IC: A New Peptide Identification Tool for Both Data-Dependent and Data-Independent Acquisition

Sangtae Kim[1], Ying S. Ting[2], Alex Hu[2], Richard D. Smith[1], William S. Noble[2], Michael J. MacCoss[2], Samuel H. Payne[1]
[1]Biological Sciences Division, Pacific Northwest National Laboratory; [2]Dept. Genome Science, University of Washington, Seattle, WA

Pacific Northwest NATIONAL LABORATORY

Overview

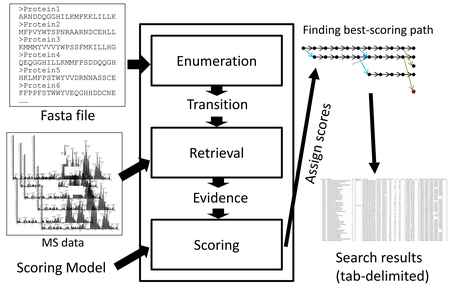
- Informed Characterization (IC) approach for peptide identification
- IC works well for both Data-Dependent Acquisition (DDA) and Data-Independent Acquisition (DIA).
- IC does not require assay libraries
- IC uses sequence graphs to decompose peptides into unique transitions, efficiently scoring peptides with modifications

Introduction

- DIA is becoming increasingly popular in proteomics as a complement to DDA.
- Several recent studies showed the usefulness of DIA for targeted proteomics [1,2,3]. However, for discovery proteomics, the full potential of DIA has not yet been achieved mainly because of the lack of effective algorithms for analyzing the data.
- We present an IC approach that works well for both DDA and DIA data analysis.
- Our prototype implementation of IC was applied to analyze DDA and DIA datasets from the same sample of high complexity and showed promising results, identifying more peptides from DIA than DDA for a complex sample.
- Furthermore, for DDA, IC identified significantly more peptides than our present state-of-the-art MS-GF+ [4] database search engine.

Methods

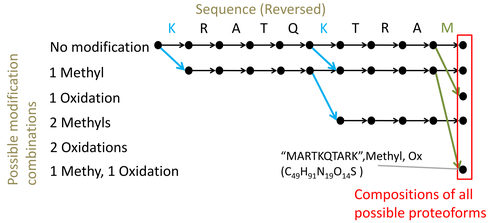
IC Workflow



Sequence Graph

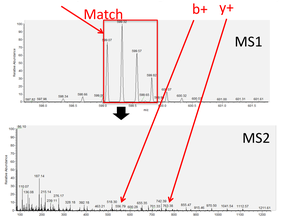
Decomposing a sequence into a set of unique transitions
Transition: (PrecursorComposition, FragmentComposition)

Sequence: MARTKQTARK ($C_{48}H_{89}N_{19}O_{13}S$)
Modification: Methyl K, (CH_2), Oxidation M (O)
Max #Modifications: 2



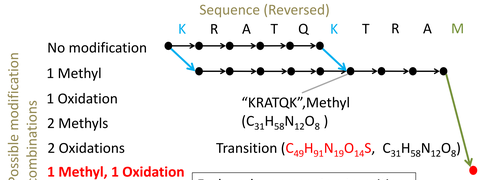
Scoring a transition

(*PrecursorComposition*, *FragmentComposition*)



This approach can be used for both DIA and DDA

Select "MARTKQTARK", 1 Methyl, 1 Oxidation ($C_{49}H_{91}N_{19}O_{14}S$)



Each node represents a transition.
Assign a score to each node.
Find the best scoring path.

Q-Exactive Dataset

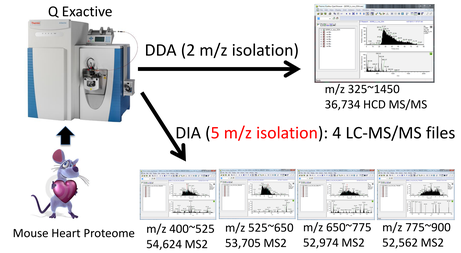


Figure 1. The Q-Exactive dataset. The DDA and DIA data were generated by a single injection and 4 injections of a mouse heart proteome sample, respectively. For the DDA data, the survey scan range was *m/z* 325-1450, the number of selected precursors was 12, and the isolation window width was *m/z* 2. The DIA spectra were generated similar to SWATH-MS, but with 5 *m/z* isolation windows.

Orbitrap Fusion Dataset

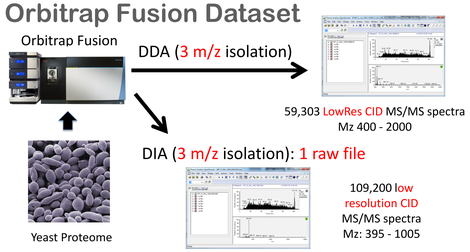
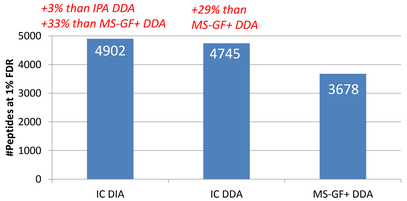


Figure 2. The Orbitrap Fusion dataset. The DDA and DIA data were generated by a single injection of a yeast proteome sample each. For the DDA data, the survey scan range was *m/z* 400-2000, the number of selected precursors was 10, and the isolation window width was 3 *m/z*. The DIA data was generated with the survey scan range of *m/z* 395-1005, and the isolation window width 3 *m/z*.

Results

Q-Exactive



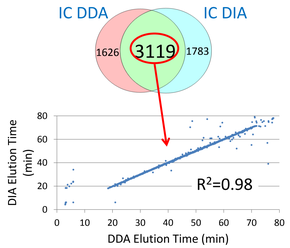
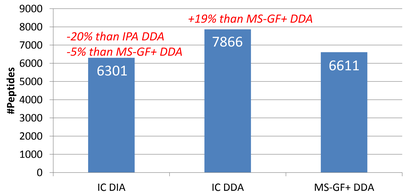


Figure 3. Results for the Q-Exactive dataset. IC identified slightly more peptides from the DIA data (4,902 peptides) than from the DDA data (4,745 peptides). 3,119 peptides with highly correlated elution times (R=0.98) were identified from both, and MS-GF+ identified only 3,678 peptides (30% less than IC) from QE-DDA, indicating that IC is effective both for DIA and DDA.

Orbitrap Fusion



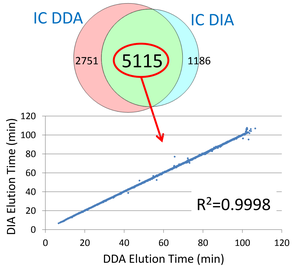


Figure 4. Results for the Orbitrap Fusion dataset, IC identified 20% less (6,301) peptides from Fusion-DIA than Fusion-DDA (7,866 peptides), mainly due to the Fusion-DIA dataset being created by a single injection. Similar to the Q-Exactive dataset, MS-GF+ identified 20% fewer peptides than IC from Fusion-DDA.

Conclusions

- IC was applied to DIA data acquired with 5 and 3 m/z isolation windows and shown to identify comparable numbers of peptides to those identified from DDA data from the same sample
- Unlike existing tools, e.g. OpenSWATH [3], IC does not require assay libraries
- For DDA data, IC outperformed MS-GF+

Acknowledgements

Samples were analyzed using capabilities developed under the support of NIH National Institute of General Medical Sciences (8 P41 GM103493-10). This project was supported by the U.S. Department of Energy (DOE) Office of Biological and Environmental Research (OBER) Pan-omics program at PNNL and performed in the Environmental Molecular Sciences Laboratory, a DOE OBER national scientific user facility on the PNNL campus. PNNL is a multiprogram national laboratory operated by Battelle for the DOE under contract DE-AC05-76RL01830.

References

1. Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, Aebersold R. Mol Cell Proteomics. 2012 Jun;11(6):O111.016717.
2. Egertson JD, Kuehn A, Merrihew GE, Bateman NW, MacLean BX, Ting YS, Canterbury JD, Marsh DM, Kellmann M, Zabrouskov V, Wu CC, MacCoss MJ. Nat Methods. 2013 Aug;10(8):744-6.
3. Röst HL, Rosenberger G, Navarro P, Gillet L, Miladinović SM, Schubert OT, Wolski W, Collins BC, Malmström J, Malmström L, Aebersold R. Nat Biotechnol. 2014 Mar 10;32(3):219-23.
4. Kim S, Mischerikow N, Bandeira N, Navarro JD, Wich L, Mohammed S, Heck AJ, Pevzner PA. Mol Cell Proteomics. 2010 Dec;9(12):2840-52.

CONTACT: Sangtae Kim, Ph.D.
Biological Sciences Division
Pacific Northwest National Laboratory
E-mail: sangtae.kim@pnnl.gov

Career Opportunities: For potential openings in the Omics Separations and Mass Spectrometry Department at PNNL please contact:
Dick Smith at rds@pnnl.gov
Josh Adkins at Joshua.Adkins@pnnl.gov

www.omics.pnl.gov