MSFragger Output Visualization & Evaluation (MOVE): An Interactive Interpretation and Graph Visualization Software for Clinical Proteomics Data

Lusha Cao, PhD MS[1], Kevin Nestler BS[2], Hossein Fazelinia PhD[1,3], Asif Chinwalla MS[1], Kathrin Bernt MD[2,4], Tina Glisovic-Aplenc PhD MS[2,4], Richard Aplenc MD PhD[2,4]

[1] Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, [2] Division of Oncology, Center for Childhood Cancer Research, Children's Hospital of Philadelphia, [3] Proteomics Core Facility, Children's Hospital of Philadelphia, [4] Department of Pediatrics, University of Pennsylvania Perelman School of Medicine

Children's Hospital of Philadelphia℠
Department of Biomedical & Health Informatics

BACKGROUND

- Large-scale mass spectrometry proteomics analyses of biologic specimens can provide insights into biological pathways
- Data manipulation and visualization of large data sets can be very challenging particularly for clinicians and researchers without extensive bioinformatics experience
- MSFragger[1] is a fast and comprehensive peptide identification software developed to generate peptide and protein identification from mass spectrometry data using proteome database searches
- Current downstream (post MSFragger) data integration tools lack the ability to display data in a user friendly, compact, and sharable dashboard

OBJECTIVE

- To develop a dashboard displaying MSFragger peptide and protein group data

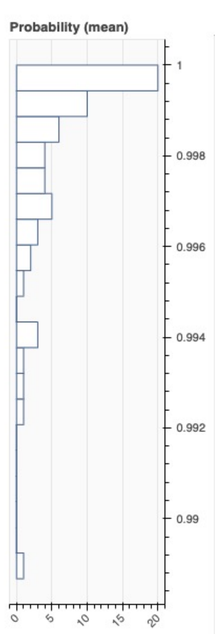
WORKFLOW

- We developed MOVE, an open source, object-oriented Python package
- MOVE uses MSFragger search results as input files to process data for downstream analysis
- MOVE performs normalization and semi-quantitative analysis, determines uniquely expressed proteins, creates explicit relational plots and interactive dashboards in a portable html format from a case vs control clinical proteomics data set
- MOVE has a noise reduction method using user-defined or default thresholds to select common proteins (i.e. proteins are detected above the threshold at least in one searching subgroup)
- A sample benchmark dataset (PXD025792)[2] was used to demonstrate the features of MOVE. Samples were from tumor growth factor beta stimulated (H1975_TGFb) or unstimulated H1975 cells (H1975_unst). Raw MS spectra were downloaded and then processed by MSFragger (version 3.3)

CONTACT: Lusha Cao, PhD MS, caol1@chop.edu

RESULTS

	All proteins		Common proteins	
	Missing	Missing (%)	Missing	Missing (%)
H1975_TGFb_Rep1_1	914	22.9	580	16.1
H1975_TGFb_Rep1_2	902	22.6	560	15.5
H1975_TGFb_Rep2_1	944	23.7	630	17.5
H1975_TGFb_Rep2_2	860	21.6	553	15.3
H1975_TGFb_Rep3_1	854	21.4	524	14.5
H1975_TGFb_Rep3_2	816	20.5	501	13.9
H1975_unst_Rep1_1	589	14.8	363	10.1
H1975_unst_Rep1_2	545	13.7	335	9.3
H1975_unst_Rep2_1	814	20.4	501	13.9
H1975_unst_Rep2_2	835	20.9	517	14.3
H1975_unst_Rep3_1	743	18.6	438	12.2
H1975_unst_Rep3_2	645	16.2	369	10.2



#	Protein ID	Gene	Number of peptide	%Prevalence	%Probability (mean)
0	Q02952	AKAP12	31	100	99.959
1	Q9P0K7	RAI14	29	100	99.802
2	Q86UU1	PHLDB1	20	100	99.354
3	Q13753	LAMC2	19	100	99.737
4	P05121	SERPINE1	18	100	99.933
5	O15327	INPP4B	12	100	99.823
6	P50281	MMP14	12	100	99.847
7	O75326	SEMA7A	11	100	99.617
8	P35052	GPC1	10	100	99.997
9	P05106	ITGB3	9	100	99.221
10	P25686	DNAJB2	7	100	99.976
11	Q6EMK4	VASN	6	100	100
12	Q6ZWT7	MBOAT2	6	100	99.595
13	Q02487	DSC2	5	100	99.794
14	Q8N4X5	AFAP1L2	5	100	99.766
15	Q92959	SLCO2A1	2	100	99.985
16	Q8WUF5	PPP1R13L	22	83.33	99.904
17	Q9ULH0	KIDINS220	11	83.33	99.694
18	Q15262	PTPRK	7	83.33	99.967
19	Q9HCN4	GPN1	7	83.33	99.686
20	O14640	DVL1	2	83.33	100
21	Q53LP3	SOWAHC	16	66.67	99.401
22	P10586	PTPRF	11	66.67	99.517

Figure 1. Results of MOVE. MOVE conducts automated data processing. After creating a MOVE object, user data can be uploaded and then normalized. In addition, a summary table showing missing values in each sample before and after noise reduction will be created based on the user-defined filter threshold. The user can specify names of case and control groups to have a Venn diagram created. An interactive html dashboard with annotated uniquely expressed proteins in case group will be generated. The user can hover over the dashboard or zoom in to evaluate the peptide numbers of a certain protein and the associated probability. The example data and code to create plots and dashboard can be found in the MOVE GitHub repository (Enterprise Cloud for CHOP). We also provide a tutorial Jupyter notebook.

CONCLUSION

- MOVE can easily be incorporated into a user's existing quantitative and statistical data analysis workflow
- MOVE will facilitate interpreting the output of MSFragger and producing an intelligible comparative analysis for large proteomics data

REFERENCES

1. *Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods. 2017;14(5):513-520. doi:10.1038/nmeth.4256.*
2. *Project PXD025792. Proteomics Identification Database. https://www.ebi.ac.uk/pride/archive/projects/PXD025792*