

Modification Site Localization Scoring Integrated into a Search Engine

Peter R. Baker¹, Jonathan C. Trinidad¹,
Katalin F. Medzihradzsky¹, Alma L. Burlingame¹
and Robert J. Chalkley¹

¹ Mass Spectrometry Facility, Dept. of Pharmaceutical Chemistry, University of California, San Francisco, USA

Introduction

Large proteomic datasets identifying thousands of modified peptides are becoming increasingly common in the literature. However, tools for measuring the confidence of modification site assignments are sparse and are not often employed. A few tools for estimating phosphorylation site assignment reliabilities have been developed^{1,2}, but these are not integral to a search engine, so require a particular search engine output for a second step of processing. They may also require use of a particular fragmentation method and are mostly only applicable for phosphorylation analysis, rather than PTM analysis in general. In this study, we present the performance of site assignment scoring that is directly integrated into the search engine Protein Prospector, which allows site assignment reliability to be automatically reported for all modifications present in an identified peptide. It clearly indicates when a site assignment is ambiguous (and if so, between which residues), and reports an assignment score that can be translated into a reliability measure for individual site assignments. This work has recently been published³.

Site Assignment Software

For the top hits saved for each spectrum Protein Prospector's Batch-Tag now calculates the scores for all permutations of the site assignments. The next best hits with different site assignments are saved.

E-values are calculated for each saved hit and converted into $-10\log_{10}E$. The difference in these log e-values is then reported for each site assignment as a SLIP (Site Localization In Peptide) score.

Example

Peptide Sequence	Score	E-Value	$-10\log_{10}E$
PET(Phospho)PPRQSHSGSIS(Phospho)PYPK	57.6	3.9×10^{-7}	64
PET(Phospho)PPRQSHSGS(Phospho)ISPYPK	51.4	4.0×10^{-6}	54
PETPPRQS(Phospho)HSGSIS(Phospho)PYPK	39.3	1.6×10^{-4}	38

Phospho@14 has a $-10\log_{10}E$ difference of (64-54) giving a SLIP score of 10

Phospho@3 has a $-10\log_{10}E$ difference of (64-38) giving a SLIP score of 26

This would be written as Phospho@3=26;Phospho@14=10

A SLIP score of 10 corresponds to an order of magnitude difference in probability scores.

Reporting Site Assignments

A score threshold can be defined below which a site is defined as ambiguous. Eg:

ES(Phospho)KSSPRPTAEK
SATTTPS(Phospho)GSPR
SS(Phospho)SFREM(Oxidation)ENQPHK
GRRS(Phospho)PS(Phospho)PGNSPSGR
Y(14.0067)IGVGER
RPS(Phospho)QDGRST(Phospho)PVYNK
QQSHFAMMHGGTGFAGIDSSSPEVK

Phospho@2=18
Phospho@5|7|9
Oxidation@7;Phospho@2|3
Phospho@4=26;Phospho@6=20
14.0067@N term|1
Phospho@3&8|3&9|8&9
Phospho@3&Oxidation@7|Phospho@3&Oxidation@8

If the site is not ambiguous a SLIP score is given.

The ‘|’ character (meaning or) is used to signify an ambiguity.

Multiple modified sites are separated by a ‘;’ character.

No SLIP score is given if there is only 1 possible site.

The ‘&’ (meaning and) character is used if there is ambiguity across multiple modified sites.

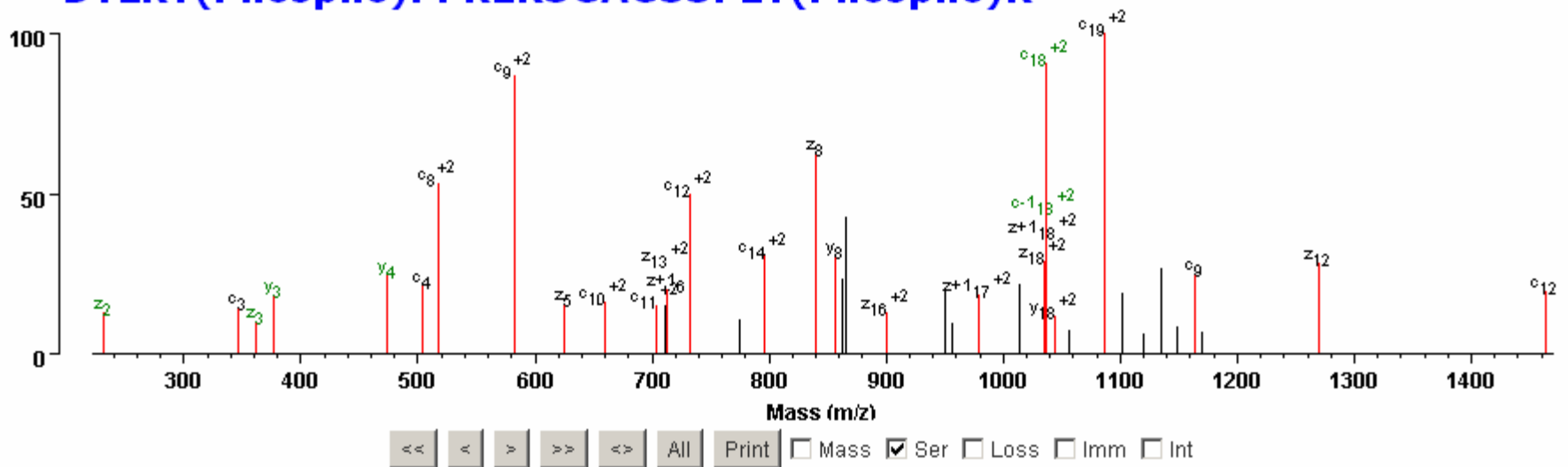
Analysis of Ambiguous Sites

If the site is ambiguous, all possible peptide sequences can be passed to MS-Product for manual analysis.

- No limit on number of sequences that can be passed
- Up to 6 sequences can be compared on color coded plot
- Sequences can be considered as alternatives or mixtures
- Discriminating ions have color coded labels
- Raw data can be shown on the same plot as the peak list
- The peak list can be plotted with different peak densities

Analysis of Ambiguous Sites

DTLRT(Phospho)PPRERSGAGSS(Phospho)PETK⁺⁴
DTLRT(Phospho)PPRERSGAGSSPET(Phospho)K⁺⁴



Max Intensity: 141592

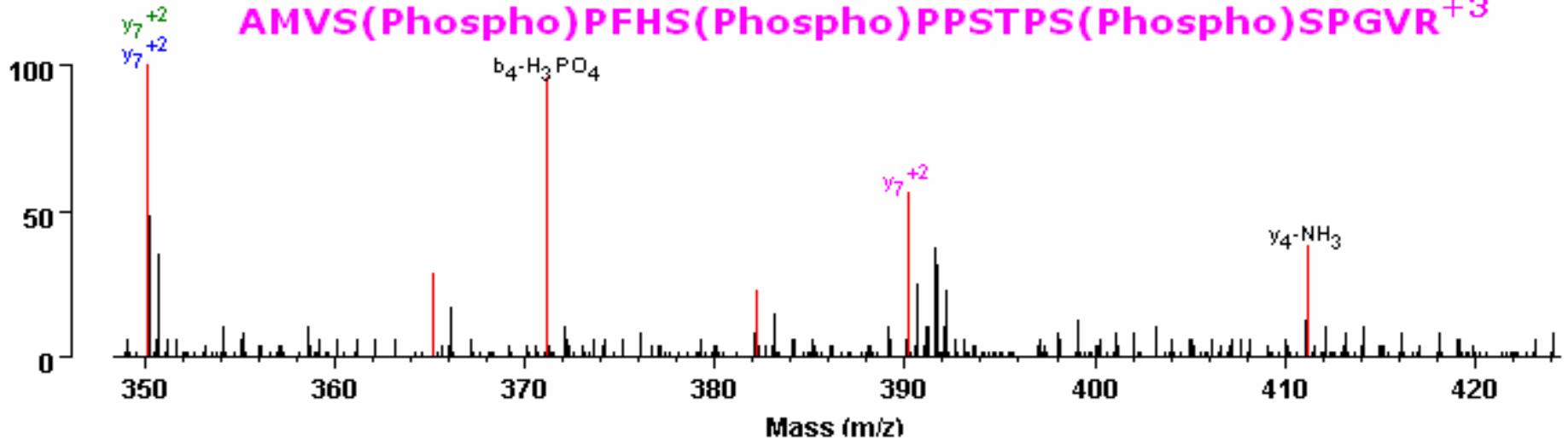
Num Matched: 25/50 (50.0% unmatched) Matched Intensity: 68.2% Matched Series Intensity: 68.2%

Num Matched: 20/50 (60.0% unmatched) Matched Intensity: 55.0% Matched Series Intensity: 55.0%

Raw Data Display

For some instruments it is possible to overlay the raw and centroid data in MS-Product to help with manual site assignment. For example it is possible to check charge states if the centroid list has been deisotoped.

AMVS(Phospho)PFHS(Phospho)PPS(Phospho)TPSSPGVR⁺³
AMVS(Phospho)PFHS(Phospho)PPST(Phospho)PSSPGVR⁺³
AMVS(Phospho)PFHS(Phospho)PPSTPS(Phospho)SPGVR⁺³



Protein Based SLIP Scores

Sites may also be reported relative to their location within the protein.

Peptide Sequence	Peptide SLIP Scores	Protein SLIP Scores
SAPSPTSGPGGASHDTDFR	HexNAc@1=9;HexNAc@6=7;HexNAc@13 16	HexNAc@683=9;HexNAc@688=7;HexNAc@695 698
SAPSPTSGPGGASHDTDFR	HexNAc@1=8;HexNAc@6=6	HexNAc@683=8;HexNAc@688=6
SPTSGPGGASHDTDFR	HexNAc@1 3	HexNAc@686 688
PTSGPGGASHDTDFR	HexNAc@2 3	HexNAc@688 689
TSGPGGASHDTDFR	HexNAc@1=30;HexNAc@2=28	HexNAc@688=30;HexNAc@689=28
TSGPGGASHDTDFR	HexNAc@8=24;HexNAc@1 2	HexNAc@695=24;HexNAc@688 689
TSGPGGASHDTDFR	HexNAc@11=28;HexNAc@1 2	HexNAc@698=28;HexNAc@688 689
TSGPGGASHDTDFR	HexNAc@1 2	HexNAc@688 689
RIKGTTPALPFAPVQAPSVILPLPGQSVDR	HexNAc@5=41;HexNAc@6=33;HexNAc@8=33	HexNAc@705=41;HexNAc@706=33;HexNAc@708=33
RIKGTTPALPFAPVQAPSVILPLPGQSVDR	HexNAc@5&6 5&8 6&8	HexNAc@705&706 705&708 706&708
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@4=60;HexNAc@5=52;HexNAc@7=52	HexNAc@705=60;HexNAc@706=52;HexNAc@708=52
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@4=44;HexNAc@5=36;HexNAc@7=36	HexNAc@705=44;HexNAc@706=36;HexNAc@708=36
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@4&5 4&7 5&7	HexNAc@705&706 705&708 706&708
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@4&5 4&7 5&7	HexNAc@705&706 705&708 706&708
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@4=8;HexNAc@7=8	HexNAc@705=8;HexNAc@708=8
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@7=8;HexNAc@4 5	HexNAc@708=8;HexNAc@705 706
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@5=7;HexNAc@7=13	HexNAc@706=7;HexNAc@708=13
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@5=7;HexNAc@7=14	HexNAc@706=7;HexNAc@708=14
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@5 7	HexNAc@706 708
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@5 7	HexNAc@706 708
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@7=8	HexNAc@708=8
IKGTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@7=6	HexNAc@708=6
GTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@2=55;HexNAc@3=55;HexNAc@5=48	HexNAc@705=55;HexNAc@706=55;HexNAc@708=48
GTTPTALPFAPVQAPSVILPLPGQSVDR	HexNAc@2=7;HexNAc@3 5	HexNAc@705=7;HexNAc@706 708

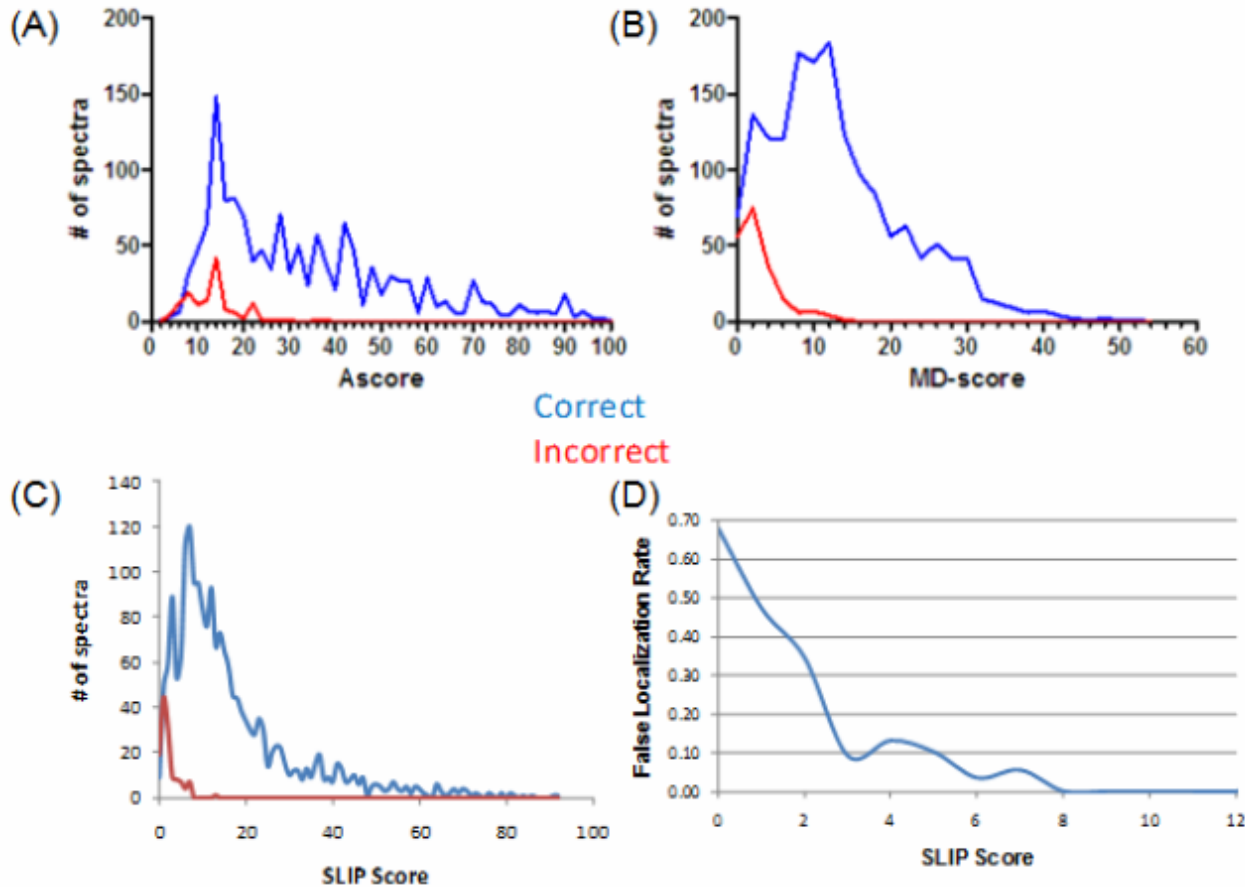
Comparison to Ascore and Mascot Delta Score

A previous study² compared the Mascot Delta Score to the phosphorylation site scoring software Ascore¹. The results of using SLIP scoring on the same QTOF micro data set of synthetic phosphopeptides are given below.

	SLIP	Ascore	MDS
Spectra	2334		
Phosphosites	2437	1584	1840
Correct	1924		
Incorrect	130	138	201
Ambiguous	220		
One Possible Site	164		
FLR (%)	6.3	9.0	11.0
Ambiguous (%)	9.0		

In the table a site is called ambiguous if different sites gave exactly the same score. FLR stands for false localization rate.

Ascore/MD-Score Comparison



A, B, C: Score histograms for correct and incorrect site assignments for Ascore, MD-score and SLIP Score.

D: FLR against SLIP score.

Testing SLIP on a Larger Scale

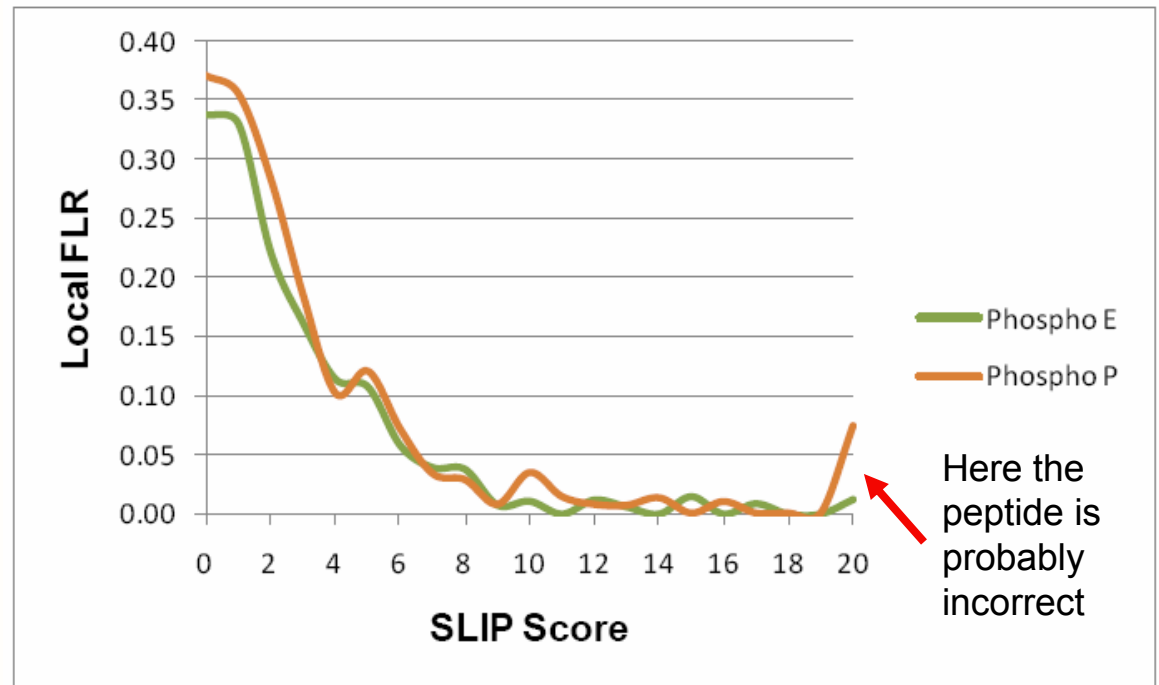
To test SLIP scoring on a larger scale we temporarily changed the Batch-Tag settings so it would consider phosphorylation and loss of phosphoric acid from Pro and Glu residues.

A large phosphopeptide dataset was acquired using ion trap CID fragmentation on an LTQ-Orbitrap VELOS. A first search was performed considering phosphorylation of S, T, Y and P and a second considering S, T, Y and E. For each search over 90000 spectra were identified at 0.1% FDR against a concatenated normal/random database. 60000 of these were phosphopeptides.

Results

The site assignment is known for spectra with only a single S, T or Y in the peptide. In a proportion of cases the site is incorrectly reported as a decoy residue allowing a global FLR to be calculated (table). Using a SLIP score threshold (graph) decreases the FLR as more sites are reported as ambiguous.

	STYP search	STYE search
Peptide IDs	5433	5415
Decoy site scores best	361	245
% Global FLR	6.6	4.5



Conclusions

- Site localization capability built into search engine without significant overhead.
- Works for all post-translational modifications.
- Outperforms Ascore and Mascot delta score.
- Adjustable score threshold for reporting ambiguous sites.
- Ambiguities may be viewed and assessed manually in a spectral viewer.
- Raw and centroid data can both be displayed in spectral viewer.
- Site localization scores are reported for all ambiguous sites.
- Sites can be reported relative to either the peptide or the protein.

Acknowledgements

This work was supported by NIH NCRR grant RR001614, SIG RR019934 and the Biotechnology and Biological Sciences Research Council of the UK. We also thank the Vincent Coates Foundation for support.

This software is freely available on the web at <http://prospector.ucsf.edu>

References

- Beausoleil SA, Villén J, Gerber SA, Rush J, Gygi SP (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. **Nat Biotechnol.** 24:1285-1292
- Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, Bantscheff M, Kuster B (2011) Confident phosphorylation site localization using the Mascot Delta Score. **Mol Cell Proteomics**, 10: M110.003830
- Baker PR, Trinidad JC, Chalkley RJ (2011) Modification Site Localization Scoring Integrated into a Search Engine. **Mol Cell Proteomics**, doi: 10.1074/mcp.M111.008078