[BIOINFO 2023]

NovoCert: Statistical validation of *de novo* peptide sequencing results

Shanji Zhang[1], Seunghyuk Choi[1], and Eunok Paek[1,2*]

[1] *Department of Computer Science, Hanyang University, Seoul 04763, Republic of Korea*

[2] *Institute for Artificial Intelligence Research, Hanyang University, Seoul 04763, Republic of Korea*

⋆Corresponding author: eunokpaek@hanyang.ac.kr

De novo peptide sequencing from tandem mass spectra can be useful to identify novel peptides. However, lack of statistical validation hinders its application in practice. We propose a method, called NovoCert, that utilizes both semi-supervised learning and statistical approach to validate the peptide spectrum matches (PSMs) inferred from *de novo* peptide sequencing. Our peptides of interest in the *de novo* peptide sequencing results are “novel” peptides in general, hence spectra that could be reliably identified through database search (*e*.g., Comet search against reference protein sequence database) were discarded at the beginning. Among the unidentified spectra, we obtained the PSMs by identifying spectra using a *de novo* sequencing tool PEAKS and used them as a positive training data set. Using BLASTp, we divided the PSMs into “exact group” and “additional group” based on whether they had at least one exact match in the protein sequence database or not, respectively. After that, we generated negative training data sets for each of the exact and additional group by sequencing “reverse-shifted and precursor-swap” and “reverse-shifted” spectra, respectively. To precisely discriminate between positive and negative training data sets, we used 14 proteomic features including spectral similarity and delta retention time. Using Percolator, we estimated each group at 1% false discovery rate (FDR). For the identified PSMs in the additional group, we further validated their quality by calculating empirical p-value of the spectral similarity. To evaluate NovoCert, we used the ProteomeTools synthetic peptide dataset (PXD004732). As a result, in the exact group, NovoCert identified 7,690 peptides that were not found in the database search. In the additional group, 62,267 peptides were identified at 1% FDR. Almost all PSMs (>99%) showed a significant p-value (<0.05), indicating that the identification was confident. These peptides in the additional group are assumed to have been rendered due to peptide synthesis errors resulting in chemical modifications and/or altered sequences.