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

*Jihun Cha[‡,1], Seunghyuk Choi[1] and Eunok Paek[*1]*

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

1. INTRODUCTION

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. 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.

2. METHODS

We adopted the precursor-swap method[1] to generate decoy spectra directly at a spectrum level. With given spectra, the precursor-swap method simply chooses two spectra with the same charge state and with a precursor difference less than pre-defined precursor tolerance, then swap their precursors. This method takes advantage of the fact that in order to identify spectra, their precursors should be obtained precisely. Target and decoy spectra are searched and assigned to peptides by de novo sequencing tools such as PEAKS[2], and the search results are used as a positive and negative training set, respectively.

NovoFit then calculates features (Figure 1) that can assess peptide-spectrum match quality for the target and decoy PSMs. Within all 27 features, the boruta algorithm[3] discards weak features which are better to be removed than used for the rest of the workflow. Positive training and negative training sets are made out of each target and decoy PSMs: unique peptide level positive and negative datasets are constructed, and confident positive PSMs and randomly sampled negative PSMs are used as inputs for iterative training (Figure 2). By using random forest, training sets are iteratively re-scored with a random forest probability score (NovoFit score), resulting in a new training set with each iteration.

Features			
1	ObservedPepMass	12-13	SumY(B)matchInt
2	CalculatedPepMass	14-15	FracY(B)matchInt
3	DeltaMass	16-17	SeqCoverY(B)ion
4	DeltaMass_abs	18-19	ConsecutiveY(B)ion
5	DeltaMass_ppm	20	NumofAnnoPeaks
6	PeptideLength	21-22	MassErrMean(SD)
7	TIC	23-24	Max(Min)MassErr
8	Num_missed_cleavage	25	Comet Cn
9	MaxIntAll	26	DeltaCn_second
10-11	MaxY(B)matchInt	27	DeltaCn_topn

Figure 1. NovoFit Features

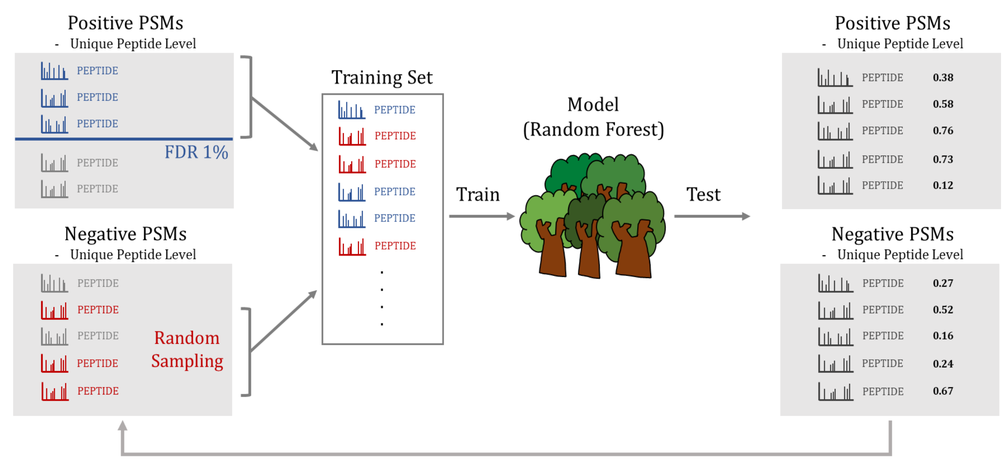
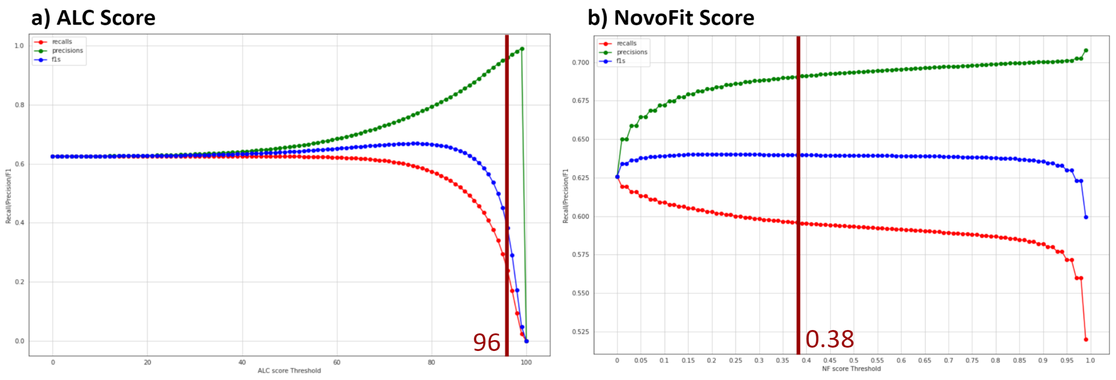


Figure 2. NovoFit Iterative training workflow

3. RESULTS

Comparison with the result using PEAKS ALC score



	ALC Score (FDR 1%)	NovoFit Score (FDR 1%)
Recall	17.09%	59.54%
Precision	96.94%	69.10%
F1 Score	**0.29**	**0.64**

Figure 3. Comparison between the result using PEAKS ALC score and NovoFit score

To evaluate the performance of NovoFit, we used synthetic peptide dataset from ProteomeTools (PXD004732)[4]. 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 (Figure 3). 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 only ~0.29 but ~0.64 with NovoFit score. Figure 3-a and 3-b show that FDR estimation with ALC score cannot get the optimal F1 score value while FDR estimation with NovoFit score can. These results show that using NovoFit can estimate FDR considering the characteristics of the dataset. Note that these results are derived under the assumption that MaxQuant[5] database search results are the ground truth.

Comparison with the MaxQuant database search result

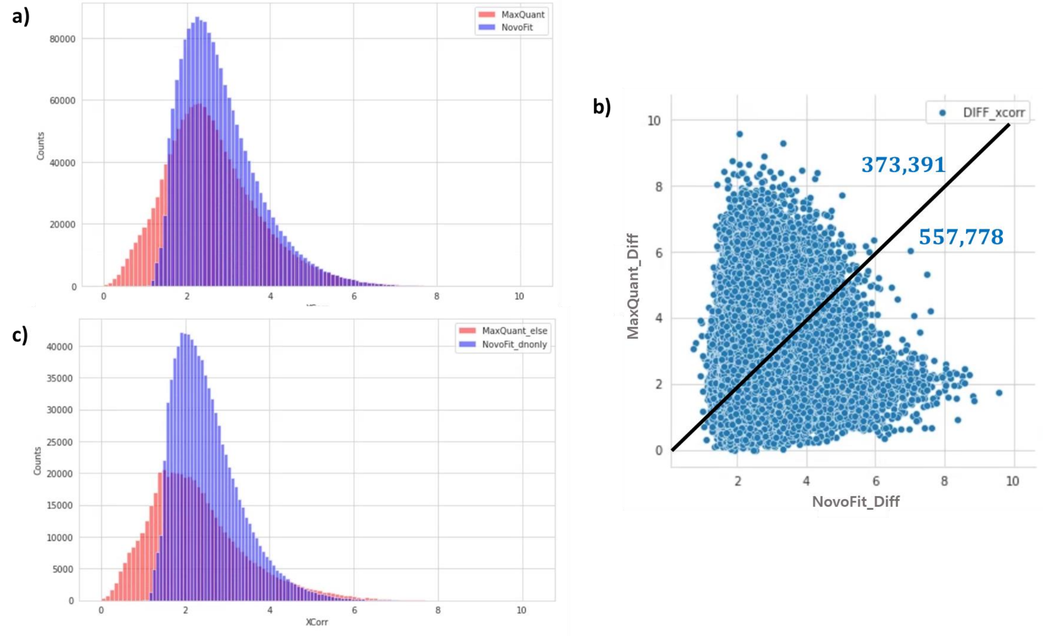


Figure 4. Distribution of cross-correlation values of PSMs identified by MaxQuant and NovoFit.

To compare NovoFit identification result and the MaxQuant identification result with an objective indicator, we used Comet[6] to obtain the cross-correlation value of PSMs identified from NovoFit and MaxQuant each(Figure 4). Figure 4-a is a comparison between the cross-correlation value distribution of all PSMs identified by MaxQuant and NovoFit. Figure 4-b is a comparison between the cross-correlation value distribution of PSMs both identified by MaxQuant and NovoFit but matched with different peptides. Figure 4-c is a comparison between the cross-correlation value distribution of PSMs identified by either MaxQuant or NovoFit. It is notable that PSMs identified with NovoFit mostly had higher cross-correlation values than ones identified with MaxQuant. These results imply that it is necessary to interpret the results with the possibility of mutation or SNV through additional analysis of the results identified by NovoFit.

4. CONCLUSIONS AND DISCUSSION

NovoFit is a tool for FDR estimation by post-processing de novo sequencing results using machine learning. We used the precursor-swap method, which is mainly used in spectral library search, to construct data. Additional features used for peptide-spectrum match quality evaluation are calculated for every PSMs. Using machine learning, NovoFit can iteratively learn and re-score the dataset, resulting in an estimate of FDR considering the characteristic of the data. The F1 score of the result derived by using PEAKS ALC score was ~0.29; using NovoFit the F1 score went up to ~0.64. Although, it gets lower precision than PEAKS; further studies are in progress for getting more reliable identification. Finally, It was verified that the results presented by NovoFit are convincing and reliable by showing match quality of PSMs presented by NovoFit is rather good comparing to those of MaxQuant. It can be implied that about results from NovoFit, additional analysis with the possibility of mutation is needed.

REFERENCES

[1] Cheng, C. Y., Tsai, C. F., Chen, Y. J., Sung, T. Y., & Hsu, W. L. (2013). Spectrum-based method to generate good decoy libraries for spectral library searching in peptide identifications. *Journal of proteome research*, *12*(5), 2305-2310.

[2] Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., & Lajoie, G. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, *17*(20), 2337-2342.

[3] Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J Stat Softw*, *36*(11), 1-13.

[4] Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., ... & Kuster, B. (2017). Building ProteomeTools based on a complete synthetic human proteome. *Nature methods*, *14*(3), 259-262.

[5] Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, *26*(12), 1367-1372.

[6] Eng, J. K., Jahan, T. A., & Hoopmann, M. R. (2013). Comet: an open-source MS/MS sequence database search tool. *Proteomics*, *13*(1), 22-24.

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01373, Artificial Intelligence Graduate School Program(Hanyang University), National Research Foundation of Korea (NRF) grants (2019M3E5D3073568), and by the BK21 FOUR (Fostering Outstanding Universities for Research) project of the National Research Foundation of Korea Grant.