Monoclonal Antibody *de novo* Sequencing by Deep Learning

Ngoc Hieu Tran[1], Xianglilan Zhang[1], Lei Xin[2], Lin He[2], Baozhen Shan[2], Ming Li[1]
[1] University of Waterloo, Waterloo, ON, Canada
[2] Bioinformatics Solutions Inc., Waterloo, ON, Canada

Introduction

The amino acid sequence of an antibody is mainly obtained by DNA sequencing of source cell line. However, *de novo* sequencing is required, when the original cell line or cDNA is unavailable [1]. For antibody *de novo* sequencing, multi-enzyme digestion followed by high-resolution mass spectrometry has been demonstrated to be a reliable approach. There, the accuracy of peptide *de novo* sequencing is crucial. In this study we propose a novel deep neural network model, DeepAB, to address this problem. The key aspect of DeepAB is its ability to auto-learn multiple levels of representation of high-dimensional data. Furthermore, DeepAB is re-trainable without needing the pre-designed domain-specific features. Preliminary results show that DeepAB outperforms current approaches of peptide *de novo* sequencing.

Method

The DeepAB model takes a spectrum as input and sequence the peptide by predicting one amino acid at each iteration. In DeepAB, a spectrum is discretized into a vector in which masses correspond to indices and intensities are values. DeepAB incorporates two classification models, a Convolutional Neural Network (CNN) and a Long Short Term Memory Network (LSTM), which uses the output of previous sequencing steps as a prefix to predict the next amino acid. The outputs of the two models are then combined via a fully-connected layer followed by a linear classifier to produce a probability distribution over the amino acid classes. Putting all together, DeepAB finally performs beam search until it found the optimum amino acid prediction.

Given the predicted peptides, they can be assembled into complete antibody protein sequences by using de Bruijn graph techniques such as ALPS. As shown in Figure 1, newly identified proteins are then used to enhance existing databases that subsequently produce more training data to further improve the performance of DeepAB. This self-reinforcing circle, with DeepAB as the main driving force, is an ideal solution to the problem of antibody *de novo* sequencing.

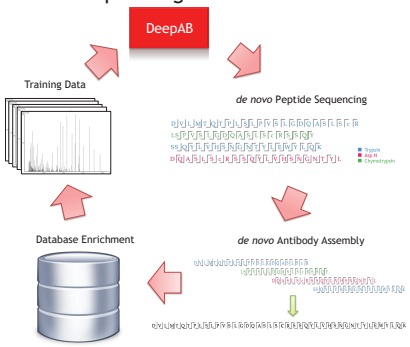

Figure 1. DeepAB and AI virtuous circle in the context of antibody *de novo* sequencing

Results

The performance of DeepAB was evaluated on LC-MS datasets from thirteen monoclonal antibody samples, MS data were acquired with Thermo Q Exeactive. The total number of MS/MS spectra in the twelve datasets is 1,549,530.

Database search was performed to find real peptide sequence of a spectrum. Each data set was analyzed by PEAKS® DB software, with fixed modification of Carbamidomethylation (C), variable modifications of Oxidation (M) and Deamidation (NQ). The mass error tolerance was set to 10 ppm for precursor ions, and 0.05 Da for fragment ions. The peptide sequences identified with 0.1% of false discovery rate (FDR) were used as ground-truth for testing the accuracy of *de novo* sequencing results.

To measure the accuracy of *de novo* sequencing results, we compared the *de novo* peptide sequence with the real sequence. A *de novo* amino acid is considered "matched" with a real amino acid if their masses are different by less than 0.05 Da and the prefix masses before them are different by less than 0.05 Da. Such approximate match is used instead of exact match because of the resolution of LC-MS/MS instruments.

Six datasets were used for training DeepAB, and the remaining five datasets and other two public mouse antibody datasets were used for testing. The accuracy was compared with current state-of-the-art *de novo* peptide sequencing tool PEAKS®.

We calculated the total recall (and precision) of *de novo* sequencing as the ratio of the total number of "matched" amino acids over the total length of real peptide sequences (and predicted peptide sequences, respectively) in the testing datasets. We also calculated the recall at peptide level, i.e. the fraction of real peptide sequences that were fully correctly predicted. As shown in Figure 2., DeepAB considerably outperformed PEAKS® in all seven testing datasets. For example, at amino acid level, compared with PEAKS®, the recall of DeepAB is increased by 16%, 16%, 9%, 8%, 22%, 19% and 19% respectively.

Most importantly, all sequencing tools report confidence scores for their predictions. The confidence scores reflect the quality of predicted peptides and are valuable for downstream analysis. Here, we used confidence scores to compute the precision-versus-recall curves and the area under the curves (AUC). The AUC is a standard metric for classification evaluation and captures both precision and recall. Figure 3 show the precision-recall curves and the AUC of PEAKS® and DeepAB. DeepAB considerably outperformed PEAKS® across all seven datasets.

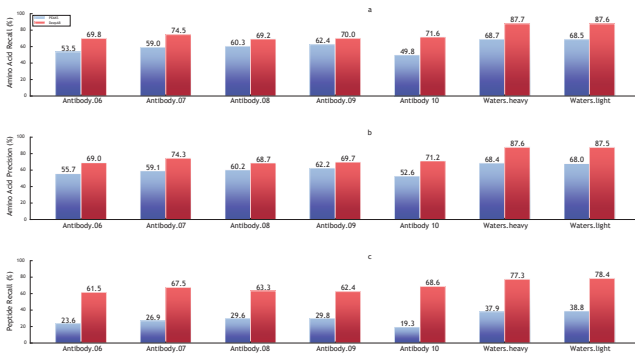

Figure 2. Total recall and precision of PEAKS® and DeepNovo on seven datasets. a. Recall at amino acid level. b. Precision at amino acid level. c. Recall at peptide level.

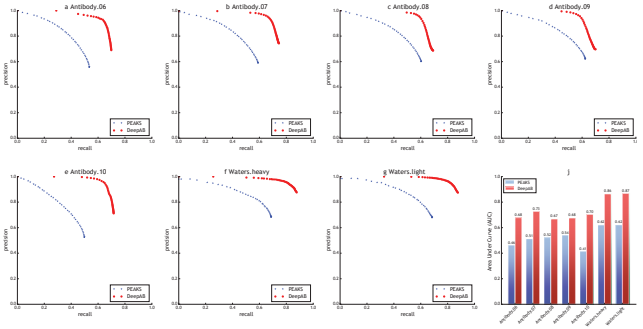

Figure 3. The preicison-recall curves and the area under the curves (AUC) of PEAKS® and DeepAB. a. Precision-recall curves on Antibody.06. b. Precision-recall curves on Antibody.07. c. Precision-recall curves on Antibody.08. d.Precision-recall curves on Antibody.09. e. Precision-recall curves on Antibody.10. f. Precision-recall curves on Waters.heavy. g. Precision-recall curves on Waters.light. h. AUC of PEAKS® and DeepAB on seven datasets.

Conclusion

In this study, we proposed DeepAB, a deep neural network model that combines recent advances in deep learning and dynamic programming to address the problem of antibody *de novo* sequencing. Our experiments on seven different antibody datasets show that DeepAB consistently surpassed state-of-the-art records in antibody *de novo* peptide sequencing.

Reference

[1] Sen K.I., et al. J. Am. Soc. Mass Spectrom. (2017) 28:803. doi:10.1007/s13361-016-1580-0.

www.bioinfor.com