New algorithm efficiently finds antibiotic candidates

October 2 2018, by Byron Spice

Credit: CC0 Public Domain

If you're looking for a needle in a haystack, it's best to know what hay looks like. An international team of researchers has applied this idea to the search for new pharmaceuticals, developing a technique that reduces

the chances of simply rediscovering known compounds.

In an article published today in the journal *Nature Communications*, researchers from Carnegie Mellon University; the University of California, San Diego; and St. Petersburg State University in Russia describe a new means of searching vast repositories of compounds produced by microbes. By analyzing the mass spectra of the compounds, they were able to identify known compounds within the repository and eliminate them from further analysis, focusing instead on the unknown variants—the needles within the haystack—that might potentially be better or more efficient antibiotics, anticancer drugs or other pharmaceuticals.

In just a week, running on 100 computers, the algorithm, called Dereplicator+, sorted through a billion mass spectra in the Global Natural Products Social molecular network at UC San Diego and identified more than 5,000 promising, unknown compounds that merit further investigation, said Hosein Mohimani, assistant professor in CMU's Computational Biology Department and first author on the article.

The algorithm that powers this molecular search engine is now available for use by any investigator to study additional repositories.

In the past, mass spectrometry data repositories have been underused because it was difficult to search through them and because those efforts to date have been plagued by high rates of rediscovery of known compounds.

"It's unbelievable how many times people have rediscovered penicillin," Mohimani said.

Analyzing the compounds' mass spectra—essentially, a measurement of

the masses within a sample that has been ionized—is a relatively inexpensive way of identifying possible new pharmaceuticals. But existing techniques were largely limited to peptides, which have simple structures such as chains and loops.

"We were only looking at the tip of the iceberg," Mohimani said.

To analyze the larger number of complex compounds that have entangled structures and numerous loops and branches, the researchers developed a method for predicting how a mass spectrometer would break apart the molecules. Beginning with the weakest rings, the method simulated what would happen as the molecules came apart. Using 5,000 known compounds and their mass spectra, they trained a computer model that could then be used to predict how other compounds would break down.

Mohimani said Dereplicator+ not only can identify known compounds that don't need to be investigated further, but it can also find less common variants of the known compounds that likely would go undetected within a sample.

More information: Hosein Mohimani et al, Dereplication of microbial metabolites through database search of mass spectra, *Nature Communications* (2018). DOI: 10.1038/s41467-018-06082-8

Provided by Carnegie Mellon University