Computing Information Content of PTM Site Assignments

David D. Shteynberg[1], Eric W. Deutsch[1], David S. Campbell[1], Michael R. Hoopmann[1], Ulrike Kusebauch[1], Dave Lee[2], Luis Mendoza[1], Zhi Sun[1], Anthony Whetton[2], and Robert L. Moritz[1]

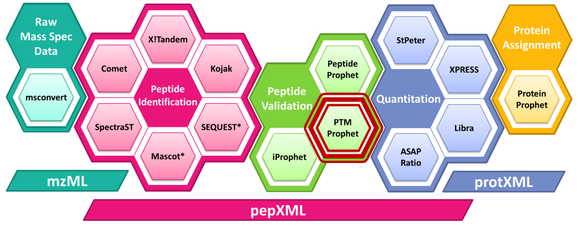
[1] Institute for Systems Biology, Seattle, WA, 98008, USA

[2] University of Manchester, Manchester, M13 9PL, UK

Institute for Systems Biology

Overview

- Search algorithms are good at assigning the peptide sequence, but not necessarily at determining the correct positions of PTMs contained in the peptide
- **PTMProphet** part of the Trans-Proteomic Pipeline (TPP) evaluates all modifications possible given user settings and the assigned peptide sequence
  - Current TPP version 5.2 is available at **www.tppms.org**

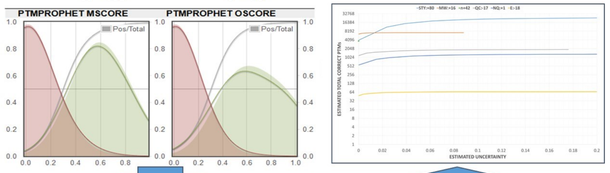


PTMProphet Method

For each modified PSM:

1. Compute accurate probabilities of each potential modification site being modified
2. Compute information content statistics, thereby allowing comparison of PSMs having different numbers of potential modification sites and number of modifications

$$Mscore = \frac{M^m}{M^m+M^u} \quad , \quad Oscore = \frac{O^m}{O^m+O^u}$$



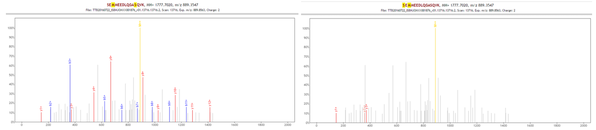
$$\frac{Pr(Mscore, Oscore|Modified)Pr(Modified)}{Pr(Mscore, Oscore)}$$

$$H_t^{norm} = -\frac{1}{m}\sum_{i=1}^{s} p_i \log_{s/m} p_i$$

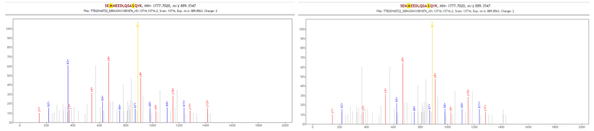
PTM Localization

Simple Peptide Example: One Oxidation, One Phosphorylation

SEMMEEDLQGASQVK vs SEMMEEDLQGASQVK



SEMMEEDLQGASQVK vs SEMMEEDLQGASQVK



1. For each PSM, evaluate all possibilities of peptide modification
2. Sum matched peak intensities: $\Psi(P)$ for peptide P
3. Compute Ψ for each peptide possibility
4. For each potential PTM site **s** on the peptide, compute:
   - P^{mod} = argmax(∀ P *with site* **s** *modified* | Ψ(P))
   - P^{unmod} = argmax(∀ P *with site* **s** *unmodified* | Ψ(P))
5. Compute "common" matched peak intensity: $C(P^{mod}, P^{unmod})$
6. Compute discretized observed maximum noncommon intensities: $O^m = \frac{\Psi(P^{mod}) - C(P^{mod}, P^{unmod})}{i}$ and $O^u = \frac{\Psi(P^{unmod}) - C(P^{mod}, P^{unmod})}{i}$
7. Compute observed maximum noncommon matched peaks: M^m and M^u
8. Compute probability for each potential PTM site
9. EM > 0 - apply expectation / maximization algorithm until probabilities remain constant
10. Normalize all probabilities by the number of modifications in the peptide
11. Record the output in pepXML

S(0.000)EM(1.000)M(0.000)EEDLQGAS(1.000)QVK

Information Content

PROBLEM: Site Probabilities May Not Be Comparable

- Different numbers of potentially modified sites in different peptides
  - SEMMEEDLQGASQVK (2 phospho sites)
  - SESSEEDLQGASQVK (4 phospho sites)
- Different numbers of modifications in different peptides
  - pSEpSEEDLQGApSQVK (3 phospho mods)
  - pSESSEEDLQGASQVK (1 phospho mod)

H_t^{norm}: Multiple Modification & Site Normalized Shannon's Entropy

Quantifies the amount of information stored in the PTM site assignment for a peptide with **s** modification sites and ***m*** modifications

- **s** modification sites with probabilities $p_{1,\ldots}, p_s$ of being modified

$$H_t^{norm} = -\frac{1}{m}\sum_{i=1}^{s} p_i \log_{s/m} p_i$$

H_t^{norm} range: *[0, 1]*

M_t: Localized Modifications Estimate

Estimates the number of modifications confidently localized that can be used to directly compare PSMs containing ***m*** modification

$$M_t = m - H_t$$

where, $H_t = -\sum_{i=1}^{s} p_i \log_{s/m} p_i$

M_t range: *[0, m]* the higher the score the greater the number of modifications localized in a PSM with ***m*** modifications

I_t: Normalized Per-Modification Information Content

Estimates the per-modification localization certainty that can be used to directly compare PSMs with different number of modifications

$$I_t = 1 - H_t^{norm}$$

where, H_t^{norm} is the normalized per modification entropy of modification of type ***t***

I_t range: *[0, 1]* the higher the score higher the localization certainty

B_t: Mean Best Probability Statistic

Easy to compute and works well in practice

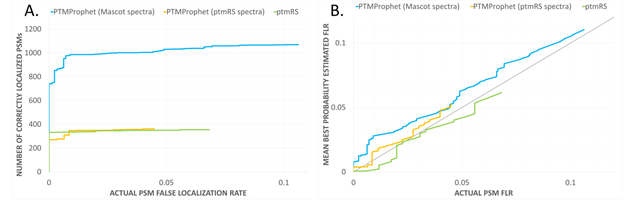
Should be considered in the context of Information Content

$$B_t = \frac{\max_{\{i_1,\ldots,i_m \in 1,\ldots,s\}} \sum_{j=1}^{m} p_{i_j}}{m}$$

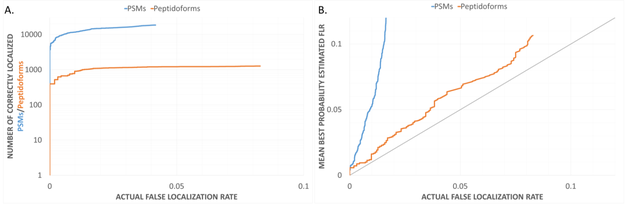
Results

- **Dataset 1:** a small reference dataset of synthetic phosphopeptides previously published (Ferries et al. Proteome Res. 2017) and used to evaluate confident site localization, measured with an Orbitrap Fusion Tribrid mass spectrometer
  - 175 distinct peptide sequences with 191 phosphorylation sites
- **Dataset 2:** a large in-house dataset of synthetic peptides with known phosphorylated sites, measured with a TripleTOF® 5600+
  - 1,342 chemically synthesized phosphopeptides with 5,329 potentially phosphorylated S, T and Y residues
- **Dataset 3:** a phospho-enriched cell lysate dataset (Nyman et al. J Proteomics. 2018) known to contain many mass modifications, measured with a Q Exactive
  - human macrophage cells infected with influenza A virus.

Dataset 1 Peptide Results: Mascot → PTMProphet vs Mascot → ptmRS



Dataset 2 Peptide Results: Mascot+X!Tandem → PTMProphet



Support & Information

Support provided by:

- NIH, NIGMS grants: R01GM087221, R24GM127667, P50GM076547
- National Institute of Allergy and Infectious Diseases grant: R21AI133335
- National Institute of Biomedical Imaging and Bioengineering grant U54EB020406
- NHLBI grant: R01HL133135
- Medical Research Council (ADW)
- Cancer Research UK major centre award (20761)

- **TPP: PTMProphet Resources**
  - **www.tppms.org/tools/ptm/**