systems into modules that can be used regardless of the specific context, such as “nuclear export“ or “membrane localization.“ It then defines those parts in a mathematical language. Researchers can use Little b to put together assorted modules to describe their system; Little b then uses those symbolic modules to write out executable code that a scientist could use in a simulation program like MATLAB. “I’ve given Little b the power to reason about biological objects,” Mallavarapu says.

Mallvarapu is excited about the possible use biologists might make of Little b. He would like to see the language help uncover the complex pathways involved in diseases. He hopes that researchers will eventually build entire virtual cells or virtual plants collaboratively, increasing their ability to study their projects *in silico*.

While the idea of breaking down biological systems into modular chunks may seem logical, Little b may not arrive in the lab immediately, says **Birgit Schoeberl, PhD**, a senior director of research at Merrimack Pharmaceuticals, Inc, in Cambridge, Massachusetts. “I’m excited about the concept and what I see, but in my own experience, it isn’t straightforward,” Schoeberl says. “I think it’s not quite ready for non-developers. I hope he keeps developing it, or someone takes it on to keep working on the idea.”

—By Molly Davis □



Little b is based on a core language, which includes the Lisp language it was created in (green) and the knowledge base, symbolic mathematics and syntax modules that allow Little b to reason about biological systems. It also includes modular libraries that describe specific biological interactions, and translators that can generate code used in simulations. Blue areas exist within the current framework; yellow areas are currently under development or are envisioned for future work. Reprinted with permission from Mallavarapu, A, et al., Programming with models: modularity and abstraction provide powerful capabilities for systems biology, **Journal of the Royal Society Interface,** ***online publication, July 23, 2008.***