DiNovo enables high-coverage and high-confidence de novo peptide sequencing via mirror proteases and deep learning

Zixuan Cao[1], Xueli Peng[1], Di Zhang[2], Piyu Zhou[1], Li Kang[3], Hao Chi[4], Ruitao Wu[2] Zhiyuan Cheng[1], Yao Zhang[3], Jiaxing Dai[3], Yanchang Li[3], Lijin Yao[2], Xinming Li[2], Yaoyu He[1], Jinghan Yang[1], Haipeng Wang[2], Ping Xu[3], Yan Fu[1*]

[1] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[2] Shandong University of Technology, Zibo, China
[3] Beijing Institute of Lifeomics, Beijing, China
[4] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[*]Presenting and corresponding author

Despite the recent advancements driven by deep learning, de novo peptide sequencing is still constrained by incomplete peptide fragmentation and insufficient protein digestion in current single protease-based proteomic experiments. Furthermore, the evaluation of de novo sequencing algorithms heavily depends on constructing restricted benchmark datasets from peptide-spectrum matches obtained by database search. we present a software system, named DiNovo, for high-coverage and high-confidence de novo peptide sequencing by leveraging the complementarity of mirror proteases. DiNovo is empowered by several innovative algorithms, including a mirror-spectra recognition algorithm independent of pre-sequencing, two sequencing algorithms based on deep learning and graph theory, respectively, and target-decoy mapping, a method for sequencing result evaluation free of prior peptide identification. We evaluate the performance of DiNovo using two pairs of mirror proteases including trypsin/LysargiNase, and Lys-C/Lys-N, which were used to digest proteins from E. coli and yeast proteomes. By taking full advantage of the complementarity of mirror spectra, DiNovo achieves much higher sequence coverage and confidence than state-of-the-art single-protease de novo sequencing algorithms. Compared with the trypsin protease used alone, DiNovo using two pairs of mirror proteases leads to two to three times high-confidence amino acids sequenced. Furthermore, DiNovo identifies a comparable number of proteins to database search at the same false discovery rate level, showing great potential as a powerful complement or even an alternative to database search for protein identification. DiNovo is designed to dramatically increase the coverage and confidence of de novo peptide sequencing and to facilitate mirror-protease proteomics.