Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

A graph-based approach for modification site assignment in proteomics

Dafni Skiadopoulou [1,2], Lukas Käll [3], and Marc Vaudel [1,2,4,*]

[1]Mohn Center for Diabetes Precision Medicine, Department of Clinical Sciences, University of Bergen, Bergen, Norway, and
[2]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway,
[3]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Stockholm, Sweden, and
[4]Department of Genetics and Bioinformatics, Health Data and Digitalization, Norwegian Institute of Public Health, Oslo, Norway.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Tandem mass spectrometry (MS/MS) has become the reference approach for the identification of proteins from a sequence database. Proteins can be modified, which alters their function, and can be detected by mass shifts at specific sites of the amino acid sequence. While the identification of modified proteins can be achieved by considering the modification of every possible acceptor site, the exponential increase of the search space makes the accurate assignment of modifications to sites intractable for highly modified proteins like histones. Here, a graph modeling approach is presented that considers the possibilities of modifications occurring at specific sites of the amino acid chains, and returns the most likely modification sites with controlled processing time. The modification assignment problem is reduced to the Maximum Weight Matching (MWM) problem in bipartite graphs, using modification localization probabilities as weights.

Availability and Implementation:

Contact: name@bio.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Tandem mass spectrometry (MS/MS) has become the reference tool for the identification of proteins from a sequence database. For that, the observed spectra of peptide fragments are matched against theoretical spectra to determine the proteins that are expressed in a biological mixture. This process becomes much more difficult when it comes to modified proteins because even a small change in the amino acids sequence can lead to a significant shift in the peaks of the corresponding spectrum. Such changes can be the product of post-translational modifications (PTMs) or mutations of the proteins which also result in alterations of the protein's functionality. Except of finding the types of the existing PTMs in a protein, another crucial aspect of the identification process is to determine the exact sites in the amino acid chain where the modifications are located.

This problem of modification site localization has been addressed by several tools starting from the Ascore (Beausoleil *et al*., 2006) that was mainly used for phosphorylation localization, which uses only site-determining fragment ions present in an MS/MS spectrum to create a probability-based score that a site is correctly localized. **..... continue with past related work overview .....**

2 Past-related work

The Ascore (Beausoleil *et al*., 2006) metric uses only site-determining fragment ions present in an MS/MS spectrum to create a probability-based score that a site is correctly localized.

ModLS (Trudgian *et al*., 2012) calculates the PTM score (proposed in (Olsen *et al*., 2006)) of each localization combination and then for the top-ranked localization the final value is the difference in PTM Score between this and the second ranked.

In the work of (Yang *et al*., 2018) a semi-supervised SVM model is used to estimate the confidence of each amino acid obtained by de novo

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

peptide sequencing which is also proved to yield accurate results for the modification site localization problem.

A Site Localization In Peptide (SLIP) scoring is proposed by (Baker *et al*., 2011) that is automatically calculated for all modifications in peptides identified by the search engine Batch-Tag in the Protein Prospector suite of tools.

Using mass accuracy and peak intensities, LuciPHOr tool (Fermin *et al*., 2013) improves site localization and false localization rate (FLR) estimation. It estimates FLR based on a target-decoy framework, in which artificial phosphorylation is used to generate decoy phosphopeptides to compare with target matches from a database search.

In the work of (Yu *et al*., 2020) the search engine MSFragger is extended with a shifted ion indexing strategy that enables localization-aware open search leading to an increased number of identified PSMs and an improvement in accuracy at the detection of single amino acid substitutions.

PTM-Shepherd (**?**) manages to determine which PSMs contain a specific mass shift and characterizes PTM profiles in open searches using attributes such as amino acid localization, fragmentation spectra similarity and retention time.

The PTMiner (**?**) tool can be used to assist in localizing the potential mass modifications discovered during open modification searches (....have to describe how this is done....).

3 Approach

In this work the modification site localization problem is addressed by a graph theoretical approach. For each studied peptide a weighted bipartite graph ($G = \{V, E, w\}$) is used to model all possible combinations of modifications on the amino acid chain. In this graph we determine two kinds of vertices, the ones that represent the different modifications that have occured in the peptide (ie $D = \{d_1, d_2, \ldots, d_k\}$) and the ones that represent all their possible acceptor sites in the amino acid chain (ie $A = \{a_1, a_2, \ldots, a_n\}$). The set of the graph's vertices V is then formed by the union of the sets A and D (ie $V = A \cup D$). For each modification, an edge is formed between the corresponding vertex and the ones that represent its possible acceptor sites. Moreover, a weight is assigned to each edge, taking the value of the confidence score (**is this really a confidence score???**, have to better describe the scores we're using...) which represents the probability of the modification to have occurred in that specific site of the amino acid chain. An example of a graph model for a modified peptide is presented in Fig. 1

Using this graph modeling we can reduce the modification site localization problem to the maximum weight matching problem on the resulting graph. This consists of finding a set of edges that are pairwise non-adjacent (without common vertices), in which the sum of weights is maximized. Solving this problem in the graph model of our application will result in the best combination of modifications localization on the peptide, based on their confidence scores.

4 Methods

(mention the usage of python's networkX library implementation of the maximum weight matching algorithm based on methods invented by Jack Edmonds 1965 **???**)

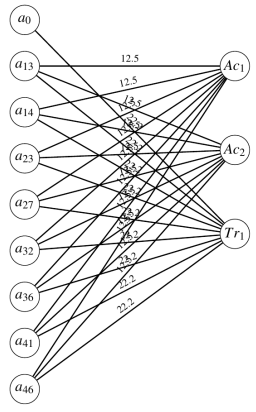

Fig. 1. Graph representation of a peptide with 2 Acetylations and 1 Trimethylation with all their possible acceptor sites.

5 Results

6 Conclusion

Funding

References

Yu,F.T. *et al*. (2020) Identification of modified peptides using localization-aware open search. *Nat Commun.*, **11** (1), 4065.

Yang,H. *et al*. (2018) pSite: Amino Acid Confidence Evaluation for Quality Control of De Novo Peptide Sequencing and Modification Site Localization *Journal of Proteome Research*, **17** (1), 119-128.

Baker,P.R. *et al*. (2011) Modification site localization scoring integrated into a search engine. *Mol. Cell. Proteomics.*, **10**, (M111. 008078)

Fermin,D. *et al*. (2013) LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. *Mol Cell Proteomics*, **12** (11), 3409-19.

Trudgian,D. *et al*. (2012) ModLS: Post-translational modification localization scoring with automatic specificity expansion. *J. Proteomics Bioinform.*, **5**, 283–289.

Beausoleil,S.A. *et al*. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.*, **24**, 1285–1292.

Olsen,J.V. *et al*. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell.*, **127** (3), 635-48.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture