Modified Uniprot Flat File Helps in Rapid Identification and Discovery of Post-Translational Modifications with ProSightPC

Shadab Ahmad[1]; Amol Prakash[1]; David Sarracino [1]; Bryan Krastins[1]; Maryann Vogelsang[1]; Jennifer Sutton[1]; Michael Athanas[1]; Alejandra Garces[1];
Victoria Lunyak[2]; Benjamin Blackwell[2]; Mary F Lopez[1]

[1]BRIMS, Thermo Fisher Scientific, Cambridge, MA; [2]Buck Institute for Age Research, Novato, CA

Overview

Purpose: To develop a strategy for high throughput, automated discovery and identification of novel post-translational modifications (PTMs) using ProSightPC in a top-down approach.

Methods: Modified Uniprot database with additional post-translational modifications was created in order to identify novel targeted Post-translational modifications (PTMs) for H2A3 histone proteins in a high throughput automated fashion. Undigested, intact mesenchymal stem cells that were derived from human adipose tissue were subjected to CID fragmentation in a LTQ Orbitrap Velos instrument, and the top-down data were analyzed with ProSightPC using the modified Uniprot database.

Results: Modified uniprot flat file for histone H2A3 protein helps in identification and discovery of various PTM on human mesenchymal stem cells H2A3 protein; several of these PTMs were not there in the original uniprot flat file.

Introduction

Post-translational modifications are a central theme in the regulation of gene expression. A growing list of modifications confirms that they play a fundamental role in cellular differentiation, cell signaling, regulation of gene expression etc [1]. Thus protein PTM identification and characterization is important in order to understand various biological processes. Mass spectrometry-based top-down proteomics approach is currently the method of choice for the identification and characterization of PTMs. The strategy involves direct fragmentation of intact protein in high resolution mass spectrometer and subsequent analysis of the fragmented backbone of the protein. Most of the software available for top down analysis requires high skilled manual input for accurate identification and characterization of PTMs. More over the task become even more challenging when multiple PTMs are present on a single protein. ProSightPC is a state of art software that effectively supports high-mass-accuracy MS/MS experiments performed on LTQ FT and LTQ Orbitrap instruments which is inevitable for PTM identification. ProSightPC generates proteome database as well as gathers information regarding intact protein sequences along with information about all known PTMs from database (uniprot). It calculates all possible combinations of known modifications (including variations and PTMs) and can identify these known PTMs in high throughput mode. However it fails to identify those PTMs that are not present in the uniprot database for a given protein. Modified uniprot database for targeted proteins can solve this problem and make ProSightPC capable to identify and characterize novel sight and location of known PTMs as well as novel PTM in high throughput fashion. In order to test this concept we created a modified uniprot database for H2A3 histone proteins in order to identify novel as well as known PTMs on human mesenchymal stem cells H2A3 protein.

Methods

Database Creation

The uniprot database contains well revised but limited PTM information, it often does not contain newly reported PTMs. More over if one wants to look for a novel PTM on a protein one is compelled to search it by looking and analyzing mass shift of fragment ions of intact protein. In order to solve this problem we took a novel approach and created a modified uniprot database (uniprot flat file) for H2A3 histone proteins. A flat file for a protein contains all information regarding that protein present in the uniprot knowledgebase including PTMs. We downloaded the flat file for human histone H2A3 protein (Q7L7L0) from the uniprot database (http://www.uniprot.org) which contains information regarding PTMs including acetylation at serine and lysine; phosphorylation at serine and threonine; Citrullination and Symmetric dimethylarginine at various positions. The file does not contain several other reported PTM information such as methylation, crotonylation, hydroxilation and formylation [2] therefore we have included these PTMs in the flat file. In addition to that we also included mono, di and tri-methylation to every possible lysine on the protein sequence in order to discover novel lysine methylation sites on the protein. A new proteomics database (ProSightPC warehouse) was created with the modified uniprot H2A3 flat file in order to search experimental data against in ProSightPC. Top down forward database was created with and without initial methionine for this 130 amino acid long protein. A general workflow for PTMs analysis through modified uniprot flat file is shown in figure 1.

Sample Preparation and LC-MS

Mesenchymal stem cells were derived from human adipose tissue and the cultured cell samples were lysed. The prepared samples containing intact proteins were injected onto a Proxeon Easy nLC system configured with a PLRPS trap (100um x 10cm) and PLRPS analytical column (100um x 25cm). Undigested, intact mesenchymal stem cells proteins were run for 120min with 0-40% acetonitrile gradient in 0.1% formic acid and subjected to CID fragmentation in LTQ Orbitrap Velos instrument which is coupled to the above said nano LC system.

Data Analysis

Thermo Scientific ProSightPC 2.0 software was used to analyze data with modified uniprot flat file through high throughput wizard. High throughput logical tree provided in the software was used for running absolute mass search and biomarker search (figure 2). A biomarker search identifies any subsequence of a protein that has an observed intact ion mass so as to facilitate identification of truncated targeted proteins (if present). Search was set for monoisotopic mass considering all PTMs from the warehouse that was formed from the modified flat file for H2A3 protein. As we don't want to miss those protein whose mass gets shifted by 1-2 Dalton by commonly occurring modifications (such as citrullination at one or more amino acid), we keep relaxed mass tolerance search window of 2.2 Dalton.

Results

Top-down analysis of H2A3 proteins from stem cell lysate using ProSightPC and the modified database aided in identification of various PTMs on histone H2A3. Several of these were not present in the uniprot database; therefore we were not able to identify those modifications using the original uniprot flat file for human H2A3 protein with same search criteria. Two lysine methylation sights have been identified on the H2A3 proteins from stem cell using modified file (figure 3A and 3B); this modification is not present in original uniprot flat file for H2A3 protein.

FIGURE 1. General Workflow for post-translational modification analysis using modified uniprot database

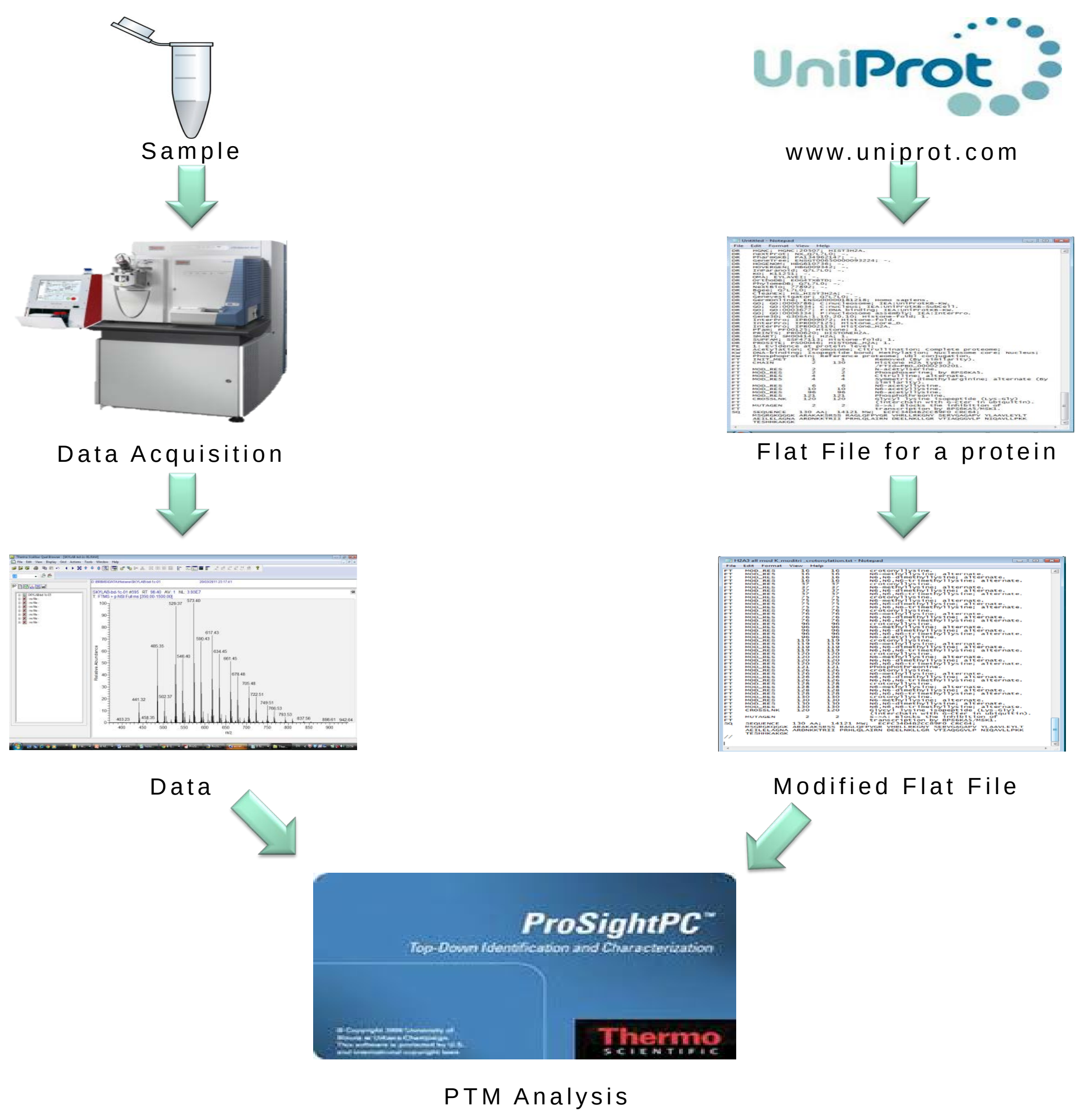
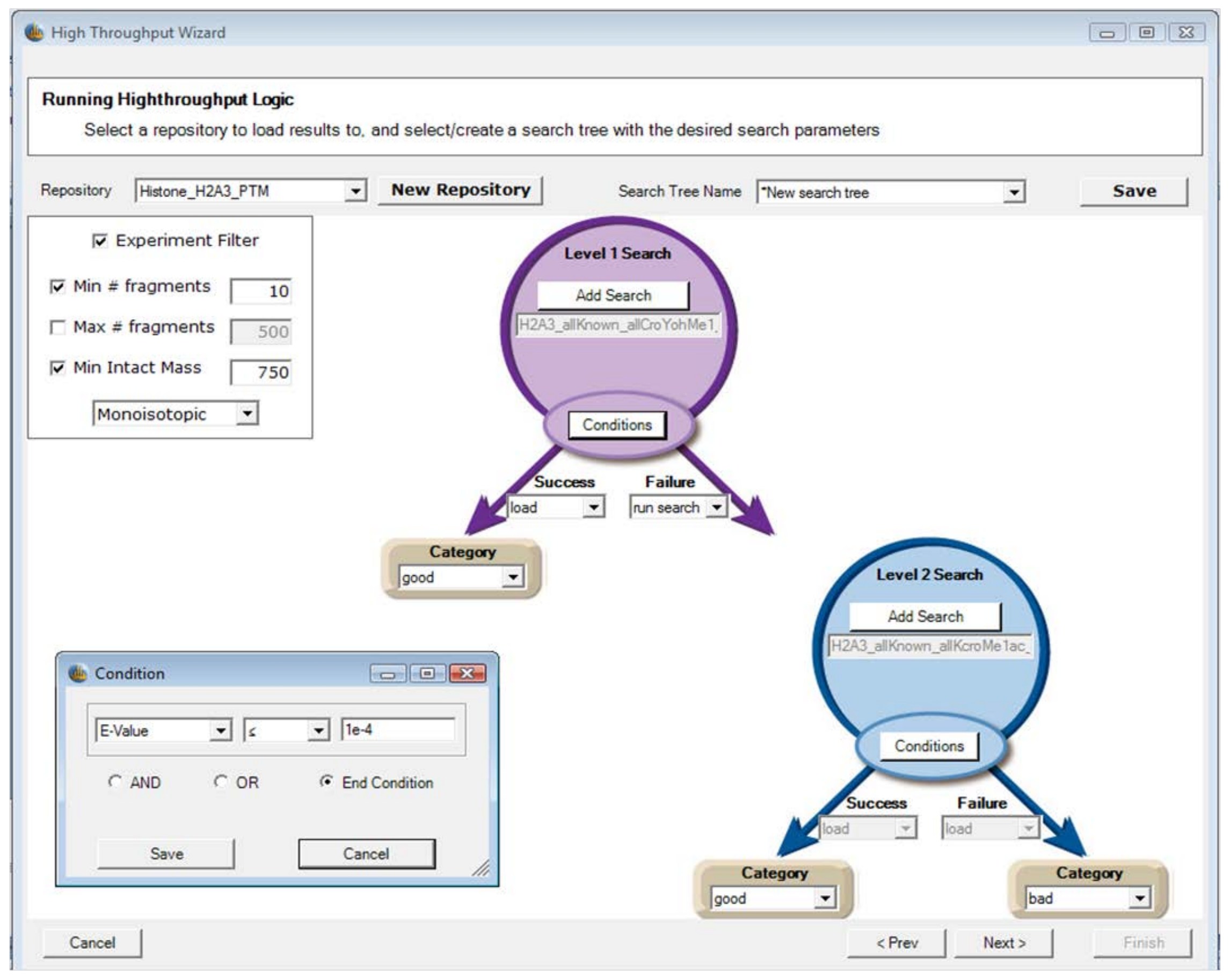


FIGURE 1. ProSightPC high throughput logic work flow



High Throughput Wizard

Running Highthroughput Logic

Select a repository to load results to, and select/create a search tree with the desired search parameters

Repository: Histone_H2A3_PTM | New Repository | Search Tree Name: *New search tree | Save

☑ Experiment Filter
☑ Min # fragments: 10
☐ Max # fragments: 500
☑ Min Intact Mass: 750
Monoisotopic

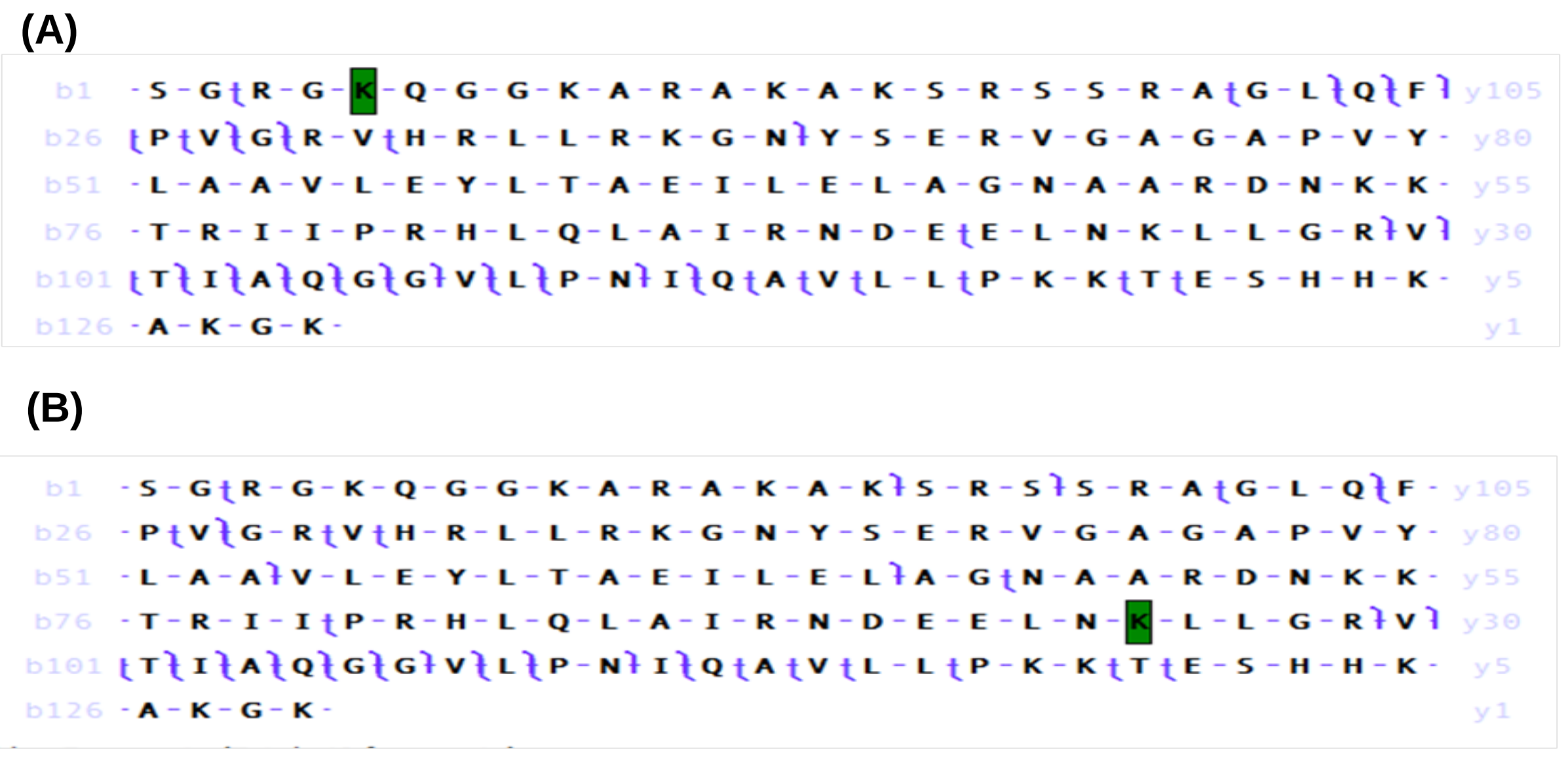
Level 1 Search — Add Search — H2A3_allKnown_allCroYohMe1_
Conditions — Success: load; Failure: run search
Category: good

Level 2 Search — Add Search — H2A3_allKnown_allKcroMe1ac_
Conditions — Success: load; Failure: load
Category: good; Category: bad

Condition: E-Value ≤ 1e-4
○ AND ○ OR ● End Condition
Save | Cancel

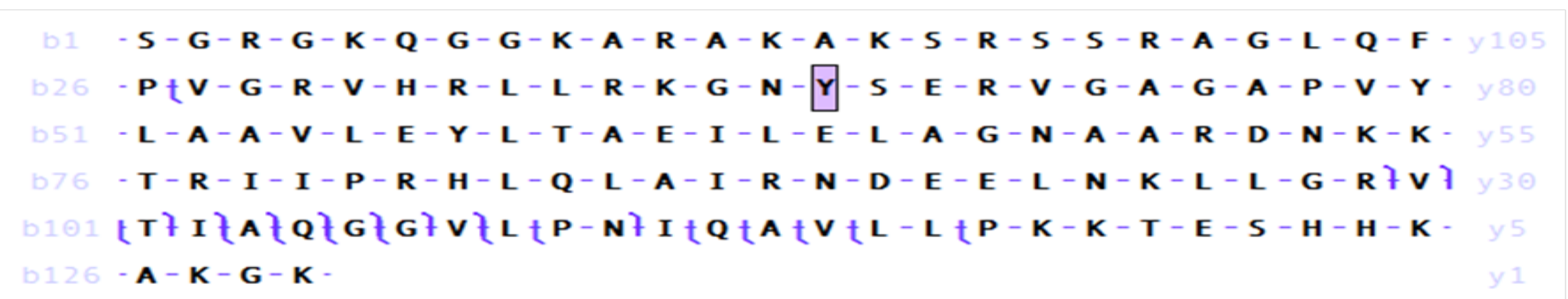
Cancel | < Prev | Next > | Finish

FIGURE 3. Lysine methylation sight on human mesenchymal stem cells H2A3 proteins


(A)
b1 S-G-R-G-K-Q-G-G-K-A-R-A-K-A-K-S-R-S-S-R-A-G-L-Q-F y105
b26 P-V-G-R-V-H-R-L-L-R-K-G-N-Y-S-E-R-V-G-A-G-A-P-V-Y y80
b51 L-A-A-V-L-E-Y-L-T-A-E-I-L-E-L-A-G-N-A-A-R-D-N-K-K y55
b76 T-R-I-I-P-R-H-L-Q-L-A-I-R-N-D-E-E-L-N-K-L-L-G-R-V y30
b101 T-I-A-Q-G-G-V-L-P-N-I-Q-A-V-L-L-P-K-K-T-E-S-H-H-K y5
b126 A-K-G-K y1
(B)
b1 S-G-R-G-K-Q-G-G-K-A-R-A-K-A-K-S-R-S-S-R-A-G-L-Q-F y105
b26 P-V-G-R-V-H-R-L-L-R-K-G-N-Y-S-E-R-V-G-A-G-A-P-V-Y y80
b51 L-A-A-V-L-E-Y-L-T-A-E-I-L-E-L-A-G-N-A-A-R-D-N-K-K y55
b76 T-R-I-I-P-R-H-L-Q-L-A-I-R-N-D-E-E-L-N-K-L-L-G-R-V y30
b101 T-I-A-Q-G-G-V-L-P-N-I-Q-A-V-L-L-P-K-K-T-E-S-H-H-K y5
b126 A-K-G-K y1

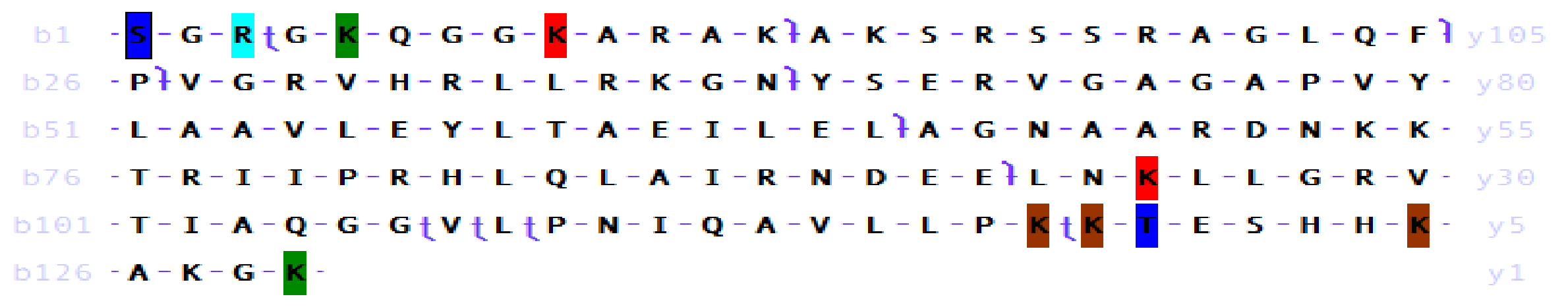
One of the recently identified PTM on histone is hydroxylation at tyrosine; this PTM is not documented in uniprot database. Hydroxylation at tyrosine was found to be present on H2A1 (accession: P0C0S8) protein at position 39 [2]. Amino acid sequence of H2A1 protein is very similar to H2A3 protein (Identical positions 127, similar position 3, identity = 97.692%) therefore we included this PTM in our modified flat file for H2A3 protein. We found hydroxylation to be present on human mesenchymal stem cells H2A3 protein, most probably at tyrosine (figure 4). However more fragment ion of the protein (may be by using HCD and ETD) is needed to confirm the location of this PTM.

FIGURE 4. Hydroxylation at tyrosine on human mesenchymal stem cells H2A3 protein


b1 S-G-R-G-K-Q-G-G-K-A-R-A-K-A-K-S-R-S-S-R-A-G-L-Q-F y105
b26 P-V-G-R-V-H-R-L-L-R-K-G-N-Y-S-E-R-V-G-A-G-A-P-V-Y y80
b51 L-A-A-V-L-E-Y-L-T-A-E-I-L-E-L-A-G-N-A-A-R-D-N-K-K y55
b76 T-R-I-I-P-R-H-L-Q-L-A-I-R-N-D-E-E-L-N-K-L-L-G-R-V y30
b101 T-I-A-Q-G-G-V-L-P-N-I-Q-A-V-L-L-P-K-K-T-E-S-H-H-K y5
b126 A-K-G-K y1

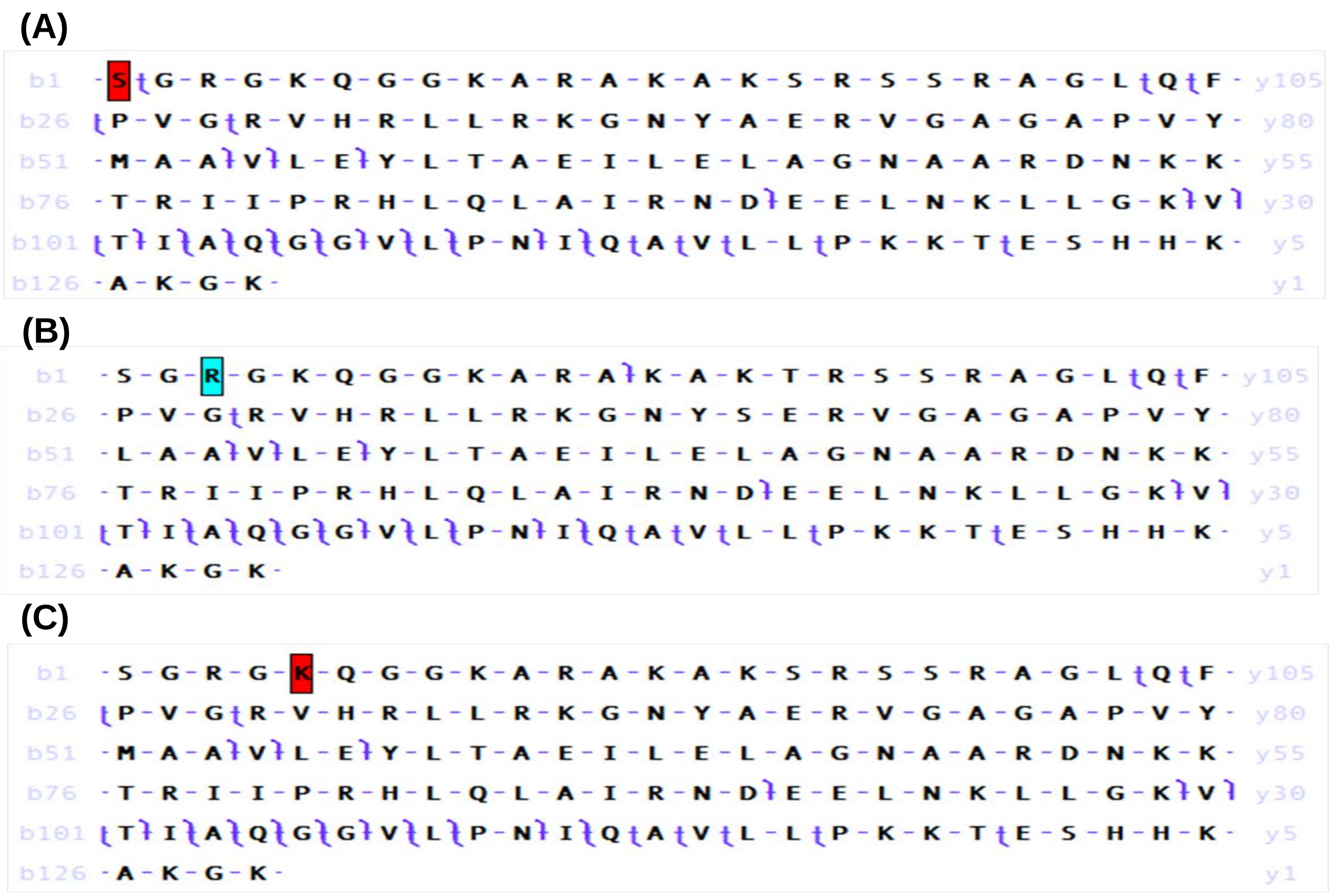
Another important recently discovered PTM on histone protein is crotonylation at lysine [1], we also found various Crotonyllysine on H2A3 protein with other PTMs such as phosphorylation at serine and tyrosine; dimethylation at arginine; methylation and acetylation at lysine (figure 5). The result shows the benefit of using modified uniprot flat file in order to identify multiple PTMs identification and discovery when present on a single protein which otherwise is a tedious job.

FIGURE 5. The figure shows presence of phosphorylation (blue box) at serine and tyrosine; dimethylation (sky blue box) at arginine; methylation (green box) and acetylation (red box) at lysine and crotonylation (brown box) at lysine on human mesenchymal stem cells H2A3 protein


b1 S-G-R-G-K-Q-G-G-K-A-R-A-K-A-K-S-R-S-S-R-A-G-L-Q-F y105
b26 P-V-G-R-V-H-R-L-L-R-K-G-N-Y-S-E-R-V-G-A-G-A-P-V-Y y80
b51 L-A-A-V-L-E-Y-L-T-A-E-I-L-E-L-A-G-N-A-A-R-D-N-K-K y55
b76 T-R-I-I-P-R-H-L-Q-L-A-I-R-N-D-E-E-L-N-K-L-L-G-R-V y30
b101 T-I-A-Q-G-G-V-L-P-N-I-Q-A-V-L-L-P-K-K-T-E-S-H-H-K y5
b126 A-K-G-K y1

In addition to above mentioned PTMs we also found various previously reported PTMs that were documented in the uniprot flat file for H2A3 protein including N-acetylserine, dimethylarginine and acetyllysine (figure 6).

FIGURE 6. The figure shows presence of (A) Acetylation at serine (B) dimethylation at arginine and (C) acetylation at lysine human mesenchymal stem cells H2A3 protein


(A)
b1 S-G-R-G-K-Q-G-G-K-A-R-A-K-A-K-S-R-S-S-R-A-G-L-Q-F y105
b26 P-V-G-R-V-H-R-L-L-R-K-G-N-Y-A-E-R-V-G-A-G-A-P-V-Y y80
b51 M-A-A-V-L-E-Y-L-T-A-E-I-L-E-L-A-G-N-A-A-R-D-N-K-K y55
b76 T-R-I-I-P-R-H-L-Q-L-A-I-R-N-D-E-E-L-N-K-L-L-G-K-V y30
b101 T-I-A-Q-G-G-V-L-P-N-I-Q-A-V-L-L-P-K-K-T-E-S-H-H-K y5
b126 A-K-G-K y1
(B)
b1 S-G-R-G-K-Q-G-G-K-A-R-A-K-A-K-T-R-S-S-R-A-G-L-Q-F y105
b26 P-V-G-R-V-H-R-L-L-R-K-G-N-Y-S-E-R-V-G-A-G-A-P-V-Y y80
b51 L-A-A-V-L-E-Y-L-T-A-E-I-L-E-L-A-G-N-A-A-R-D-N-K-K y55
b76 T-R-I-I-P-R-H-L-Q-L-A-I-R-N-D-E-E-L-N-K-L-L-G-K-V y30
b101 T-I-A-Q-G-G-V-L-P-N-I-Q-A-V-L-L-P-K-K-T-E-S-H-H-K y5
b126 A-K-G-K y1
(C)
b1 S-G-R-G-K-Q-G-G-K-A-R-A-K-A-K-S-R-S-S-R-A-G-L-Q-F y105
b26 P-V-G-R-V-H-R-L-L-R-K-G-N-Y-A-E-R-V-G-A-G-A-P-V-Y y80
b51 M-A-A-V-L-E-Y-L-T-A-E-I-L-E-L-A-G-N-A-A-R-D-N-K-K y55
b76 T-R-I-I-P-R-H-L-Q-L-A-I-R-N-D-E-E-L-N-K-L-L-G-K-V y30
b101 T-I-A-Q-G-G-V-L-P-N-I-Q-A-V-L-L-P-K-K-T-E-S-H-H-K y5
b126 A-K-G-K y1

We did not have enough fragment ion to exactly pinpoint the site of PTM in some instances. Use of complimentary fragmentation technique such as Electron-transfer dissociation (ETD) and Higher collision energy dissociation (HCD) can further help in locating exact position of a PTMs on protein. We are currently extending our workflow to include this.

Nevertheless the results strongly support the capability and efficiency of using modified flat files for identification of PTMs. Thus this methodology is useful for high throughput automated identification of known PTMs as well as for searching for novel targeted PTMs on intact proteins.

Conclusion

- Modified uniprot flat file helps in identification and discovery of various PTM on histone H2A3 protein using ProSightPC
- Several PTMs that were identified by the modified uniprot files were not present in the original uniprot database for human H2A3 protein
- Our methodology is fast, accurate, user friendly and broadly applicable for identification of multiple PTM on any protein of interest.
- The method is also helpful in exploring novel targeted PTMs

References

1. Lunyak VV, Rosenfeld MG. Epigenetic regulation of stem cell fate. Hum Mol Genet. 2008 Apr 15;17(R1):R28-36.
2. Tan M, Luo H, Lee S, Jin F, Yang JS, Montellier E, Buchou T, Cheng Z, Rousseaux S, Rajagopal N, Lu Z, Ye Z, Zhu Q, Wysocka J, Ye Y, Khochbin S, Ren B, Zhao Y. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell. 2011 Sep 16;146(6):1016-28.

Thermo SCIENTIFIC