





Electronic access offers the promise that computers might rapidly process and integrate this wealth of information. But the information is recorded in natural language and pictures, which are hard for computers to make sense of. General-purpose text mining tools make a stab at it. But despite substantial progress during the past half century, they are far from giving computers the ability to “read” and understand language in any human sense. Plus, tools developed for general English don’t work well when applied to papers containing bioscience jargon.

Fortunately, computational linguists and computer scientists are teaming up with biologists and physicians to develop text-mining tools for biomedicine. “There’s been a huge expansion of the field in the past six or seven years,” Ananiadou says, including a flurry of papers, competitions, and conference sessions.

Researchers have developed a range of approaches. Some rely on minimal language processing, such as statistical algorithms that look at word counts. Others, by contrast, dig deeper to discern basic language structure and meaning (such as identifying noun phrases or genes) or even reveal the complete grammatical structure of millions of sentences. The latter approach is the most sophisticated and (if perfected) promises to deliver the most precise and comprehensive information, but lower level approaches can deliver a big payoff with much less complexity. Besides mining text, other researchers are working on an arguably more difficult problem for a computer—mining images and diagrams.

The potential applications are as wide-ranging as the biomedical literature itself. Researchers are not simply retrieving and repackaging what is already known, but are also deriving new knowledge by discovering connections that were previously unnoticed. Systems can already generate novel hypotheses by connecting missing links in the literature; predict unknown features of genes and proteins; help researchers make sense of microarray data; extract information to fill biological databases; build large networks of protein, gene, molecule, and disease interactions; evaluate the literature-wide evidence for scientific facts; and trace the evolution of scientific ideas.

Someday, text-mining may even make connections that bridge entire disciplines—from physics to statistics to biology, for example, says **Andrey Rzhetsky, PhD**, professor of medicine and human genetics at the University of Chicago. “You may be able to discover connections between ideas that are far, far away in the knowledge universe,” he says.

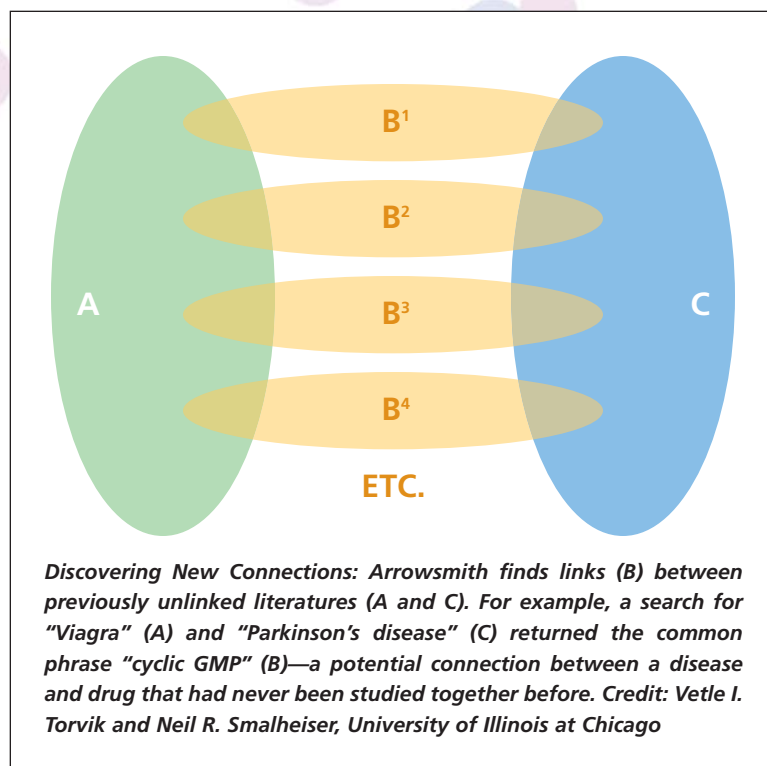
## Word Hopping

### Connecting the Missing Links

In the 1980s, **Don Swanson, PhD**, now professor emeritus at the University of Chicago, made an early successful attempt to generate novel hypotheses by mining the biomedical literature. He was able to identify indirect connections between therapies and diseases that had never been explicitly linked in the literature. For example, he tied fish oil to Raynaud's disease, and magnesium to migraines. Both treatments were later tested and proven effective.

"One experimental paper may report explicitly that A influences B; another paper, published in some other journal at some other time, may report that B influences C. The inference 'A influences C' will represent an implicit assertion that may be novel, non-trivial and worthy of investigation," explains **Neil R. Smalheiser, MD, PhD**, assistant professor in psychiatry at the University of Illinois at Chicago. He teamed up with Swanson in the 1990s to automate this strategy, creating an online tool called Arrowsmith, which has since been updated and expanded.

Arrowsmith has about 1200 unique users per month. And even though it only parses the titles of papers, Smalheiser says it has already helped researchers formulate new experiments. Among the documented successes, **John Goudreau, DO, PhD**, an associate professor of neurology and toxicology/pharmacology at Michigan State University used Arrowsmith to link Parkinson's disease and Viagra (which had never been studied together): Viagra increases cyclic GMP levels in cells, and cyclic GMP are neuroprotective in several model systems for Parkinson's. He subsequently received a grant from Pfizer to study the association.



### Arrowsmith: Word Matching

How it works: Arrowsmith performs separate PubMed searches for user-entered "A" and "C" terms and seeks common words or phrases ("B-terms") in the retrieved titles (excluding common English words such as "the" and "patient"). Using filters, the B-term search can be limited to certain categories—for example, diseases. ([http://arrowsmith.psych.uic.edu/arrowsmith\\_uic/index.html](http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html))

## Word Profiles

### Classifying Genes and Proteins

Anyone who has browsed books at Amazon.com has seen some basic text mining in action; for example, books are tagged with “Statistically Improbable Phrases”—terms that occur significantly more frequently in a particular book relative to other books—to give customers one snapshot of a book’s essence. A similar strategy can be applied to the medical literature to capture the essence of a gene or protein.

For example, a team led by **Hagit Shatkay, PhD**, an associate professor in the School of Computing at Queen’s University in Ontario created a text-based tool to predict where proteins localize in the cell. The idea is that the biomedical literature available for a protein can give clues about its cellular location even before that protein has been localized experimentally. For example, because mitochondrial proteins and nucleus proteins play different roles in the cell, research publications describe them using different

terms, Shatkay says. Their program finds terms that occur significantly more frequently in abstracts associated with proteins of a particular location compared with other locations—for example ‘bind’, ‘dna’, ‘control’, ‘histone’, and ‘transcript’ for nucleus proteins.

They combined this text-based tool with MultiLoc, a tool that predicts protein localization based on sequence data, which was created at the University of Tübingen in Germany (by a team of scientists led by **Oliver Kohlbacher, PhD**, professor for simulation of biological systems). “MultiLoc, as far as I know, was the most extensive and accurate system at the point where we joined forces,” Shatkay says. “The question was could we use text to make it even better?”

Indeed, the integrated tool, SherLoc, gave significantly better predictions of protein localization than MultiLoc alone. Across all organelles, average accuracy for MultiLoc was 74.6 percent and for SherLoc was 85.1 percent (as estimated by cross-validation). “We show that by using text, you can really get an improvement,” Shatkay says.

A similar text-based strategy can also be applied to help researchers interpret microarray data, Shatkay says. When a biologist identifies a cluster of co-expressed genes, she can then predict whether they share biological function based on the similarity of their literature profiles. “The advantage of doing it in document space is it gives you some idea of semantics,” Shatkay says. If genes cluster based on shared function, the resulting word profile will betray the function (for example, with informative terms such as “fatty acid metabolism”). When clusters form for reasons besides function—such as shared experimental methods—this will be similarly transparent. It’s been almost a decade since she and others—including **Steven Edwards, PhD**, **Mark Boguski, MD, PhD**, and **John Wilbur, MD, PhD**, then all at National Center for Biotechnology Information

### SherLoc: Classifying Genes and Proteins

How it works: SherLoc, (<http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc/>) combines a sequence-based tool (MultiLoc) and a text-based tool. The text-based tool trains a machine-learning algorithm on abstracts associated with already localized proteins. The program reduces abstracts to a “bag of words”—a list of all words and all two-term phrases (consecutive pairs of words) and their frequencies, excluding common words like “the.” Then it finds terms that appear significantly more often in abstracts associated with proteins of a particular location and it assigns weights to these terms based on their importance in classification. Once trained, the resulting algorithm can be applied to the literature associated with a new protein to predict its location.

(NCBI)—pioneered this strategy, and other scientists are now rediscovering it, she says.

Shatkey's approach involves little natural language processing; it doesn't identify 'this is a gene' or 'this is a noun', for example. But the simplicity is what makes the work elegant. "There are some text-related problems that are relatively easy to solve," Shatkey says. "And the question is, if we solve these problems, can we get anything out of it or do we need to solve the really hard problems before we can get any leverage from text?"

"It won't be as clean, it won't be as nice as natural language processing, but it's really readily available. It's low-hanging fruit," she says.

"There are some text-related problems that are relatively easy to solve," says Hagit Shatkey. "And the question is, if we solve these problems, can we get anything out of it or do we need to solve the really hard problems before we can get any leverage from text?"

## Toward Language Teaching Computers To Read Biology

While statistical approaches yield a big payoff for less effort, researchers in natural language processing are after the holy grail of text mining—getting computers to understand language in some way.

If you just use machine learning and count bags of words while ignoring linguistic structure and meaning, "there's stuff that's just going to stay out of reach," says **Kevin Cohen**, lead artificial intelligence engineer at The MITRE Corporation and biomedical text mining group lead at the University of Colorado School of Medicine.

But tackling natural language is enormously difficult. "There's this sense, this assumption, that it should be easy. You can talk and understand things and read things really easily. But of course your whole brain is designed for that," says **Alex Morgan, MS**, a doctoral student in biomedical informatics at Stanford University. "And you think that things like analytical chemistry and scheduling of flights are really complicated problems, but those are trivial computer problems [compared with natural language processing]."

To incorporate language, text-mining researchers use extensive lexical resources (including word lists, thesauri, and ontologies) to look up word variants and meanings; manually created rules about grammar and language; and machine-learning algorithms trained on collections of text marked up with linguistic information (annotated corpora).

In the world of news and journalism, considerable progress has been made on two key language-based tasks: identifying simple entities such as places, organizations, and people; and extracting simple facts such as "Company A took over Company B." Researchers have achieved near human proficiency on the first task, evaluated with F-measures—a quantity that combines precision (getting it right) and recall (not miss-

ing anything)—above 95 percent; and reasonable performance on the second task, with F-measures of 70 to 80 percent. But when off-the-shelf systems were applied to biology, they did poorly. Having been trained on text from the news world, such as The Wall Street Journal corpus, they were ill-equipped to tackle biomedical journal articles written by scientists and containing considerable jargon and nonstandard grammar.

Fortunately, in the past decade, several key events have advanced natural language processing in the biomedical domain. First, researchers in Japan—led by **Junichi Tsujii, PhD**, professor of computer science at the University of Tokyo and professor of text mining at the University of Manchester in the United Kingdom—created the GENIA corpus, a collection of

state of the art with respect to text mining for biology? And, if we can do 90-plus percent accuracy on newswire, why don't we get that performance in biology?" says **Lynette Hirschman, PhD**, director of biomedical informatics at The MITRE Corporation. There was also a need for a standard way to assess text-mining tools and a need to assess them on datasets other than the ones that were used to train them. Results on researchers' private datasets were all over the map, says Hirschman. The BioCreAtIvE competition was intended to fix that problem.

As described below, BioCreAtIvE has also addressed key challenges in bio-text mining—including promoting the development of tools to find gene and protein mentions in text and extracting basic facts, such as protein-protein interactions.

"There's this sense, this assumption, that it should be easy [for computers to read natural language]. You can talk and understand things and read things really easily. But of course your whole brain is designed for that," says Alex Morgan.

PubMed abstracts annotated with both linguistic and biological information. Corpus-based techniques had revolutionized natural language processing, Tsujii says. "I thought I should apply a similar approach to bio-text mining."

"Then we made that corpus available to all the researchers in the world," he says. "And I think that contributed quite a lot to the progress of bio-text mining."

Another driving force was the creation of a series of challenge evaluations (competitions) for text mining in biology called BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology), which started in 2003 and is run by The MITRE Corporation and the Spanish National Cancer Research Center (CNIO). Challenge evaluations can help drive a field forward by creating resources, building a community of researchers, and providing standards for assessment.

"At the time, I found myself asking: What's the

## The Name Game Tagging Genes and Proteins

One of the most basic tasks in natural language processing is to recognize important entities in running text. In biology, this means identifying genes, gene products, diseases, drugs, and cells and linking them to a unique identifier (such as an EntrezGene or SwissProt ID). If this foundational task is done poorly, the accuracy of higher-level tasks suffers.

"It's a very pragmatic problem, but it's very hard in the biomedical domain. It's surprisingly much easier

## Sentence Slicing and Dicing

### Mining for Relationships

in economics or business or news, because categories are better defined and less overlapping,” Rzhetsky says. “But here it’s essentially a mess.”

Gene and protein names present a particular challenge. Historically, scientists have used whimsical names that are not readily distinguishable as genes, for example: cheap date, heartless pinhead, and Indy (short for “I’m not dead yet”). A gene or protein may also have multiple name variants, for example, S-receptor kinase with and without the hyphen; or nuclear factor kappa B and NFkB. Further, it may be hard to distinguish between a gene and its gene product; for example, ‘p53’ could refer to a gene, protein, or mRNA. Gene and protein names may also be shared across species. “Those things never happen in the newswire domain. Bill Clinton is always Bill Clinton,” Tsujii says.

The bio-text mining community has focused considerable attention on this problem in the past five years. Several named entity recognition tools for biology are publicly available, such as those provided by the United Kingdom’s National Centre for Text Mining (<http://www.nactem.ac.uk/>), a centre created to provide text-mining tools and services for biologists. Existing tools draw on dictionary look-up (matching strings in text with lists of names); manually constructed rules, such as ‘any word ending in ase is a protein’ or ‘any phrase containing the word receptor is a protein’; and machine-learning techniques.

The top systems in BioCreAtIvE—which all include machine-learning components—achieve F-measures of 80 to 90 percent for finding gene mentions and normalizing these genes to a unique ID, which represents the state of the art for this task.

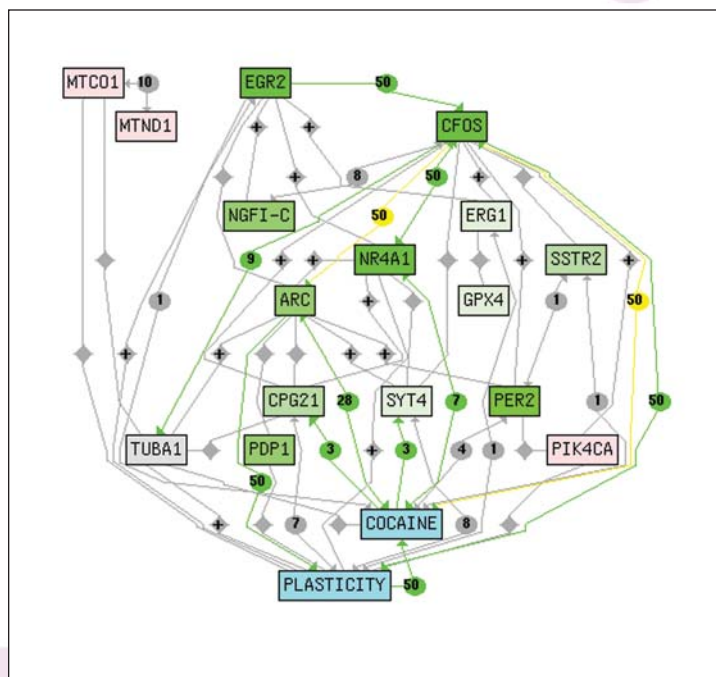
What’s promising is that the best systems in BioCreAtIvE 2006-2007 surpassed those in 2004, Hirschman says, and progress should continue.

The next step is to extract simple facts, such as protein-protein interactions, gene-disease relationships, and drug-gene relationships. These facts can be used to fill biological databases or to reconstruct biological pathways.

Fact extraction systems in biology use various degrees of “parsing”—teasing out a sentence’s grammatical structure. Early systems used no parsing, but simply inferred interaction when two proteins (or other entities) appeared in the same sentence (co-occurrence). Later systems used shallow parsing—identifying noun and verb phrases—to find telltale patterns such as two noun phrases around the verbs “phosphorylate,” “bind,” or “activate.” The latest trend is

### Chilibot: Shallow Parsing

**How it Works:** The user enters gene names and other keywords (such as addiction or nicotine). For each possible pair of terms, Chilibot (<http://www.chilibot.net>) queries PubMed to retrieve abstracts and then sentences where the pair co-occurs. The system does a shallow parse of each sentence and, based on the presence of verbs such as “activate,” “enhance,” “reduce,” and “suppress,” infers a broad relationship for each pair—stimulatory, inhibitory, or neutral. Then Chilibot presents all the pairs and relationships on a graph, with colors to represent the relationship type—red for inhibition, green for stimulation, and yellow for unresolved. From the graph, the user can jump back to the sentence that generated the relationship, and, from there, to the PubMed abstract.



*Building Networks: The Chilobot program mines PubMed abstracts for broad relationships (stimulatory, inhibitory, or neutral) between genes, proteins, drugs, and biological processes and presents them graphically. Here, Chilobot summarizes how a group of genes relate to each other and to the biological concepts of plasticity and cocaine. Credit: Hao Chen, University of Tennessee Health Science Center*

to use deep parsing—specifying the full grammatical structure of a sentence—to unravel nested and complex relationships and deal with more complex grammar (such as passive voice).

A widely used fact-extraction system that employs shallow parsing is Chilobot, short for “chip literature robot.” The program constructs relationship networks among biological concepts, genes, proteins, and drugs, and presents them in graphical form.

“You provide a list of terms and then you retrieve a graph of highly summarized relationships between the terms,” says **Hao Chen, PhD**, assistant professor of

But Chilobot probably is one of the most user-accessible interfaces of this technology on the web,” he says.

Shallow parsing has limitations, though. “Basically, you get what you pay for,” says Ananiadou. Deeper parsing delivers more precision and handles complex, nested chains of interactions. For example, the sentence “Phosphorylated Cbl coprecipitated with CrkL, which was constitutively associated with C3G” involves several nested relationships that can only be correctly mapped out with deep parsing. “If you want to work on systems biology, with pathways, you need to go to a much deeper level,” Ananiadou says. “So

“Anything that can synthesize all the literature is presumably better than something that only looks at a little bit,”  
Alex Morgan says.

pharmacology at the University of Tennessee Health Science Center. Chen, a neurobiologist, wrote the program to help him interpret microarray data. It will show how a list of co-expressed genes connects with each other and with a biological process, such as addiction.

Chilobot has been used in the design, interpretation, and validation phases of experiments, Chen says. “There are many programs with a similar function.

this is where the community is moving now.”

Not only are researchers trying to achieve depth, they are also trying to achieve breadth. Some groups have actually parsed the whole of MEDLINE and beyond. “Anything that can synthesize all the literature is presumably better than something that only looks at a little bit,” Morgan says.

For example, Tsujii’s lab developed a deep parsing tool called Enju (<http://www.tsujii.is.s.u-tokyo>).

Open access publishers—such as PubMed Central and PLoS—have unlocked a critical door for bio-text miners by providing full-text articles in a computer-friendly format.

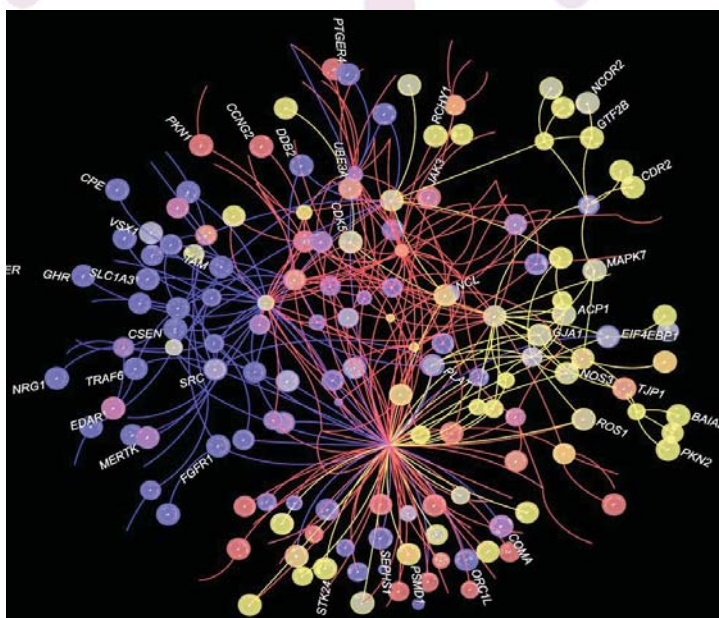
ac.jp/enju/index.html), named after a Chinese tree of wisdom (since it generates syntax trees). Enju is a cutting-edge system, one of the few text-mining programs in biology that does full parsing, Tsujii says. Tsujii's team used Enju to parse all 70 million sentences in MEDLINE in about eight days (using a 350 PC cluster). From there, they extracted all the biomedical events (such as protein-protein interactions) reported in MEDLINE; these results form the basis of an intelligent search tool called MEDIE. "Our next goal is to map all the events reported in MEDLINE to some

kind of complex network," Tsujii says.

Another cutting-edge program that employs literature-wide deep parsing is GeneWays (<http://geneways.genomecenter.columbia.edu/>), developed at Columbia University. GeneWays extracts knowledge on biological relationships in signal transduction pathways and puts these facts into a database that biologists can download. The latest run (which took about 3 months) parsed about one-third of a million full-text articles from 100 peer-reviewed journals as well as all of PubMed, and generated about 8 million redundant

## MEDIE and GeneWays: Deep Parsing

**MEDIE:** MEDIE uses the deep-parsing results from Enju to index MEDLINE with biomedical relationships and events. Using MEDIE, "biologists can retrieve all the papers in which some specific protein activates some specific biological process," Tsujii says. For example, a biologist can search for "What does p53 activate?" or "What causes cancer?" MEDIE is provided jointly by the United Kingdom National Centre for Text Mining and the University of Tokyo, <http://www-tsujii.is.s.u-tokyo.ac.jp/medie/>. **GeneWays:** GeneWays (<http://geneways.genomecenter.columbia.edu/>) employs a deep parsing tool called GENIES to extract knowledge on about 500 different types of binary relationships between genes, gene products, small molecules, diseases, and drugs in signal transduction pathways. GeneWays stores these facts in a downloadable database that biologists can use to generate large graphical networks and help them interpret experimental data.



**Pulling Out Pathways:** The GeneWays program parses the biomedical literature and returns millions of published relationships between genes, gene products, small molecules, diseases, and drugs. Here, researchers mapped the relationships between genes believed to be involved in autism (blue), bipolar disorder (yellow), and schizophrenia (red) to look for genetic overlaps between the three diseases. Credit: Ivan Iossifov, Columbia University, and Andrey Rzhetsky, University of Chicago

and 4 million unique facts, says Rzhetsky (who helped develop GeneWays). In addition to helping researchers interpret experimental data, Geneways can be used to trace the evolution of ideas in the scientific literature or help derive consensus from conflicting statements in the literature. For example, Rzhetsky's team is working on an algorithm that evaluates the weight of evidence supporting or contradicting a particular fact and generates a probability that the fact is true. "You can try to reconstruct truth," Rzhetsky says.

The performance of state-of-the-art fact extraction systems in biology is unknown, but—on a fact-by-fact basis—it may be low. The top systems on a protein-protein interaction task in BioCreAtIvE 2006-2007 achieved F-measures of only about 35%. Though fact extraction remains largely a research problem, tools in use today are benefiting users. These systems exploit redundancy (looking at multiple mentions of a fact) to increase recall and accuracy.

If you combine information extracted from 50,000 paragraphs, "you're going to get the right answer," Morgan says. "Eventually you're going to have seen that fact so many times that it must be true and all the ones that are wrong disappear, because they're random instances that don't happen that often."

## Full Text Ahead Advancing Biology and Medicine

The bio-text mining community faces several key challenges. To date, tools have focused on mining abstracts, which are more readily available than full-text articles. But the bulk of information, as well as the tables and figures, are contained in full text.

Many full-text articles require a subscription for access; and even when available, they may be in formats that don't work well for text-mining applications.

"Trying to process a PDF document is a nuisance. You can convert it to plain text but it doesn't convert very well," Hirschman says.

Open access publishers—such as PubMed Central and PLoS—have unlocked a critical door for bio-text miners by providing full-text articles in a computer-friendly format. But much of the literature still remains inaccessible.

Another challenge is making tools that are useful to biologists. Systems are typically evaluated as to their recall and accuracy in handling canned problems, but usability to biologists may actually be a more important benchmark.

"We've been pushing for evaluations that will let us quantify the value of a particular system and its performance to a biologist. How much will it help you do your job?" Cohen says. "I'm happy to say that in the last couple years, for the first time, we've actually seen productive research in that area."

Despite such challenges, bio-text mining has advanced considerably in a short amount of time. Rather than scientists tracking down one journal article at a time in the library, computers are now doing the legwork—surveying millions of abstracts and hundreds of thousands of full-text articles at once and returning insights that don't exist in a single article.

"We've made enormous progress," Hirschman says. "We have a very vibrant community of researchers now. The results are getting better. We understand where we are and what resources we need." And if key challenges, such as full-text access and usability, can be met, she and others expect the field to advance rapidly.

Moreover, says Rzhetsky, "I strongly believe that text mining can speed up scientific progress." □

# Getting the Picture

## Mining Images and Diagrams

While considerable effort has gone into processing biomedical text, much less attention has been paid to processing figures. Yet figures and figure-related text (captions and text referring to figures) make up 50 percent of a typical biomedical paper, says **Robert P. Futrelle, PhD**, associate professor of computer and information science at Northeastern University.

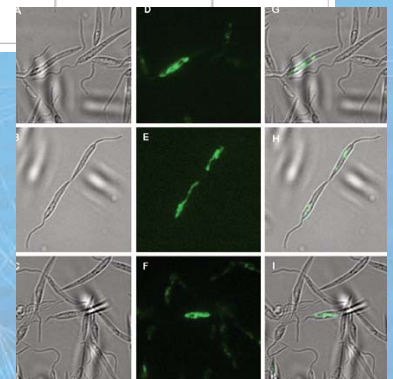
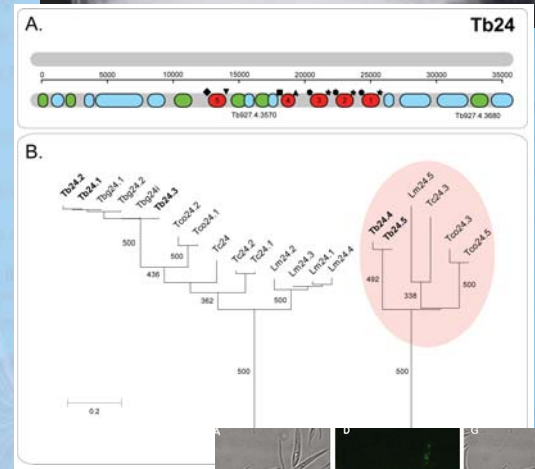
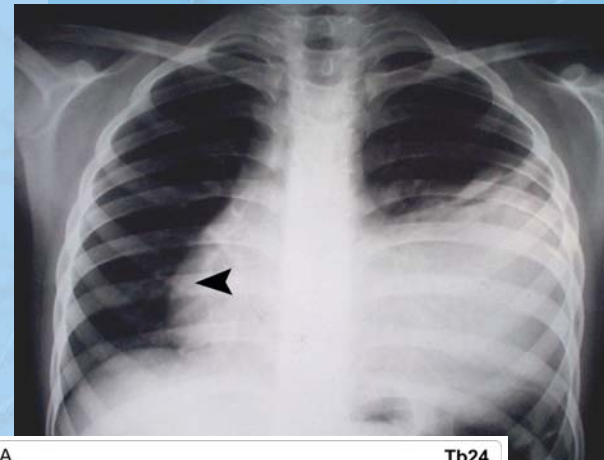
Technology for mining figures is still in its early stages, and most of the work is focused on information retrieval—the ability to search for images and diagrams the way we search for documents. As with text mining, the simplest approach is to use statistical methods in which programs look for patterns of pixels rather than patterns of words. But so far this just allows researchers to classify images in broad terms. For example, in a search of medical images, this technique might separate a chest X-ray from a CT scan. “The technology of image processing is not nearly as advanced as text processing,” says **William Hersh, MD**, professor of medical informatics and clinical epidemiology at Oregon Health & Science University.

Besides medical images, other researchers are working on classifying biological images. For example, a team led by **Robert Murphy, PhD**, professor of biological sciences, biomedical engineering, and machine learning at Carnegie Mellon University, developed SLIF (Subcellular Location Image Finder, <http://slif.cbi.cmu.edu>), a tool that divides multi-part figures into individual panels and picks out fluorescence microscope images (using a machine-learning classifier). Beyond classification, the tool also extracts facts about protein subcellular localization from image features and caption text. SLIF was used to automatically extract fluorescence microscope images from 15,000 PNAS papers and to store them in a searchable database indexed (where possible) by protein, cell type, and subcellular location.

Futrelle is trying to do even deeper processing—akin to parsing sentences—to extract meaning from diagrams (line drawings and graphs). In diagrams, the lowest level items are not words, but individual lines that have essentially no meaning on their own, he says. “If you just had the lines in a bag and pulled them out it wouldn’t mean anything; but if you put them in place, all of a sudden, ‘bingo,’ you have something,” he says.

Rather than look for nouns, verbs, and prepositions, his team looks at the lengths, positions, and connectivity of lines to detect standard pictorial expressions, such as plus signs, arrows, and error bars. “So, we have parsed diagrams. We have taken data graphs and pulled out everything—all the little tick marks and the scale lines and the data points,” he says. His lab is now redeveloping the approach in a newer programming language.

In the future, Futrelle says he hopes to build tools that perform intelligent searching for particular types of diagrams (such as a bar graph about a specific topic) and that automatically add metadata—tags that identify: “this is a gene diagram” or “this is a bar graph”—to figures in the literature. Beyond information retrieval, Futrelle’s work could also form the basis of systems that actually mine figures for new knowledge, similar to current text-mining systems.



Scientists are making progress mining information from figures such as these. Chest Xray reprinted from Marashi SM, Eghtesadi-Araghi P, Mandegar MH. A large left ventricular pseudoaneurysm in Behçet’s disease: a case report. *BMC Surg.* 2005 Jun 14;5:13. Fluorescence Microscope image reprinted from: Zamora-Veyl FB, Kroemer M, Zander D, Clos J. Stage-specific expression of the mitochondrial co-chaperonin of *Leishmania donovani*, CPN10. *Kinetoplastid Biol Dis.* 2005 Apr 29;4(1):3. Diagram example reprinted from: Jackson AP. Tandem gene arrays in *Trypanosoma brucei*: comparative phylogenomic analysis of duplicate sequence variation. *BMC Evol Biol.* 2007 Apr 4;7:54.