NovoFit: A Machine Learning Based Tool to Estimate False Discovery Rate of De Novo Sequencing

Jihun Cha, Seunghyuk Choi and Eunok Paek

Department of Computer Science, Hanyang University, Seoul 04763, Republic of Korea

In mass spectrometry-based proteomics, de novo sequencing has an advantage of identifying peptides whose sequences are not known. Its utility has been low, however, even in applications like novel peptide discovery in proteogenomic studies where we expect to find peptides whose sequences are not in the reference databases. One obstacle to its broad application is that there is no well-established method for statistical validation of the results. Here, we propose a machine learning based post-processing tool, called NovoFit, that can re-score peptide spectrum matches (PSMs) using target and decoy PSMs as a training set. We generate decoy spectra by swapping the precursors of the original target spectra. Target and decoy spectra are searched and assigned to peptides by de novo sequencing tools such as PEAKS, and the search results are used as a positive and negative training set, respectively. NovoFit then calculates features that can assess peptide-spectrum match quality for the target and decoy PSMs followed by re-scoring both PSMs iteratively using random forest. To evaluate the performance of NovoFit, we used ProteomeTools synthetic peptide dataset (PXD004732). We searched 6,274,999 target and 5,737,254 decoy spectra by PEAKS and then re-scored them using NovoFit and their identifications were estimated at 1% false discovery rate (FDR) based on two different scores, respectively: ALC score (PEAKS score) and NovoFit score. The peptide level recall by using ALC score was ~17.09%. With NovoFit score, the recall has increased to ~59.54%. Precision with the use of ALC score was ~96.94% and ~69.10% with NovoFit score. The F1 score with the use of ALC score was ~0.29 and ~0.64 with NovoFit score. It is notable that recall, precision, and F1 score derived from NovoFit were similar to those using the ad hoc threshold with ALC score 50, i.e., ~62.46%, ~65.67%, and ~0.64, respectively.