Computationally instrument-resolution-independent de novo peptide sequencing for high resolution devices

Rui Qiao[1], Hieu Tran[2], Lei Xin[2], Ming Li[1], Ali Ghodsi[1]

[1]University of Waterloo, Waterloo, Ontario, Canada

[2]Bioinformatic Solutions Incorporation, Waterloo, Ontario, Canada

Introduction

Over the past decade, the resolution and accuracy of mass spectrometers have been improved by orders of magnitude and mass spectra of higher resolution have been generated. How to effectively take advantage of those high-resolution data without significantly increasing the computational complexity remains a challenge for de novo peptide sequencing tools. Existing tools often explicitly or implicitly discretize mass spectra at certain resolution (m/z bin method). For spectra of higher resolution, this will result in higher computational complexity and memory usage. Here we present PointNovo, a neural network based de novo peptide sequencing model that can robustly handle any resolution levels of mass spectrometry data while keeping the computational complexity unchanged. Our extensive experiment results show PointNovo outperforms existing de novo peptide sequencing tools by capitalizing on the ultra-high resolution of the latest mass spectrometers.

Methods

PointNovo does not vectorize mass spectra, instead, it directly represent a mass spectrum as a set of peaks. At each step of prediction, the m/z difference between each observed peak and the fragment ions of each potential partial peptide is computed. The resulting m/z difference matrix D contains useful information of what next amino acid residue could be. We point out here that since a spectrum is a set, the model should be order invariant, i.e. output the same prediction regardless of the order of peaks in the input spectrum. PointNovo adopts an order invariant network structure: T Net to process D and the output is guaranteed to be the same for any row permutations of D. Since PointNovo directly exrtract features from m/z between observed and expected peaks, it can robustly handle spectra of any resolution.

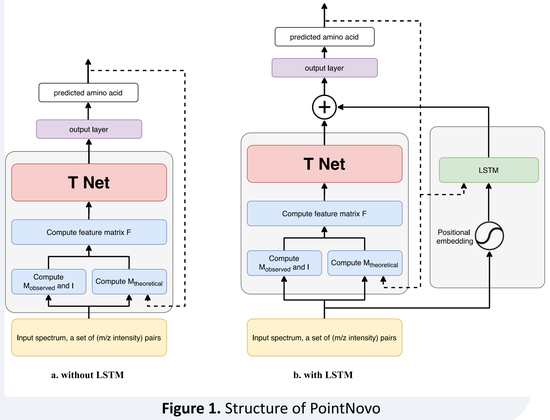

Figure 1. Structure of PointNovo

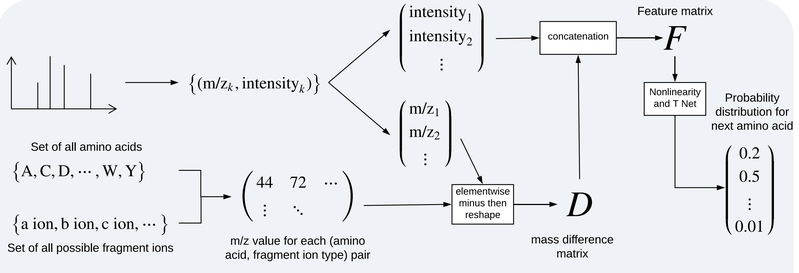

Figure 2. Overview of PointNovo

Results

PointNovo has better de novo peptide sequencing accuracy

we collected three high-resolution MS/MS spectra datasets provided by different labs (Hela samples from ABRF, PXD008844 and PXD010559). On each of the three datasets, we first ran a database search using PEAKS X. The identified PSMs at 1% FDR, on each dataset, are split into train, validation and test set in the ratio of 8:1:1. During the split, we made sure that no common peptide sequences are shared among the train, validation and test sets. Then for each of the three high resolution MS/MS spectra datasets, two PointNovo models are trained from scratch on the train set. The weights that show the best validation loss during training are saved as the trained model weights. Finally, trained models are evaluated on the test set. The amino acid level accuracy, amino acid level recall and peptide level recall on the test set are reported in Figure 3.

PointNovo has better identification between amino acid pairs of similar mass

To further demonstrate that our proposed PointNovo model could take full advantage of the high-resolution data and better discriminate between amino acids pairs that have similar masses, we calculate the precision and recall for amino acid pairs F and M(Oxidation) (the mass difference is smaller than 0.035 Da), Q and K. In this analysis, a predicted amino acid is considered as matching the ground truth amino acid in the target sequence if and only if the amino acids are exactly the same and the prefix masses before them are different by less than 0.5 Da.

Both DeepNovo and PointNovo are trained without the LSTM modules, since we want to compare their ability of learning from spectra, not their ability to remember the sequence patterns. The precision-recall curves for two datasets are shown in Figure 4 and 5.

The above results demonstrate that PointNovo outperforms previous state of the art de novo peptide sequencing tools by a significant margin and could better discriminate between similar amino acids pairs. Also, unlike previous neural network based de novo peptide sequencing tools, PointNovo does not include any spectrum vectorization, thus is ready to be applied on the more precise MS data generated in the future. The unique feature extraction method used by PointNovo could be further applied on other important problems in MS like database search and spectra library search.

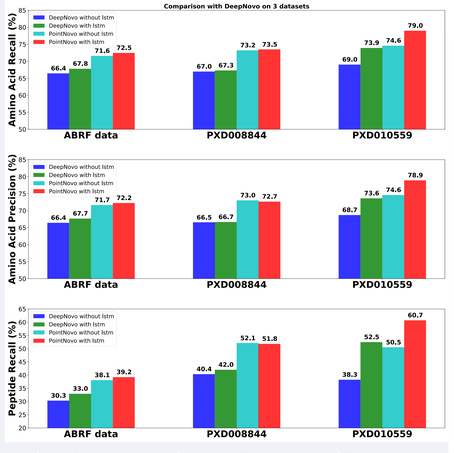

Figure 3. Comparison between DeepNovo and PointNovo

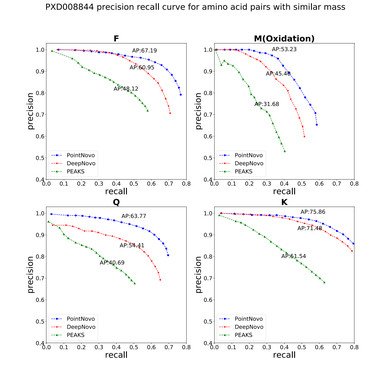

Figure 4. Precision recall curve on PXD008844

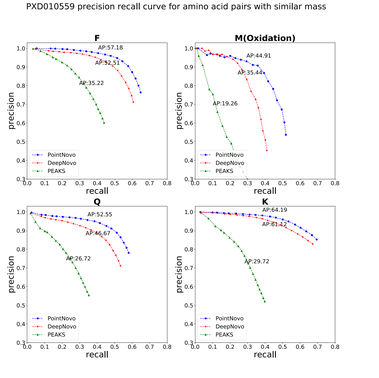

Figure 5. Precision recall curve on PXD010559

Contact

Rui Qiao rqiao@uwaterloo.ca

Bioinformatics Solutions Inc. http://www.bioinfor.com peaks@bioinfor.com