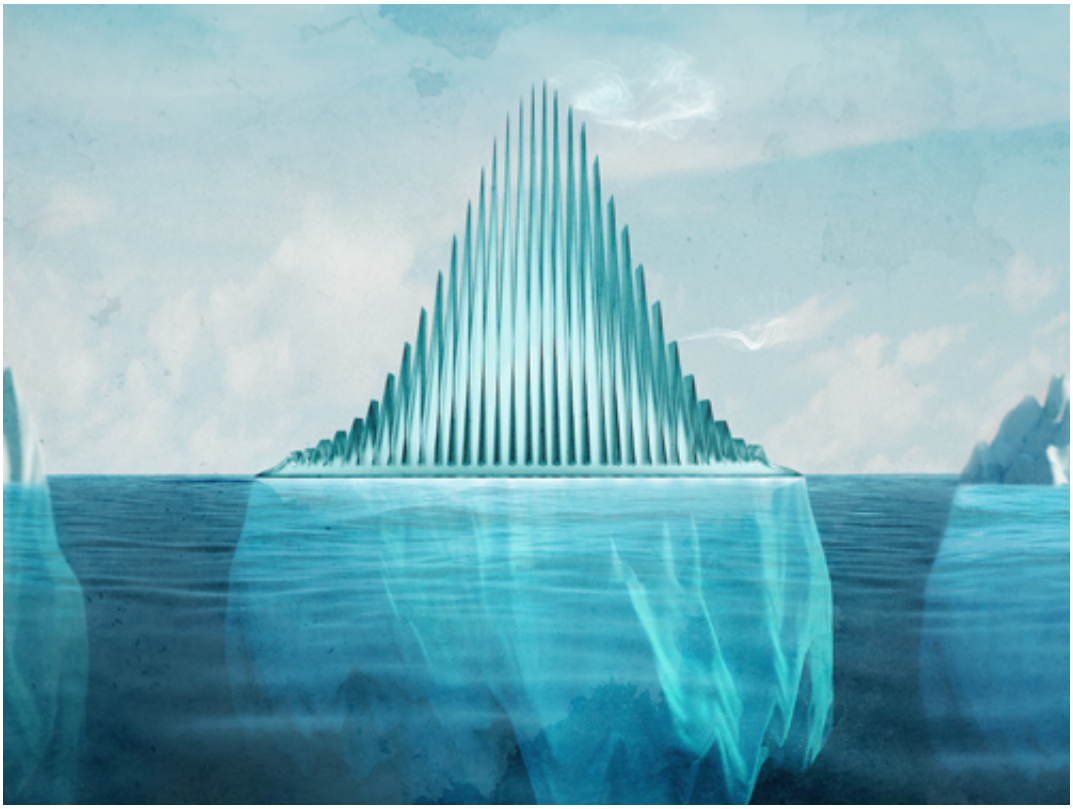
Unidentified spectra detector

June 28 2016



"Creating a sensible subset of spectra to start an in-depth analysis of unidentified spectra has been very challenging," says Juan Antonio Vizcaino of EMBL-EBI. Credit: Spencer Phillips, EMBL-EBI

A new algorithm clusters the millions of peptide mass spectra in the PRIDE Archive public database, making it easier to detect millions of consistently unidentified spectra across different datasets. Published in *Nature Methods*, the new tool is an important step towards fully

exploiting data produced in discovery proteomics experiments.

On average, almost three quarters of spectra measured in discovery proteomics experiments remain unidentified, regardless of the quality of the experiment, as they cannot be interpreted by standard sequence-based search engines. Alternative approaches to improve the rate of identification exist, but are fraught with disadvantages including ambiguous results. In today's study, researchers working on the PRIDE Archive public repository of proteomics data present a large-scale 'spectrum clustering' solution that takes advantage of the growing number of mass spectrometry (MS) datasets to systematically study millions of unidentified spectra.

"MS experiments produce huge amounts of data, but identifying meaningful sequences that could be assigned to specific biological functions can be troublesome," says Johannes Griss, formerly at EMBL-EBI in the UK and now at the Medical University of Vienna, Austria.

"Discovery proteomics is a mature technology, and it's crucial that we are able to exploit the data efficiently."

One of the challenges with these technologies is that a large proportion of the data generated can't be interpreted, as they correspond to peptides that have not yet been observed and are not available in databases. Such spectra could correspond to peptide variants derived from individual generic variation, or to peptides containing post-translational modifications, which are essential for the biological functions of proteins.

"What we have now is an algorithm that shows us patterns, or groups of spectra, that we've consistently missed, and helps us figure out which ones are good enough to pursue," adds Johannes. "It's a valuable tool that helps us unpick what's going on in proteomics, so we can better

understand basic biological processes."

The team used the approach to recognise 9 million consistently unidentified spectra, which can make post-translational modifications and peptides containing sequence variants more discoverable. They identified three distinct sets of spectra: those that have been incorrectly identified, those that are not of high enough quality to identify properly, and those that are truly unidentified. They also combined their new approach with other methods to identify roughly 20% of the originally unidentified spectra in the public archive.

"Discovery proteomics is a mature technology, and it's crucial that we are able to exploit the data efficiently - but creating a sensible subset of spectra to start an in-depth analysis of unidentified spectra has been very challenging," says Juan Antonio Vizcaíno, who leads the Proteomics team at EMBL-EBI. "We developed a comparatively lightweight computational approach that makes it much easier to detect sequences that have been incorrectly identified, or consistently observed but not identified. These ready-to-use collections of commonly unidentified spectra are a resource for the community, so that we can all pool our efforts to find lasting solutions for proteomics research."

The new algorithm will be used to improve quality control in the PRIDE Archive. The complete spectrum clustering results are available through the PRIDE Cluster resource, which aims to simplify further investigation into unidentified spectra.

More information: Johannes Griss et al, Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets, *Nature Methods* (2016). DOI: 10.1038/nmeth.3902

Provided by European Molecular Biology Laboratory