Hypothesis generation for desiccation tolerance research
Serena Lotreck

Have you ever Googled something, and then been overwhelmed by the amount of information there is on the topic?

This scenario of being bombarded with information is referred to as *information overload*; when there is too much information on a given topic for the human brain to process and understand. While it's commonly discussed in the context of general internet searches, information overload is also a problem for scientists.

To answer research questions, scientists rely on a process called *hypothesis generation*, or coming up with potential answers to the question that we can test with experiments. Hypothesis generation typically involves reading as many scientific papers related to our research question as we can – however, this is where information overload starts to become an issue.

Google Scholar, a major search engine for academic papers, contains at least 114 million records, which is estimated to be only 88% of the total English-language articles available on the web — not to mention those in other languages! Published papers represent a major resource for contemporary scientists, but the size of humanity's science database poses a challenge to using that information.

The approach I'm testing to help scientists build better hypotheses is called *automated hypothesis generation*. Automated hypothesis generation can help scientists combine the power of their own intuition with the ability to process a larger volume of scientific information to generate hypotheses.

But how does this method work? First, we need a research question. I study the way plants react to *desiccation*. Desiccation is when plants get such little water that they almost completely dry out. Desiccation kills most plants, but some plants are *desiccation tolerant*. A desiccation tolerant plant can lose almost all its water, enter into a dormant state, and then return to its normal green state the next time water is introduced. To illustrate how automated hypothesis generation works, let's start with the question: "How do plants become desiccation tolerant?"

To find a potential answer to our question, we gather some information from previous research. Typically, this means reading scientific articles related to ideas we may have from our previous experience working on the topic. We know that the seeds of many plants are able to survive desiccation; it's why we can buy seeds for our gardens in packets. Perhaps plants repurposed the mechanism of desiccation tolerance from seeds to use in their leaves?

However, the search "seed desiccation tolerance" in Google Scholar returns 89,900 results. Additionally, plants aren't the only organisms that are desiccation tolerant. Many fungi, as well as microscopic animals like nematodes, are desiccation tolerant. What if desiccation tolerant plants share mechanisms with these other organisms? The only way to find out is to read more

papers! A human brain can only contain a certain amount of information, so we aren't able to go much beyond the first few pages of a Google Scholar search and still add more useful information to form a coherent hypothesis.

Instead, we can use automation to synthesize our search results. From a large set of papers, we use an algorithm to extract words and phrases that represent biological entities, like animals, plants or proteins, as well as the relationships between them, like whether one protein regulates another.

Once we have this information, we can visualize it like a network, where the entities are connected by relationships. We then use an algorithm to identify the pattern of connections between biological entities. The algorithm predicts what connections might plausibly exist between entities, but weren't demonstrated in the literature. For example, maybe we see that two proteins interact in nematodes to help the animal survive desiccation, and one of those proteins exists in plants, but hasn't been shown to be involved in desiccation tolerance. We could formulate a hypothesis that the protein is involved in desiccation tolerance in plants, and use our prediction algorithm to add more detail to that hypothesis by predicting connections to other proteins found in plants.

Automated hypothesis generation can help us add more detail to our hypothesis and relieve the burden of manual literature searches. However, we're not looking to replace the research – we're looking to *augment* them. An automatically generated hypothesis will need to be validated by a human, and humans will need to provide scientific context to how the hypotheses are used. Our hope is that augmentation with automated hypothesis generation can help researchers more quickly design experiments to test complex hypotheses, leading to better scientific outcomes for applications in improving agriculture for a future in a changing climate.