

# Probabilistically Assigning Sites of Protein Modification with Scaffold PTM

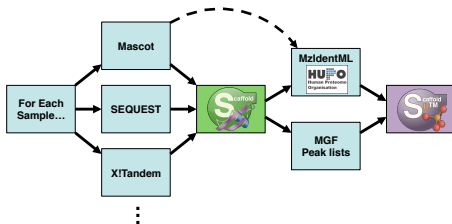
Brian C. Searle and Mark Turner

Proteome Software Inc., 1340 SW Bertha Blvd Suite 10, Portland, OR, 97219-2039, United States  
E-mail: Brian.Searle@ProteomeSoftware.com, Telephone: (503) 244-6027, Fax: (503) 245-4910

## Abstract

Accurate interpretation of MS/MS data containing post-translational modifications (PTMs) such as phosphorylation continues to be a difficult problem for proteomics researchers. One challenge involves knowing which data are reliable for assigning specific amino acid sites of modification and separating that from data that are less conclusive. This is particularly difficult when peptide identifications contain multiple possible sites of modification. Some work towards improving phosphorylation site assignment has been implemented in a software program called Ascore by Beausoleil et al<sup>1</sup>. Ascore is more accurate than traditional database search engines at localizing a modification because it only considers the fragment ladder ions that can differentiate between the two possible sites.

Here we present Scaffold PTM, a program that extends the Ascore algorithm to consider all types of variable PTMs that can be identified with MS/MS. Scaffold PTM also uses overlapping data from several peptides simultaneously to improve confidence in specific site assignments. To accomplish this, Scaffold PTM uses the Ascore algorithm to interpret sites of modification on a peptide-by-peptide basis. Peptide identifications can be read from Scaffold, Mascot, or any database search engine that supports the open community standard format, MzIdentML. Once the data has been collated, Scaffold PTM calculates site localization probabilities from the Ascores, which are peptide independent. Scaffold PTM then collapses the peptide interpretations into protein specific sites of modification, comprised of multiple pieces of overlapping evidence.



**Figure 1:** The Scaffold PTM workflow. Scaffold PTM imports search results through HUPRO standard MzIdentML format and peak lists through the MGF text file format. Although the easiest way to create MzIdentML files for most search engine results is through Scaffold 3, Mascot 2.3 can also generate valid MzIdentML files for individual samples.

## Open Data Formats and Standards

Not only does Scaffold PTM utilize the MzIdentML standard, it also saves data in an open SQLite database that can be browsed and queried independently. Scaffold PTM contains a separate console that allows programmatic access to the Ascore algorithm, among other important features including complete SQL support.

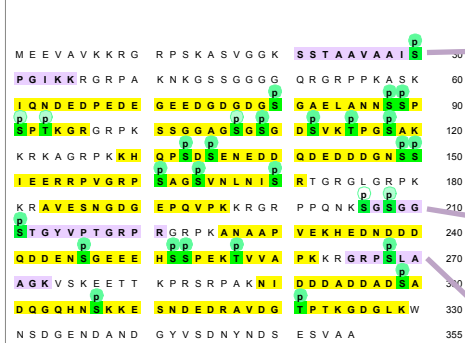
## Acknowledgements

We thank Steven Gygi of Harvard for allowing the use of the Ascore algorithm, for helpful discussions on the details of the algorithm and for making data available for validation and demonstration<sup>2</sup> on Tranche at: <https://proteomecommons.org/dataset.jsp?i=73370>

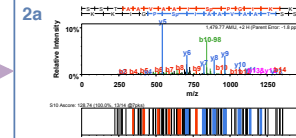
## References

1. Beausoleil, et al., Nat. Biotechnol. 2006, 24, 1285-1292.
2. Zhu, et al., J Proteome Res. 2008, 7, 1675-1682.

**(CPD1\_DROME )** Chromosomal protein D1 OS=Drosophila melanogaster GN=D1 PE=1 SV=3

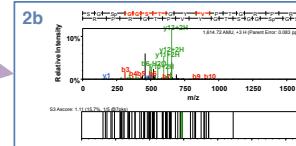


**Figure 2:** A phosphorylated protein (CPD1\_DROME) identified by Mascot, interpreted by Scaffold, with sites localized by Scaffold PTM. Confident localizations are represented as solid circles while unlocalized sites are hollow. Sites with some, but not definite localization confidence are represented as half shaded. Modifications of other types are localized separately and drawn in different colors.



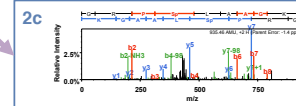
## Confident Localizations

Some spectra contain an overwhelming amount of evidence to assign a particular site. In this case, peaks that distinctly suggest site localization to S10 (over S1, S2 and S3) are drawn in blue and red in the spectrum "barcode".



## Unlocalized Phosphorylations

Not all confidently identified spectra contain enough data to localize a phosphorylation site. In this case there is very little evidence (a single peak assigned to y14+2H) to differentiate S3 from S1, resulting in a low Ascore. While it is clear that one of these sites is phosphorylated, it is impossible to determine which one is correct.



## Undisputed Sites

If there are no neighboring serines, threonines or tyrosines, then a site localization is considered to be undisputed. These sites are automatically assigned with the highest confidence.

## The Ascore Calculation

The Ascore calculation is interesting because it only considers fragment ions that can differentiate between two possible PTM sites. Ascore computes the localization p-value as:

$$P\text{-value}_{\text{unphosphorylated}} = \sum_{i=1}^N p^i (1-p)^{N-i}$$

where  $N$  is the total number of peaks that could possibly differentiate between the sites,  $n$  is the number of those peaks found, and  $p$  is an indication of the probability that a peak could be identified by chance. Scaffold PTM uses a Pascal's triangle optimization to speed up calculation of binomial probabilities. Between this and other optimizations, Scaffold PTM is approximately 1000x faster than the original Ascore implementation.

The actual Ascore score, as described in Beausoleil et al, is made up of the difference between two component p-values:

$$\text{Ascore}(S) = [-10^4 \log(P\text{-value}_{\text{phosphorylated}})] - [-10^4 \log(P\text{-value}_{\text{unphosphorylated}})]$$

Where  $P\text{-value}_{\text{phosphorylated}}$  is the probability the site was assigned by chance and  $P\text{-value}_{\text{unphosphorylated}}$  is the probability that the site next most likely was assigned by chance.

## Calculating PTM Localization Probabilities

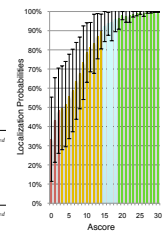
Converting Ascores to probabilities puts site localizations on even footing for comparing across peptide identifications or overlapping evidence. The method we propose here requires more stringent criteria if many possible site configurations exist in a peptide, or if localization peaks can be identified for multiple configurations.

Since  $P\text{-value}_{\text{phosphorylated}}$  and  $P\text{-value}_{\text{unphosphorylated}}$  represent the two states for phosphorylation at that site (present and absent), we can calculate the likelihood the phosphorylation site was assigned by chance as:

$$P(S) = \frac{p - \text{value}_{\text{unphosphorylated}}}{p - \text{value}_{\text{phosphorylated}} + p - \text{value}_{\text{unphosphorylated}}}$$

and the inverse likelihood that the phosphorylation site was actually assigned as:

$$P(S|I) = \frac{p - \text{value}_{\text{unphosphorylated}}}{p - \text{value}_{\text{phosphorylated}} + p - \text{value}_{\text{unphosphorylated}}}$$



Since it is known that  $N$  amino acid sites can be phosphorylated in the peptide (S,T,Y) and the total number of phosphorylations ( $k$ ) can be assigned using the parent ion mass measurement, we can narrow the number of peptide candidates down to  $\binom{N}{k}$  peptide interpretations. For example, if there are three possible sites of phosphorylation, S1, S2, and S3, and two phosphorylations assigned, then the three possible peptide interpretations are S1p-S2p-S3, S1p-S2-S3p, S1-S2p-S3p. We can calculate the likelihood that any of the interpretations exists as:

$$P(I) = \prod_{S \in I} P(S) * \prod_{S \notin I} (1 - P(S))$$

In the case of the interpretation S1p-S2p-S3, for example, we can calculate . We can use the limited list of peptide interpretations to calculate the probability that site  $i$  is phosphorylated as:

$$P(S|I) = \frac{\sum_{I \text{ containing } S} P(I)}{\sum_{I} P(I)}$$

Localization probabilities are mirrored, so 100% indicates confident localization, 0% indicates negative localization (definitely not here), and 50% indicates inconclusive data. Probabilities above 99% suggest that the site is confidently phosphorylated and probabilities above 95% suggest that strong evidence is present.