

# Speed reading

Scientists are struggling to make sense of the expanding scientific literature. **Corie Lok** asks whether computational tools can do the hard work for them.

In 2002, when he began to make the transition from basic cell biology to research into Alzheimer's disease, Virgil Muresan found himself all but overwhelmed by the sheer volume of literature on the disease. He and his wife, Zoia, both now at the University of Medicine and Dentistry of New Jersey in Newark, were hoping to test an idea that they had developed about the formation of the protein plaques in the brains of people with Alzheimer's disease. But, as newcomers to the field, they were finding it almost impossible to figure out whether their hypothesis was consistent with existing publications.

"It's really difficult to be up to date with so much being published," says Virgil Muresan. And it's a challenge that is increasingly facing researchers in every field. The 19 million citations and abstracts covered by the US National Library of Medicine's PubMed search engine include nearly 830,000 articles published in 2009, up from some 814,000 in 2008 and around 772,000 in 2007. That growth rate shows no signs of abating, especially as emerging countries such as China and Brazil continue to ratchet up their research.

The Muresans, however, were able to make use of Semantic Web Applications in Neuromedicine (SWAN), one of a new generation of online tools designed to help researchers zero in on the papers most relevant to their interests, uncover connections and gaps that might not otherwise be obvious, and test and generate new hypotheses.

"If you think about how much effort and money we put into just Alzheimer's disease research, it is surprising that people don't put more effort into harvesting the published knowledge," says Elizabeth Wu, SWAN's project manager.

SWAN attempts to help researchers harvest that knowledge by providing a curated, browseable online repository of hypotheses in Alzheimer's disease research. The hypothesis that the Muresans put into SWAN, for example, was that plaque formation begins when amyloid- $\beta$ , the major component of brain plaques, forms seeds in the terminal regions of cells in the brainstem that then nucleate the plaques in the other parts of the brain into which the terminals reach. SWAN provides a visual, colour-coded display of the relationships between the hypotheses, as derived from the published literature, and shows where they may agree or conflict.

The connections revealed by SWAN led the Muresans to new mouse-model experiments designed to strengthen their hypothesis. "SWAN has advanced our research, and focused it in a certain direction but also broadened it to other directions," says Virgil Muresan.

The use of computers to help researchers drink from the literature firehose dates back to the early 1960s and the first

experiments with techniques such as keyword searching. More recent efforts include the striking 'maps of science' that cluster papers together on the basis of how often they cite one another, or by similarities in the frequencies of certain keywords.

As fascinating as these maps can be, however, they don't get at the semantics of the papers — the fact that they are talking about specific entities such as genes and proteins, and making assertions about those entities (such as gene X regulates gene Y). The extraction of this kind of information is much harder to automate, because computers are notoriously poor at understanding what they are reading. Even so, informaticians and biologists are working together more and making considerable progress, says Maryann Martone, the chairwoman of the Society for Neuroscience's neuroinformatics committee. Recently, a number of companies and academic researchers have begun to create tools that are useful for scientists, using various mixtures of automated analysis and manual curation (see 'Power tools', page 418).

## Deeper meaning

The goal of these tools is to help researchers analyse and integrate the literature more efficiently than they can do through their own reading, to hone in on the most fruitful experiments to do and to make new predictions of gene functions, say, or drug side effects.

The first step towards that goal is for the text- or semantic-mining tool to recognize key terms, or entities, such as genes and proteins. For example, academic publisher Elsevier, headquartered in Amsterdam, has piloted Reflect in two recent online issues of its journal *Cell*. The technology was developed at the European Molecular Biology Laboratory in Heidelberg, Germany, and won Elsevier's Grand Challenge 2009 competition for new tools that improve the communication and use of scientific information.

Reflect automatically recognizes and highlights the names of genes, proteins and small molecules in the *Cell* articles. Users clicking on a highlighted term will see a pop-up box containing information related to that term, such as sequence data and molecular structures, along with links to the sources of the data. Reflect obtains this information from its dictionary of millions of proteins and small molecules.

Such 'entity recognition' can be done fairly accurately by many mining tools today. But other tools take on the tougher challenge of recognizing relationships between the entities. Researchers from Leiden University and Erasmus University in Rotterdam, both in the Netherlands, have developed software called Peregrine, and used it to predict an undocumented interaction between two proteins:

"Somebody staring at the data or using existing tools would never come up with this hypothesis."

— Lawrence Hunter

calpain 3, which when mutated causes a type of muscular dystrophy, and parvalbumin B, which is found mainly in skeletal muscle. Their analysis found that these proteins frequently co-occurred in the literature with other key terms. Experiments then validated that the two proteins do indeed physically interact (H. H. van Haagen *et al. PLoS One* 4, e7894; 2009).

### Development role

At the University of Colorado in Denver, bioinformatician Lawrence Hunter and his research group have developed a tool called the Hanalyzer (short for 'high-throughput analyser'), and have used it to predict the role of four genes in mouse craniofacial development. They gathered gene-expression data from three facial tissues in developing mice and generated a 'data network' showing which genes were active together at what stage of development, and in which tissues. The team also mined relevant abstracts and molecular databases for information about those genes and used this to create a 'knowledge network'.

Using both networks, the researchers homed in on a group of 20 genes that were upregulated at the same time, first in the mandible (lower jaw area) and then, about 36 hours later, in the maxilla (upper jaw). A closer look at the knowledge network suggested that these genes were involved in tongue development, because the tongue is the largest muscle group in the head and is in the mandible. Further analysis led them to four other genes that had not been previously linked to craniofacial muscle development but that were active in the same area at the same time. Subsequent experiments confirmed that these genes were also involved in tongue development (S. M. Leach *et al. PLoS Comput. Biol.* 5, e1000215; 2009).

"I don't see that there is any way that somebody staring at the data or using existing tools would have ever come up with this hypothesis," says Hunter.

Although extracting entities and the relationships between them is a common approach for literature-mining tools, it is not enough to pull out the full meaning of research papers, says Anita de Waard, a researcher of disruptive technologies at Elsevier Labs in Amsterdam. Scientific articles typically lay out a set of core claims, together with the empirical evidence that supports them, and then use those claims to argue for a conclusion or hypothesis. "Generally that's where the real, interesting science is," de Waard says.

Capturing the higher-level argument is an even more difficult task for a computer, but a small number of groups, such as the SWAN group, are trying to do so.

The SWAN website, which opened to the public in May 2009, was developed by two Boston-based groups, the Massachusetts General Hospital and the Alzheimer Research Forum, a community and news website for Alzheimer's researchers. For each hypothesis in the system, SWAN shows the factual claims that support it, plus links to the papers supporting each claim. Because claims from the various hypotheses are linked together in a network, a user can browse from one to the next and see the connections between them. The visualization tool uses a red icon to show when two claims conflict and a green icon to show when they're consistent, allowing the user to see at a glance which hypotheses are controversial and which are



In SWAN (Semantic Web Applications in Neuromedicine), hypotheses are labelled 'H', and supporting claims are marked 'C'. Relationships between the claims are indicated as 'inconsistent' (red box) or 'consistent' (green box). See SWAN at [go.nature.com/nNz93Y](http://go.nature.com/nNz93Y).

well supported by the literature (see graphics, above).

At the moment, this information is unlikely to surprise experts in Alzheimer's disease. In its current stage of development, SWAN may be more useful for newcomers trying to get up to speed on the subject. Beneficiaries could include more established scientists such as the Muresans who want to move into a different field, or researchers with a pharmaceutical or biotech company who have just been put on an Alzheimer's disease project.

### Building up

SWAN also has scalability issues. The vast majority of the hypotheses, claims and literature links in SWAN have been annotated and entered by the site's curator, Gwen Wong, with the help of authors. This curation is a painstaking process that has so far produced only 1,933 claims and 47 fully annotated hypotheses. But the intent is for these early hand-curation efforts to set a 'gold standard' for how the SWAN knowledge base should be built by the community as a whole. The SWAN developers plan to improve the user interface to encourage scientists to submit their own hypotheses, post comments and even do some of the curation themselves.

The need for some level of manual curation is common to the various literature tools, and limits their scalability. The SWAN team is working to automate parts of the curation process, such as extracting gene names. Elsewhere,

POWER TOOLS			
Tool	Description	Developed/marketed by	Available at
Knowledge Dashboard	Text mines and groups papers by concepts, looking for connections between them. Requires subscription.	Collexis Holdings/Thomson Reuters	go.nature.com/Xk27YW
ConceptWeb	Extracts 'triplets' (relationships between two concepts) from papers; clusters related triplets, looking for new connections.	Knewco	go.nature.com/BZHFLW
Arrowsmith	Identifies links between sets of seemingly unrelated articles in PubMed, looks for new inferences.	University of Illinois, Chicago; University of Chicago	go.nature.com/x2ze28
Neuroscience Information Framework	Semantic-based search of neuroscience literature. Also searches databases, other resources and an indexed set of relevant websites.	Collaborators from University of California, San Diego; California Institute of Technology; George Mason University; Yale University School of Medicine; and Washington University in St Louis	go.nature.com/kPH9Ak
MEDIE	Semantically searches MEDLINE using text-mining and natural language-processing techniques.	University of Tokyo	go.nature.com/VmFZiM
Information Hyperlinked over Proteins (iHOP)	Searches and links PubMed abstracts by gene and protein name.	Robert Hoffmann, Memorial Sloan-Kettering Cancer Center	go.nature.com/Dr8dMb
Textpresso	Text mines the literature related to a variety of model organisms.	California Institute of Technology	go.nature.com/v8mYel

de Waard and other researchers are investigating ways of automatically recognizing hypotheses — for example, by looking for specific word patterns.

For most of these tools, however, curation is unlikely to become fully automated. "Literature mining is hard to do in a way that is both high scale and high accuracy," says John Wilbanks, director of Science Commons, a data-sharing initiative in Cambridge, Massachusetts. Developers say a more likely solution, at least in the short term, is that papers will have to be curated and annotated through some combination of automated tools, professional curators and the papers' authors, who might, for example, be prevailed on to write their abstracts in a more structured machine-readable form.

### The right people

Are authors willing to add to the already arduous task of writing an article? And are authors even the best people to do this job? The journal *FEBS Letters* experimented in 2009 with structured digital abstracts to see how authors would respond and perform in shaping their own machine-readable abstracts. The results were not encouraging. Authors presented their abstracts about protein-protein interactions as structured paragraphs describing entities, the relationships between the entities and their methods using specific, simple vocabularies (for example, 'protein A interacts with protein B'). But the curators of a protein database didn't accept them, says de Waard. "Authors are not the right people to validate their own claims," she says. The community — referees, editors, curators, readers at large — is still needed.

This could be a business opportunity for the publishers, says Wilbanks: they could curate and mark up their publications for text and semantic mining and provide that as a value-added service.

"There's a lot of business out there for the publishers, but it's not the same business," says Allen Renear, associate dean for research at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. "If they keep making PDFs, that's not going to work for them. They have to get into more

of the semantic side of this."

Perhaps the largest challenge is getting scientists to use these tools. It will be up to the developers to demonstrate the benefits and make their wares easy to use.

That's going to be difficult, says Hunter. Academic informaticians are rewarded more for coming up with new algorithms, and less for making their programs usable and widely adoptable by biomedical scientists, he says. Only a few tools are being developed by companies for more widespread use.

Major issues that all technology developers will need to tackle are transparency, provenance and trust. Scientists won't trust what a computer is suggesting in terms of new connections or hypotheses if they don't know how the results were generated and what the primary sources were. "We as informaticians are going to have to take on these more user-driven and less technology-driven problems," says Hunter.

Even if researchers do start to trust the new tools, it's not clear how much of their reading they will delegate. "As reading becomes more effective," says Renear, "some people have speculated that we won't do as much because we'll get done what we need to do sooner." Or, he says, "it may be that we'll do more reading because it's more valuable. Which one is true is actually an empirical question."

Analysing articles in new ways leads to the larger question of whether the articles themselves should change in structure. If an article is to be boiled down into machine-readable bits, why bother writing whole articles in the first place? Why don't researchers just deal with statements and facts and distribute and mash them up to generate hypotheses and knowledge?

"Human commentary and insight are still extraordinarily valuable," says Martone. "Those insights don't immediately fall out of data without human ingenuity. So you need to be able to communicate that and that generally means building an argument and a set of supporting claims. These things are not going to go away any time soon." ■

**Corie Lok is Research Highlights editor for Nature, based in Cambridge, Massachusetts.**