of computer-literate biologists to change the way biology is done," says Iain Wallace, a postdoctoral fellow in Gary Bader's group at the University of Toronto, which has been active in the development of databases of protein interactions. PubChem, which is funded by the NIH, brings to academics data that until now have been accessible primarily only to those in deep-pocketed pharmaceutical companies.

In the case of PubChem or Vanderbilt's electronic medical record database, careful statistical analyses will be required to robustly analyze these potential treasure troves of information. But rather than the algorithmic advances typically pursued in computational biology, according to Butte, "Ninety-nine percent of the work is not in software engineering or coding; it's in coming up with the right kind of question: given this data set, what question are we newly able to ask that everyone would love to know the answer to, but no one even realizes we can ask today?" Exposure is key, says Butte, "What I would love to see is a computational person going to surgical grand rounds at a hospital to figure out what the unsolved questions are, hearing about this tumor that spreads like crazy and saying, 'I can solve this problem computationally.' That would be the ideal." Unlike problems requiring clever new algorithms or massive clusters of computers, increasing exposure may be a particularly manageable challenge facing the field.

1. Denny J.C. *et al. Bioinformatics* **26**, 1205–1210 (2010).
2. Wang, Y. *et al. Nucleic Acids Res.* **38** database issue, D255–D266 (2010).

Learning to see

Why have computer scientists long endeavored to create software capable of accomplishing tasks humans can already do? In the case of biological research, one advantage of computational analysis is automation and fidelity. Whereas a trained person can look at one confocal microscope image and readily identify where a fluorescently labeled protein is localized in the cell, that person cannot hope to analyze the millions of images that can be gathered with automated technology. And even if several people were enlisted to the task, each may interpret the same image in different ways. This problem provides an apt introduction to machine learning, a technology that is finding success in biology.

In machine learning, computer programs are trained to pick out patterns, which may be predefined by human supervisors or learned by the program directly from data. Such 'unsupervised' machine-learning tasks are often the hardest, in part because there are many possibilities for the computer to consider. Notably, many machine-learning tasks in disparate problem domains can be articulated using a common set of concepts. In this way, techniques developed for one problem, say mining data from text, can inspire solutions to other problems.

Robert Murphy thinks that when computer science and biology come together "inside one person's head, that is a much more efficient process."

The advance. Robert Murphy and colleagues[1,2] at Carnegie Mellon University devised machine-learning algorithms that could accurately classify whether a pattern of fluorescent staining represents localization to one subcellular organelle or to a mixture of locations. Moreover, this 'pattern unmixing' can be done in an unsupervised way, without introducing bias from a human who predefines the categories. The need for this method is supported by studies in yeast in which up to a third of all fluorescently tagged proteins appeared to localize to several places in the cell.

The key to the approach is to segment an image into objects or shapes with quantifiable features. Then a pattern of objects can be defined as the probability that certain objects are found together. The best-performing algorithm identified patterns of objects using a technique called latent Dirichlet allocation, which has been successfully used to identify patterns of words representing conceptual topics from text documents. By analogy, visual objects representing the nucleus or Golgi apparatus are 'words' in an image, and patterns of protein localization that characterize the content of an image correspond to sets of words that co-occur in documents and define the topics in the text.

What it means. "This represents the first step toward a new way of thinking about interpreting images that is generative rather than descriptive," says Murphy. Whereas a descriptive approach may take an image of a cell expressing fluorescently tagged protein and tell you that the protein is in the nucleus, a generative approach builds a model that can produce images that look like other images, and in the process of building that model (that is, determining the parameters of the model), you learn about what characterizes a pattern in a way that is meaningful across a variety of situations. For instance, a drug in a screening assay may cause a protein to partially redistribute from one subcellular location to another, but given that organelles may look different in different cell types, without Murphy's approach, if the same screen is done on a different cell type, it is difficult to know that the same process is occurring. Machine learning has been previously applied to biology, but recent increases in the data-generation capacity of technology suggest that these kinds of approaches may play a growing role in biological discovery in the future.

Does this mean that more collaboration needs to occur between biologists and computer scientists classically trained in machine learning? Not necessarily, according the Murphy. "That's been going on for a long time already. In fact, there is a group of people who are knowledgeable in many of these different domains. There are people who in general may not push the frontier of computer science, but who use state-of-the-art techniques, and in some cases do end up pushing frontiers and identifying new problems that others in the field can then solve." The role of computational biologists is to be able to straddle domains. Murphy continues: "When the field started, it often grew by adventitious 'collisions' between computer scientists and biologists—over lunch, at a faculty meeting. That is a very inefficient way of moving forward. When those collisions can happen inside one person's head, that is a much more efficient process."

1. Coelho, L.P. *et al. Bioinformatics* **26**, i7–i12 (2010).
2. Peng, T. *et al. Proc. Natl. Acad. Sci. USA* **107**, 2944–2949 (2010).

CompBio 2.0

Businesses and broad segments of society have recently embraced decentralized mechanisms of information processing based on interactions among large groups of people. This advance in computing has not relied on new algorithms or clever data structures in the traditional sense