Deep Learning-derived Evaluation Metrics for Benchmarking Computational Pipelines for the Analysis of Large-scale Phosphoproteomics datasets

Wen Jiang[1,2], Kai Li[2], Bo Wen[2], Felipe da Veiga Leprevost[3], Alexey Nesvizhskii[3], Jamie Moon[4], Tao Liu[4], Bing Zhang[1,2]

[1]Center of Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX 77030, USA

[2]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX 77030, USA

[3]Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA

[4]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99354, USA

Introduction

Phosphorylation, one of the most common post-translational modifications (PTMs), is a reversible mechanism that regulates cellular processes such as cell growth, development and aging through protein kinases and phosphatases. Tandem mass spectrometry (MS/MS)-based phosphoproteomics has emerged as a powerful platform for global phosphorylation analysis. Different peptide identification and site localization pipelines can lead to different interpretations of the same dataset, which in turn, affects all downstream analyses. However, it is difficult to compare the performance of these pipelines on complex datasets without proper evaluation metrics. Here, we propose three novel, deep learning-derived metrics, including retention time (RT) differences between observed and predicted RTs (Delta RT), spectral similarities (Spearman's Correlation Coefficient, SPC) between predicted MS/MS spectra and observed spectra, and predicted phosphorylation probabilities, to evaluate the performance of three peptide identification and localization pipelines on multiple CPTAC datasets. In this poster, we report preliminary results from this study.

Methods

We developed three quantitative metrics based on deep learning approaches to systematically evaluate the quality of phospho-peptide identification:

(1) RT differences between observed RTs and the RTs predicted by AutoRT.

(2) Spectral similarities between experimental spectra and the spectra predicted by pDeep2.

(3) Phosphorylation probability for a given site predicted by MusiteDeep.

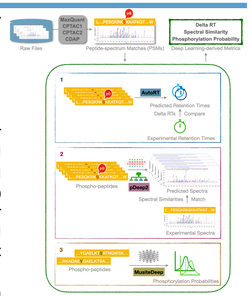

Figure 1. Study design.

Firstly, one large-scale synthetic proteomics dataset (PXD000138) was used to evaluate our benchmarking approach. We used base AutoRT model trained with unmodified peptides and performed first-step transfer learning with modified peptides (PXD015087). Afterwards, we used data from 10 raw files and split them into training data (6198 peptides), test data (689 peptides) and validation data (700 peptides). The training data were used for AutoRT second-step transfer learning and for pDeep2 base model transfer learning. The validation data and wrongly identified PSMs were used as test PSMs to evaluate the ability of Delta RT and SPC to distinguish correct and different types of wrong identifications.

Next, one large-scale phospho-proteomics dataset from CPTAC (TMT-10plex labeled data from >100 uterine cancer samples, UCEC) was used for benchmarking. The dataset including 16 experiments which were searched against RefSeq human protein database using four pipelines, including MaxQuant and three independent CPTAC pipelines, CPTAC1, CPTAC2 and CPTAC 3 pipeline. Search results were filtered based on the recommended cutoffs of each pipeline. One experiment was used for benchmarking. The overlap PSMs among three pipelines (MaxQuant, CPTAC1, CPTAC3) were treated as ground truth and be split into training data (7174 peptides), test data (797 peptides) and validation data (851 peptides). The training data were used for AutoRT second-step transfer learning. The PSMs identified by only one tool (test PSMs) were used for comparison.

Then, since CPTAC1 and MaxQuant outperformed the other two pipelines, and MaxQuant is the most popular tool in recent years, more large-scale phospho-proteomics datasets from different species were searched by MaxQuant and CPTAC1 in order to further compare their performance on different types of datasets. Here we showed the results of PXD015282 (Mouse, TMT). The dataset was searched against Uniprot mouse protein database. Search results were filtered based on the recommended cutoffs of each pipeline. One experiment was used for benchmarking. The overlap PSMs between the two pipelines were treated as ground truth and were split into training data (6244 peptides), test data (694 peptides) and validation data (738 peptides). The training data were used for AutoRT second-step transfer learning. The PSMs identified by only one tool (test PSMs) were used for comparison.

Data

Table 1. Test PSMs from PXD000138

Phosphopepitdes						Unmodified peptides					
Seq	Phos Number	Site	Group Alias	TP/ FP	Peptide num	Seq	Dtscore	Score	Group Alias	TP/ FP	Peptide num
T	T	T	SeqT_NumT_LocT	TP	3939	T	>=6	>=40	SeqT_ScoreH	TP	3696
T	T	F	SeqT_NumT_LocF	FP	488	F	>=6	>=40	SeqF_ScoreH	FP	567
T	T	F	Simulation1_LocF	FP	4499	T			SeqT	TP	5471
T	T	F	Simulation2_LocF (closest)	FP	3252	F			SeqF	FP	4316
T	F	F	SeqT_NumF	FP	240	T	<6	Or <40	SeqT_ScoreL	TP	2404
T	F	F	Simulation3_NumF	FP	587	F	<6	Or <40	SeqF_ScoreL	FP	3897
F			SeqF	FP	7572						

Table 1. Correctly identified and wrongly identified synthetics peptides from PXD000138. Validation peptides and wrongly identified peptides were split into different groups to evaluate the ability of Delta RT and SPC to distinguish wrongly identified or localized peptides. Dtscore and Score are identification scores generated by MaxQuant.

Table 2. UCEC search results from four pipelines

UCEC			
Experiment	Tool	Identified localized phospho-peptides	Identified PSMs
16 Experiments	MaxQuant	82911	784940
	CPTAC1	110739	1175527
	CPTAC2	103775	935065
	CPTAC3	62451	1081290

Table 2. UCEC phospho-proteomics dataset including 16 experiments was searched by MaxQuant, CPTAC1, CPTAC2 and CPTAC3. The results from MaxQuant and CPTAC1 were filtered by phosphorylation localization probability 0.75. CPTAC1 identified the most localized phospho-peptides and PSMs. CPTAC2 ranked second. MaxQuant ranked third, while CPTAC3 identified the least phospho-peptides and PSMs.

Results

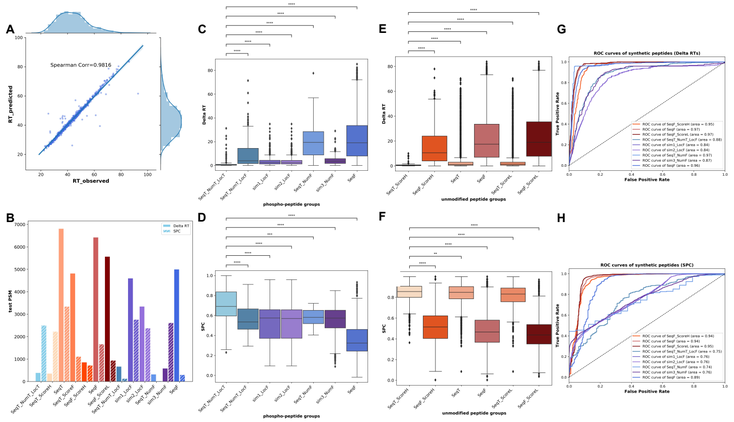

Figure 2. Evaluation of the two deep learning-derived benchmarking metrics. (A) Scatter plot comparing predicted retention time and experimental retention time from Seq_NumT_LocT (validation PSMs). The SPC is 0.9816 showing good performance of AutoRT. (B) Bar plot of test PSMs (validation PSMs and wrongly identified PSMs) used to calculate Delta RT and SPC. (C-D) Box plots of Delta RT and SPC from test PSMs with modified peptides. The peptides were split into correctly identified and localized peptides, correctly identified and wrongly localized peptides, correctly identified peptides with wrong number of phospho-sites, wrongly identified peptides and simulated peptides. Delta RT and SPC were able to distinguish different types of wrong PSMs with modified peptides from correct ones. (E-F) Box plots of Delta RT and SPC from test PSMs with unmodified peptides. The peptides were split into correctly and wrongly identified peptides with different levels of scores. Delta RT and SPC were able to distinguish wrongly identified unmodified PSMs with different scores from correct ones. (G-H) ROC curves of Delta RT and SPC. Delta RT and SPC were able to distinguish correct PSMs and different types of wrong PSMs.

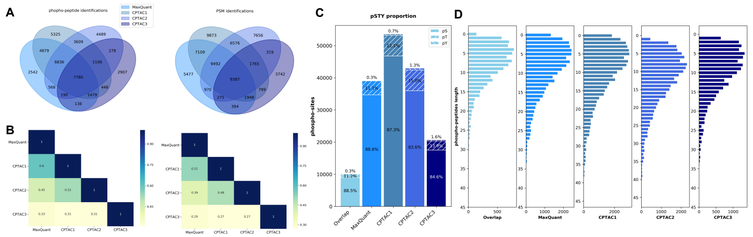

Figure 3. Summary of search results from one UCEC experiment by MaxQuant, CPTAC1, CPTAC2 and CPTAC3. (A-B) Venn plots and Jaccard Index heatmaps of phospho-peptides and PSMs identified by MaxQuant, CPTAC1, CPTAC2 and CDAP from one UCEC experiment. Different tools would have different search and localization results even for the same phospho-proteomics dataset. (C) Bar plot of phosphorylated Serine, Threonine and Tyrosine from each search engine and their overlap. (D) Density plot of phospho-peptide length from each search engine and their overlap.

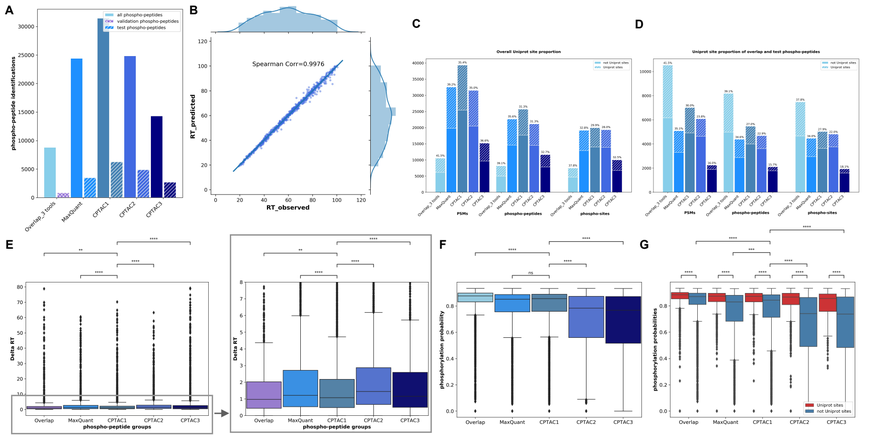

Figure 4. Benchmark four pipelines using one UCEC experiment with Delta RT and phosphorylation probability. (A) Bar plot of all phospho-peptides from overlap (MaxQuant, CPTAC1, CPTAC3) and each tool, validation phospho-peptides from overlap and test phospho-peptides from each tool. Test phospho-peptides from each tool are those identified by only one tool. (B) Scatter plot comparing predicted retention time and experimental retention time from validation data. The spearman correlation coefficient is 0.9976 showing good performance of AutoRT. (C-D) Bar plots of overall Uniprot site proportion and Uniprot site proportion of overlap and test phospho-peptides from each pipeline. (E) Box plots of Delta RT from validation phospho-peptides and test phospho-peptides. Validation data performs the best. CPTAC1 outperforms the other three. (F-G) Box plots of phosphorylation probability from validation phospho-peptides and test phospho-peptides. CPTAC1 outperforms the other three after excluding known Uniprot sites. We excluded Uniprot sites, since MusiteDeep was trained by Uniprot sites.

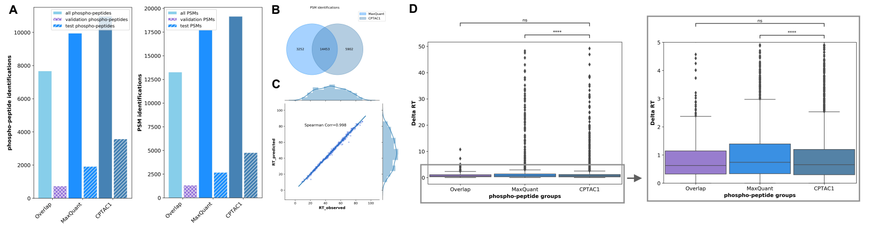

Figure 5. Benchmark of MaxQuant and CPTAC1 using one experiment from PXD015284. (A) Bar plots of all phospho-peptides, validation phospho-peptides, test phospho-peptides, all PSMs, validation PSMs and test PSMs. The test phospho-peptides and test PSMs are those identified by only one tool. (B) Venn plot of PSMs identified by MaxQuant and CPTAC1. (C) Scatter plot comparing predicted retention time and experimental retention time from validation data. The spearman correlation coefficient is 0.998. (D) Box plots of Delta RT from test phospho-peptides and validation phospho-peptides. CPTAC1 outperforms MaxQuant.

Conclusions

The three new deep learning-derived metrics — Delta RT, spectral similarity, phosphorylation probability, can distinguish wrongly identified and localized phospho-peptides which addresses a critical challenge of method evaluation in complex datasets. The novel benchmark method can help guide users when choosing computational pipelines for the analysis of large-scale phosphoproteomics datasets.

References

1. Wen et al. (2020). Cancer neoantigen prioritization through sensitive and reliable proteo-genomics analysis. Nat Commun. 11 (1759)
2. Wang et al. (2020). MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Research*, 48: 140–146
3. Dou et al., (2020). Proteogenomic Characterization of Endometrial Carcinoma. Cell. 180: 729–748
4. Leprevost et al., (2020) Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat Methods. 17(9): 869-870