Thesis for the Master of Science

NovoRank: Machine Learning Based Post-processing for Performance Improvement in De Novo Peptide Sequencing

Jangho Seo

Graduate School of Hanyang University

August 2022

Thesis for the Master of Science

NovoRank: Machine Learning Based Post-processing for Performance Improvement in De Novo Peptide Sequencing

Thesis Supervisor: Eunok Paek

A Thesis submitted to the graduate school of
Hanyang University in partial fulfillment of the requirements
for the degree of Master of Science

Jangho Seo

August 2022

Department of Artificial Intelligence
Graduate School of Hanyang University

This thesis, written by Jangho Seo,
has been approved as a thesis for the Master of Science.

August 2022

Committee Chairman:	Heejin Park	(Signature)
Committee member:	Eunok Paek	(Signature)
Committee member:	Je-Hyun Baek	(Signature)

Graduate School of Hanyang University

Table of Contents

Figures

Tables

Abstract

NovoRank: Machine Learning Based Post-processing for Performance Improvement in *De Novo* Peptide Sequencing

Jangho Seo
Department of Artificial Intelligence
The Graduate School
Hanyang University

To identify peptides in mass spectrometry-based proteomics, tandem mass (MS/MS) spectra are analyzed using database search or *de novo* sequencing tools. In contrast to database search approaches, *de novo* sequencing directly deduces peptide sequences from MS/MS spectra without any reference to sequence databases. *De novo* sequencing method often generates incorrect peptide identifications due to its practically unlimited search space and its peptide identification performance does not reach that of database search methods. Instead, *de novo* sequencing has the advantage of finding novel peptides that are not a part of the sequence database, thus is an essential method for discovering peptides of as yet unknown, biologically important functions.

Here, we propose a machine learning based post-processer for de novo sequencing tools, named NovoRank, that can improve the performance of *de novo* sequencing and is applicable with any *de novo* peptide sequencing tools. NovoRank uses DBSCAN, a well-known density-based clustering

algorithm, and adopts deep learning techniques so that candidate peptide reordering can give a better top-ranked sequence.

Given a large-scale synthetic peptide dataset (ProteomeTools), NovoRank increased the peptide recall by 8.63~12.66% when applied with *de novo* sequencing results from three different software tools.

1 Introduction

Proteins are important substances that perform various functions of cells as a component in organism. Proteomics, which studies proteome, is interested in identifying proteins in the first place. The current shotgun proteomics technology allows protein sequence characterization (from a protein mixture) by identifying peptides and then reconstruct the original protein sequence from its constituent peptides. Proteins are digested into peptides by an enzyme and the digested peptides are analyzed by tandem mass spectrometry [1]. Finally, the acquired tandem mass (MS/MS) spectra can be analyzed to identify peptide sequences.

There are two major approaches to identify peptides from an MS/MS spectrum: database search and *de novo* sequencing. Database search is a method that uses a sequence database and compares an experimental spectrum with theoretical spectra generated from the peptide sequences in the database to find the best match. On the other hand, *de novo* sequencing infers peptide sequences solely based on an MS/MS spectrum without any reference to a sequence database.

De novo sequencing has the advantage of finding novel peptides because it does not use prior knowledge, but its huge search space makes it more sensitive to the noise and missing peaks of MS/MS spectra, compared with a database search approach [2], sometimes resulting in false identifications. To overcome such performance limitation and achieve more reliable results, there have been efforts to post-process de novo sequencing results.

In this research, we propose a post-processing method NovoRank that improves the performance of *de novo* sequencing, which uses machine

learning techniques such as clustering and deep learning [3]. NovoRank takes top N candidate sequences, provided by *de novo* sequencing tools such as PEAKS [4], pNovo3 [5], and DeepNovo [6], as its input and then reorders their ranks based on various features such as peptide-spectrum match quality. Instead of processing each spectrum independently, we first cluster tandem mass spectra so that similar spectra can share their candidate sequence information and then reorder candidate ranking per cluster. Finally, cluster-level ranking is re-evaluated per spectrum based on a deep learning model. Six additional features are adopted as an input to the deep learning network so that a single candidate is finally assigned to each spectrum.

2 Related work

2.1 pNovo3

pNovo3 is a framework that improves precision of *de novo* sequencing, including post-processing that re-ranks *de novo* sequencing results.

In order to get candidate peptides in pNovo3, pNovo [7] is used. 6 features such as the original peptide-spectrum match score provided by pNovo, 3 similarity scores, 2 gap information were extracted for re-ranking. The similarity score is calculated by comparing between the experimental and the theoretical spectrum with the cosine, Pearson and Spearman methods, where a theoretical spectrum was generated using pDeep [8]. The gap information is a feature that can help determine amino acid ordering when there is no fragment ion peak between two consecutive amino acids and is obtained from pre-calculated probability values of missing the fragment ion. A support vector machine [9] model was trained to re-rank the candidate peptides originally provided by pNovo.

To further improve the performance, an additional process is performed to merge re-ranked results that are similar in precursor mass, spectrum and rank 1 peptide sequence under pre-set conditions.

2.2 Learning to rank

Learning to rank (LTR) [10] is a machine learning methodology that applies supervised learning to solve ranking problems that determining the order of search results. For peptide sequence identification, when providing search results to users, it is more important to present the result most relevant to a query in the top rank, rather than simply show a list containing all results similar to the query. So far, LTR is a subject that has been studied a lot in the field of information retrieval and recommendation systems.

To solve the LTR problem, there are pointwise, pairwise, and listwise approaches. First, a pointwise approach takes one item as input and calculates a score. After calculating the scores for all the items, the results are provided in a descending order. A pairwise approach handles two items at the same time and aims to sort each pair of items rather than scoring each. Lastly, a listwise approach receives an item list as input at once and then determines the total order of the items in the list.

3 Materials

3.1 Datasets

ProteomeTools [11] datasets from ProteomeXchange with the accession PXD004732 containing synthetic human peptides were used in this experiment. The dataset consists of 123 HCD raw files consisting of 6,359,460 MS/MS spectra, and we converted all the raw files to MGF format using MSConvert [12]. MaxQuant [13] database search results were also downloaded from the PRIDE archive and only the PSMs with PEP (Posterior Error Probability) score less than or equal to 0.01 were selected as reliable PSMs, resulting in 3,506,774 PSMs of 134,615 peptides, which were taken as the ground truth and used for performance evaluation and model training.

3.2 *De novo* peptide sequencing results

The 123 MGF files of the ProteomeTools datasets contain 6,359,460 spectra. *De novo* sequencing results were obtained from three different tools – PEAKS and DeepNovo were applied against the entire 6,359,460 spectra, while pNovo3 was applied only to 4,279,546, because some files were abnormally terminated during execution. We obtained the top 10 candidate results using PEAKS and pNovo3, while DeepNovo provides only the top 1 peptide per spectrum. *De novo* sequencing using DeepNovo was performed with the pre-trained model provided by its github repository (https://github.com/nh2tran/DeepNovo).

The peptide identification results and search parameters of each *de novo* sequencing tool are shown in Table 1 and 2.

	PEAKS	pNovo3	DeepNovo
Peptide Identification Result	6,274,999	3,594,052	6,248,624
Ground truth in top 1 result	2,190,757	2,411,900	1,416,950
Ground truth in top 10 results	2,738,266	2,648,639	X

Table 1. PEAKS, pNovo3, DeepNovo identification result

	PEAKS	pNovo3	DeepNovo
De novo sequencing result	Top 10 Candidates	Top 10 candidates	Top 1 candidate
Precursor tolerance	10 p.p.m	10 p.p.m	pre-trained model is used
Fragment tolerance	0.025 Da	0.025 Da	
Fixed modification	C (Carbamidomethylation)	C (Carbamidomethylation)	
Variable modification	M (Oxidation)	M (Oxidation)	

Table 2. Search parameters of each *de novo* sequencing

4 Methods

4.1 NovoRank

NovoRank is a post-processing tool that tries to assign the correct peptide using the top N ($N \geq 1$) candidates from *de novo* sequencing results. The workflow of NovoRank is shown in Figure 1. NovoRank consists of two major steps: (1) new candidate generation and (2) re-ranking.

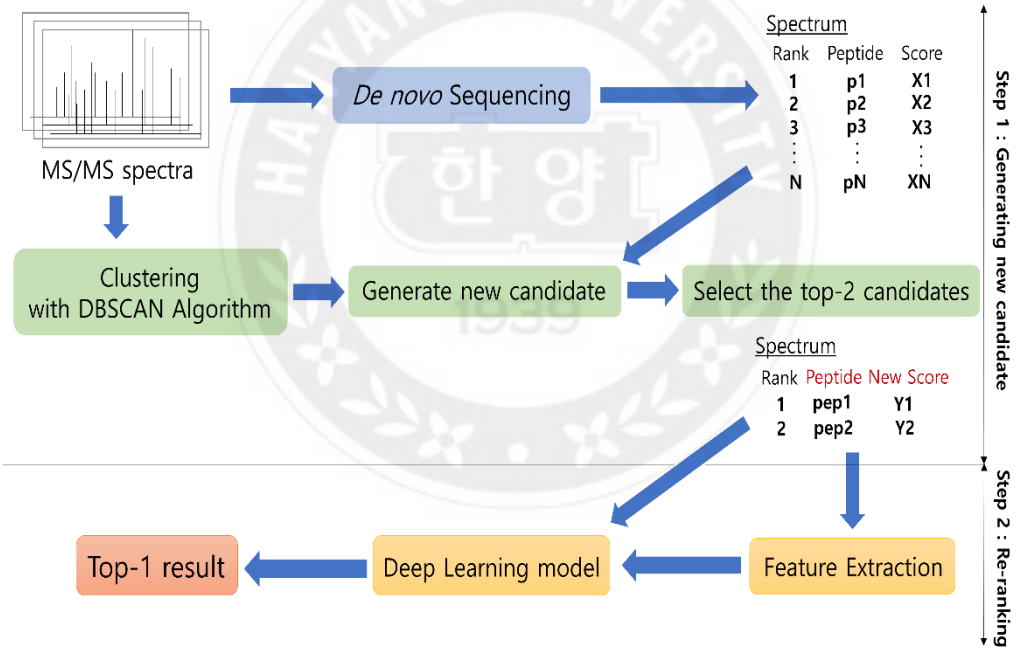

Figure 1. The workflow of NovoRank

4.2 Generating new candidate step

First, *de novo* sequencing tools such as PEAKS, pNovo3, and DeepNovo were applied to get the top N candidates for each spectrum. Simultaneously, multi-dimensional spectral clustering was conducted, so that candidate sequences can be pooled and shared among similar spectra. MS/MS spectra clustering was performed by applying DBSCAN [14] algorithm using three features - precursor m/z, charge and retention time, assuming that MS/MS spectra produced from the same peptide should have been observed with similar precursor masses and eluted at similar retention times, and that the spectra would look very different if the charge states of precursors were different even if they had resulted from the same peptide. We merged the candidates of all spectra in the same cluster and retained top 2 candidate peptides at the cluster level based on a new score calculated based on their original score and identification frequency. From PEAKS, pNovo3 and DeepNovo results, we used 'ALC', 'Final score of this result', and 'predicted_score' as its original score, respectively. The new score is calculated for all unique peptides in a cluster, by summing all of its original scores in the cluster.

Table 3 shows the percentage of correct answers in the newly determined candidates. From Table 3, we can expect that the performance can be further improved by 5.8%~9.6% at most, if we perfectly re-rank using new (cluster-assigned) top 2 candidates. If we consider the top 5 or top 10 candidates, there could be more performance improvements, but the performance gain is not so much bigger than the improvement obtained by re-ranking the top 2 while it is a much more difficult problem to solve. For this reason, we performed re-ranking using newly assigned top 2

candidates.

Peptide Recall	PEAKS	pNovo3	DeepNovo
Top 1	72.70 %	77.21 %	52.93 %
Top 2	82.26 %	82.79 %	59.98 %
Top 5	87.74 %	86.12 %	64.55 %
Top 10	90.05 %	87.83 %	65.76 %

Table 3. *De novo* sequencing performance after post-processing only with clustering

4.3 Re-ranking step

In the re-ranking step, additional features were extracted for each spectrum so that each peptide spectrum match is re-evaluated between the top 2 candidate peptides from the previous step. The additional features included peptide ranking, new score, delta score, cluster size, number of matched internal fragment ions divided by peptide length, and absolute value of the difference between real RT(Retention Time) and predicted RT obtained using DeepLC [15] (Table 4). All features except for peptide ranking were transformed using logarithmic function.

Additional Feature	Description
Peptide ranking	0 (Rank 1) or 1 (Rank 2)
Score	Sum of original *de novo* sequencing scores
Delta score	Difference between rank 1 and rank N scores (N = 1 or 2) (if N = 1, delta score is always 0)
Cluster size	Size of Cluster
Internal fragment ion	Number of matched internal fragment ions / Peptide length
Delta RT	Abs (real RT – predicted RT)

Table 4. Description of 6 additional features for NovoRank

4.4 Deep learning model architecture for peptide re-ranking

Figure 2 shows the deep learning model architecture for selecting the correct peptide. It is designed similarly to Siamese Network [16] and RankNet [17], one of the pairwise approaches, as a model for learning to rank based on two identical neural networks that share weights. Each network receives three different inputs from each of rank 1 and 2 candidates.

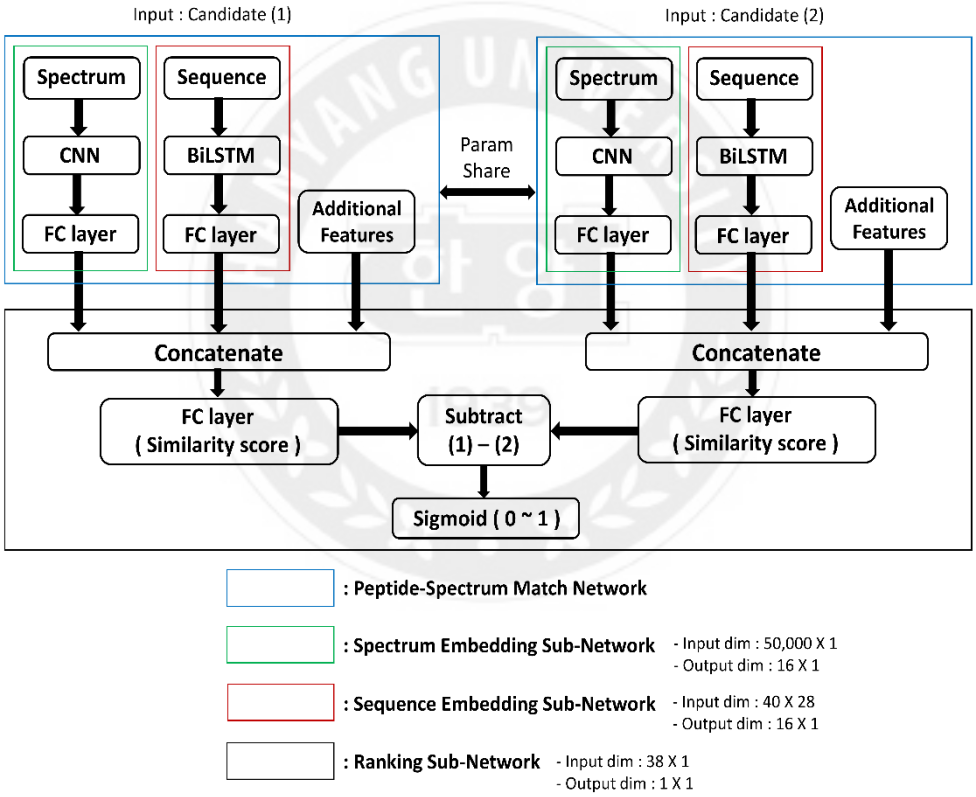

Figure 2. Deep learning model for re-ranking

4.5 Spectrum and Sequence embedding sub-network

The spectrum embedding sub-network has weight layers consisting of two convolutional (Conv) layers and three fully connected (FC) layers. The two Conv layers with 1×30 filter with stride of 1 are used, with 8 and 16 filters, respectively. The FC layers has 16 channels. A max-pooling layer is used for Conv layers with the activation function leakyReLU [18] and a dropout rate of 0.2.

The sequence embedding sub-network has weight layers consisting of one Bidirectional LSTM (BiLSTM) [19] and three FC layers. The sizes of BiLSTM and FC layers were 8 and 16, respectively. All hidden layers use the activation function of leakyReLU. A dropout rate of 0.2 is used after the FC layer.

4.6 Ranking sub-network

The spectrum and sequence are transformed into 16-dimensional vector, respectively, and the two embedding vectors and 6 additional features are concatenated. The similarity scores are calculated from each peptide-spectrum match network, using FC layers of three, and their difference is fed to a sigmoid function, which finally outputs the correct peptide.

4.7 Input data encoding

We allowed the maximum mass of a peptide to be 5,000 Da and its resolution 0.1 Da. So, a spectrum is represented as a 50,000-dimensional vector. Each element of a vector has a normalized intensity value between 0 and 1. A peptide sequence is transformed into a matrix. The maximum length of a sequence is 40 and zero values are padded if a peptide is shorter than 40. An amino acid is represented as a vector of size 28 including 6 charges, 20 amino acids, and 2 modifications – carbamidomethylation at Cysteine and oxidation at Methionine. Thus, a sequence is represented as a 40×28 matrix.

4.8 Training, Validation and Test

We use 80% of the reliable PSMs for training and the rest for testing. The validation set for the evaluation of the deep learning model was obtained by partitioning the training set, and 5-fold cross validation was conducted. Train/Validation/Test set was split so that peptides never overlap amongst the three. Among *de novo* peptide sequencing results from PEAKS, pNovo3, and DeepNovo, we randomly select one search tool result at a cluster level to create training and validation set so that each spectrum is assigned a peptide from a single *de novo* search tool.

Our deep learning model solves the binary classification problem of selecting the correct peptide given the two candidates. Output label is set to 0 if candidate (1) is correct and 1 otherwise. We used the Adam optimizer [20] and binary cross entropy loss function. Epoch and batch size were set to 50 and 64, respectively.

5 Results

5.1 Clustering quality

We tried several clustering methods to merge the identification results of similar spectra. To evaluate whether clustering was successful, the purity of a cluster was checked. Clustering evaluation was conducted on the reliable PSMs described in 3.1. The purity of a cluster was defined as the number of unique peptides in the cluster. If the smaller number of unique peptides was contained in a cluster, its purity is considered higher. Our experiment shows that it is better to perform clustering using only DBSCAN algorithm than to conduct a multi-stage clustering with MS-Cluster [21], or applying MS-Cluster alone. For multi stage clustering, the initial clustering of MS/MS spectra was performed using MS-Cluster software and the resulting clusters were further partitioned by retention time using DBSCAN algorithm to increase the purity of the cluster.

Detailed figures are shown in Table 5. In Table 5, only clusters with up to 3 unique peptides in the cluster are shown. When we compare MS-Cluster and DBSCAN results, we can see that there is a tradeoff between the cluster purity and the number of clusters. If all the clusters consist of a single spectrum, then the purity of such clustering will be the highest. Thus, we want clustering results that show higher purity but consist of a smaller number of clusters. In consideration of both purity and the number of clusters, we decided that clustering using only DBSCAN was our best choice.

	MS-Cluster	MS-Cluster + DBSCAN	DBSCAN
The number of unique peptides: The number of clusters	1: 155,288 2: 14,426 3: 2,059	1: 262,920 2: 9,260 3: 269	1: 260,533 2: 5,387 3: 147
The number of unique peptides: The number of scans	1: 2,890,734 2: 470,223 3: 101,823	1: 3,294,407 2: 203,343 3: 8,529	**1: 3,370,491** 2: 130,385 3: 5,587
Maximum number of unique peptides in cluster	11	4	4
Total number of clusters	172,568	272,462	266,073
Number of clusters of size of 1	3179	8129	0

Table 5. The clustering results

5.2 NovoRank Evaluation

In Table 6, the performance is evaluated by comparing the NovoRank results with original *de novo* sequencing results by peptide recall (Table 6a) and amino acid recall (Table 6b) on the test set. The NovoRank results are shown by dividing them into two results: Clustering and Re-ranking using the deep learning model.

We show that NovoRank increased performance by 8.63~12.66 % and 5.17~10.53 % and in peptide recall and amino acid recall, respectively, when compared with using only the *de novo* sequencing tool.

a)

	Original	Clustering	Re-ranking
PEAKS	62.68 %	72.86 %	74.29 %
pNovo3	69.30 %	77.49 %	77.93 %
DeepNovo	41.08 %	53.37 %	53.74 %

b)

	Original	Clustering	Re-ranking
PEAKS	87.14 %	91.93 %	92.31 %
pNovo3	80.19 %	89.97 %	90.72 %
DeepNovo	69.96 %	77.93 %	78.69 %

Table 6. Peptide recall and amino acid recall of PEAKS, pNovo3, DeepNovo on the test set in each step.
a) Recall at peptide level b) Recall at amino acid level

6 Conclusion

De novo sequencing is an essential method for finding novel peptides. However, it is not easy to find the optimal peptide sequence because of its large search space.

In this work, we proposed a machine learning based post-processing tool, called NovoRank, that improves the match quality of *de novo* sequencing. We proposed several processes that help us find the correct peptide sequence among candidate sequences returned by the existing de novo sequencing tools. We show that there is a significant performance improvement when the top 1 sequence is selected based on its original score by the *de novo* software and its frequency within the cluster, which is obtained by merging *de novo* sequencing results of similar spectra. We also achieved additional performance improvement using the newly designed deep learning model for re-ranking using six additional features.

References

[1] Steen, Hanno, and Matthias Mann. "The ABC's (and XYZ's) of peptide sequencing." *Nature reviews Molecular cell biology* 5.9 (2004): 699-711.

[2] McDonnell, Kevin, Enda Howley, and Florence Abram. "The impact of noise and missing fragmentation cleavages on de novo peptide identification algorithms." *Computational and Structural Biotechnology Journal* 20 (2022): 1402-1412.

[3] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436-444.

[4] Ma, Bin, et al. "PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry." *Rapid communications in mass spectrometry* 17.20 (2003): 2337-2342.

[5] Yang, Hao, et al. "pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework." *Bioinformatics* 35.14 (2019): i183-i190.

[6] Tran, Ngoc Hieu, et al. "De novo peptide sequencing by deep learning." *Proceedings of the National Academy of Sciences* 114.31 (2017): 8247-8252.

[7] Chi, Hao, et al. "pNovo: de novo peptide sequencing and identification using HCD spectra." *Journal of proteome research* 9.5 (2010): 2713-2724.

[8] Zhou, Xie-Xuan, et al. "pDeep: predicting MS/MS spectra of peptides with deep learning." *Analytical chemistry* 89.23 (2017): 12690-12697.

[9] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20.3 (1995): 273-297.

[10] Liu, Tie-Yan. "Learning to rank for information retrieval." *Foundations and Trends® in Information Retrieval* 3.3 (2009): 225-331.

[11] Zolg, Daniel P., et al. "Building ProteomeTools based on a complete synthetic human proteome." *Nature methods* 14.3 (2017): 259-262.

[12] Holman, Jerry D., David L. Tabb, and Parag Mallick. "Employing ProteoWizard to convert raw mass spectrometry data." *Current protocols in bioinformatics* 46.1 (2014): 13-24.

[13] Cox, Jürgen, and Matthias Mann. "MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification." *Nature biotechnology* 26.12 (2008): 1367-1372.

[14] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *kdd*. Vol. 96. No. 34. 1996.

[15] Bouwmeester, Robbin, et al. "DeepLC can predict retention times for peptides that carry as-yet unseen modifications." *Nature methods* 18.11 (2021): 1363-1369.

[16] Koch, Gregory, Richard Zemel, and Ruslan Salakhutdinov. "Siamese neural networks for one-shot image recognition." *ICML deep learning workshop*. Vol. 2. 2015.

[17] Burges, Chris, et al. "Learning to rank using gradient descent." *Proceedings of the 22nd international conference on Machine learning*. 2005.

[18] Maas, Andrew L., Awni Y. Hannun, and Andrew Y. Ng. "Rectifier nonlinearities improve neural network acoustic models." *Proc. icml*. Vol. 30. No. 1. 2013.

[19] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." *IEEE transactions on Signal Processing* 45.11 (1997): 2673-2681.

[20] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[21] Frank, Ari M., et al. "Clustering millions of tandem mass spectra." *Journal of proteome research* 7.01 (2008): 113-122.

국문 요지

머신러닝 기반 후처리를 통한 드노보 펩타이드 시퀀싱 성능 개선

질량 분석 기반 단백체학에서 펩타이드를 동정하기 위해, 탠덤 질량 스펙트럼을 데이터베이스 검색이나 드노보 시퀀싱 도구를 사용하여 분석한다. 데이터베이스 검색 방법과 달리 드노보 시퀀싱은 서열 데이터베이스를 사용하지 않고, 탠덤 질량 스펙트럼으로부터 직접 펩타이드 서열을 추론한다. 드노보 시퀀싱 방법은 실질적으로 무한에 가까운 탐색 공간으로 인해 종종 펩타이드를 잘못 동정하고, 펩타이드 동정 성능은 데이터베이스 검색 방법에 미치지 못한다. 그러나 드노보 시퀀싱은 서열 데이터베이스에 존재하지 않는 신규 펩타이드를 찾을 수 있다는 장점을 갖고 있어서, 아직 알려지지 않았지만, 생물학적으로 중요한 기능을 가진 펩타이드를 발견하는데 필수적인 방법이다.

본 연구에서는 드노보 시퀀싱의 성능을 향상시킬 수 있고, 다양한 드노보 시퀀싱 도구에 적용할 수 있는 기계학습 기반의 후처리 도구인 NovoRank를 제안한다. NovoRank는 밀도기반 군집화 알고리즘으로 잘 알려진 DBSCAN 알고리즘을 사용하고, 더 나은 재순위 결과를 제공하기 위해 심층학습 기술을 적용한다.

대규모 합성 펩타이드 데이터 집합인 ProteomeTools에 대해서 NovoRank는 세 종류의 드노보 시퀀싱 결과의 펩타이드 재현율을 8.63~12.66 % 증가시킴을 보였다.

Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Copykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

MAY 09, 2022

Degree :	Master
Department :	DEPARTMENT OF ARTIFICIAL INTELLIGENCE
Thesis Supervisor :	Paek, Eunok
Name :	SEO JANG HO (Signature)

연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

2022년05월09일

학위명 : 석사

학과 : 인공지능학과

지도교수 : 백은옥

성명 : 서장호 (서명)

한 양 대 학 교 대 학 원 장 귀 하