A Graphical User Interface (GUI) for Phosphorylation Site Assignment of Protein Mass Spectrometry Data [1]

Fahad Saeed
Epithelial Systems Biology Laboratory (ESBL)
National Heart Lung and Blood Institute (NHLBI)
National Institutes of Health (NIH)
Bethesda, MD USA

July 5, 2013

[1]This material is made possible by the support of operating budget of Division of Intramural Research, National Heart, Lung and Blood Institute, National Institutes of Health (NIH), Project ZO1-HL001285. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of National Institutes of Health (NIH). The Mass spectrometry data generation was conducted in the National Heart Lung and Blood Institute, Proteomics Core Facility.

Abstract

Correct phosphorylation site assignment for high throughput tandem mass spectrometry (LC-MS/MS) data is one of the most common and critical aspects of phosphoproteomics. In this report, we present a graphical user interface (GUI) implemented in JAVA for phosphorylation site assignment. The GUI implements the PhosSA algorithm and is tested on a variety of operating systems and computing platforms. The GUI is divided into two parts: The first part takes input from multiple search engines (i.e. Sequest and Mascot) and converts it into PhosSA compatible format. The second part of the GUI runs the site assignment algorithm using varying thresholds for HCD and CID fragmentation methodologies. The software is accessible for free at: `http://helixweb.nih.gov/ESBL/PhosSA/` for all non-commercial purposes.

1 Introduction

Mass spectrometry is a fundamental part of any modern proteomics research platform for large-scale protein identification and quantification [1][2][3]. In order to get peptide sequence information, a number of fragmentation methods are employed, such as CID (collision induced dissociation) and HCD (Higher Energy Collisional Dissociation). Search engines that map the fragmentation spectra to the peptides using databases are used for protein identification [4][5].

Phosphoproteomics is an emerging area in protein mass spectrometry and has useful applications in biology [6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. As mass spectrometers get more efficient, computational tools are required that can deal efficiently with large data sets. Historically, correct phosphorylation site assignment has been done using manual inspection. However, with the advent of high throughput mass spectrometers, the data generated is generally vast and manual site assignment is no longer practical. Therefore a number of site assignment tools have been recently proposed [16, 17, 18, 19, 20, 21].

Previously, our lab presented the first dynamic programming solution to phosphorylation site assignment called PhosSA[22]. The main goal of the current report is to present a user friendly graphical user interface that implements the PhosSA algorithm [22]. We discuss different characteristics of the developed GUI and a step by step procedure that researchers can find useful for their phosphorylation site assignment problems. At the end of the report we also present a method that would allow users to decide the site assignment for composite spectra. PhosSA algorithm has been used successfully in a number of studies[23, 24].

2 Graphical User Interface

PhosSA has been implemented using the JAVA programming language. The software assumes an input from Sequest search results (.out and .dta files). There is a wide variability in the formats representing mass spectrometry data. Therefore, we have developed convertors embedded in the GUI to make it compatible with multiple formats. Two main convertors included on our webpage is a convertor that transforms Proteome Discoverer(.msf) files and another convertor that converts PepXml files to .out files. Both Sequest and Mascot search results can be exported as PepXml format. In both of the cases, spectral (.dta) files can be imported using Proteome Discoverer. The layout of the GUI is shown in Fig. 1. The following section describes step-by-step user instructions.

2.1 Installation and Configuration

Since, PhosSA GUI has been implemented using the JAVA programming language no installation of the program is required. The executable of the program is a .jar file which is equivalent to an .exe file familiar to most PC users. Although there is no need to install the program (only have to double click the distributed jar file), the computer system should be configured correctly for the program to operate. The algorithm has been implemented in Java(TM) SE Runtime Environment (build1:6:0). Therefore, a JAVA SUN version of 1.6 or higher is required on the system. Since, JAVA versions are backward compatible, the program will run correctly with a newer version of JAVA. For the uninitiated, the users can go to `www.java.com/getjava/`, which will take you to a java download site and recommend the correct version of java considering your computer hardware and operating system.

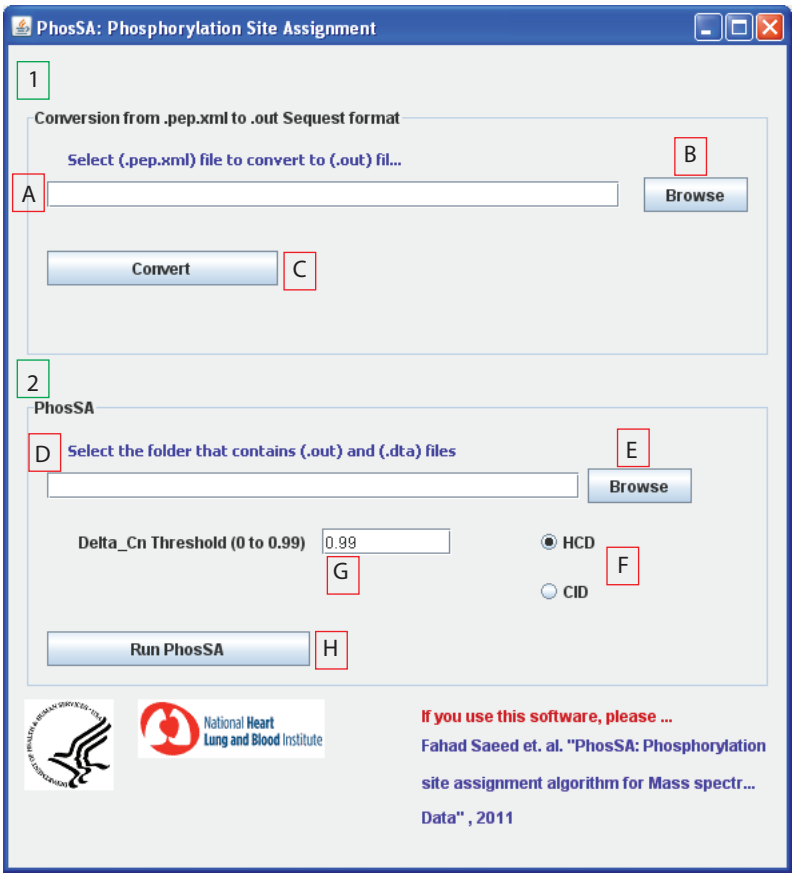

Figure 1: Graphical User Interface developed for phosphorylation site assignment is shown

2.2 File Formats Conversion

Three file format convertors are available from our webpage `http://helixweb.nih.gov/ESBL/PhosSA/`. One of the convertors is shown in Fig. 1 in part 1, which converts a .PepXml file to .out files. This allows both the Sequest and Mascot search results to be compatible with PhosSA since both of the search results can be transformed in Pepxml using Proteome Discoverer and then converted into .out files. There is another convertor available from our webpage which converts a .msf file to .out file. The user would have to click "browse" (B) which will open up the file system on the machine. The user then chooses the .PepXml file that he/she wishes to convert. Once the file is selected, the complete path of the file is shown in the status bar labeled with (A). Thereafter, the user hits "Convert" (labeled C) to convert the PepXml file to .out file. Once the conversion is complete a message will appear alerting the user. Note that the .out file(s) would be produced in the same folder where .PepXml file resides. One can use (export $\rightarrow$ spectra $\rightarrow$ .dta) schema in Proteome discoverer to convert .msf to .dta. Note that .out and .dta files should be in one folder for the GUI to run correctly. The flowchart that researchers can use for different format conversions is shown in Fig. 2.

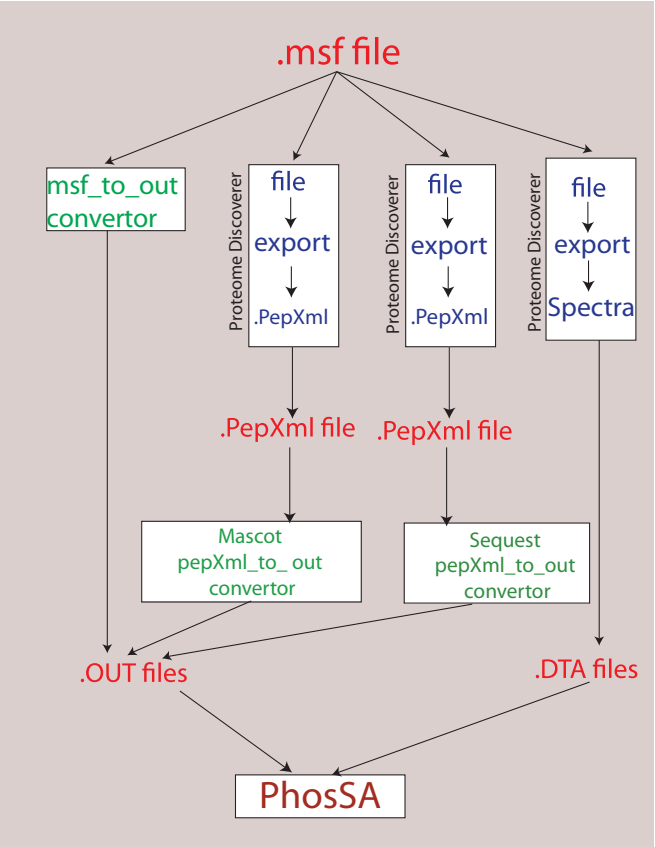

Figure 2: A flowchart shows the various ways in which convertors can be used provided on the webpage for the user. The figure shows that .msf file can be converted into .out files using convertor 1 (msf-to-out convertor). Different convertors 2 (Mascot PepXml-to-out convertor) and 3 (Sequest PepXml-to-out convertor) are provided for conversion of Mascot and Sequest search files using .PepXml files that can be generated using built in Proteome Discoverer utility. The .dta files can be converted using the built in utility in Proteome Discoverer. The .out and .dta files are then fed into the PhosSA algorithm.

2.3 Executing the site assignment routine

Figure 1 (2nd section) shows the part that is used for executing the site assignment algorithm with the assumption that .dta and .out files are present in the folder. The user can choose a folder that has both .out and .dta files (produced using the method described in the previous section) using browse function (E). Once the user has chosen a local folder, the status of the file and the path appears on the bar (labeled D). The user can then choose two other attributes that would aid PhosSA in making a superior assignment. The two attributes are briefly explained below:

1. *Fragmentation methodology:* There are two fragmentation options provided to the user, namely HCD and CID (labeled F in Fig. 1). The user can choose the fragmentation methodology for the specified data using the radio buttons labeled as "HCD" and "CID". It is important to select the correct fragmentation methodology because of the differences in the peaks and noise levels considered by PhosSA for different fragmentation methodologies. Incorrect selection would result in less accurate site assignment.

2. $DeltaC_nthreshold$: For PhosSA algorithm, threshold is similar to the $deltaC_n$ (dC_n) function used in algorithms like Sequest [5][25][4][26]. The dC_n threshold is defined as

the difference of the highest and second highest scoring sites which is normalized with the highest scoring peptide. If there are two scores very close to each other, then it is very difficult to decide if the peptide that has the highest score is correct. Thus, for our purposes the higher dC_n, the higher the accuracy of the assigned sites. The user can set a dC_n threshold by inserting numbers from 0(lowest quality) to 0.99(highest quality) in the bar labeled as G in Fig.1.

Once the user has chosen the folder that has .out and .dta files, the fragmentation methodology, and the dC_n appropriate for the dataset, PhosSA can be executed by using "Run PhosSA" labeled as H in the figure. After execution of the algorithm, the program generates a dialog box that informs the user that the algorithms is done computing. The results from PhosSA are stored in *outresult2* file in the same folder where .dta and .out files resided in the user computer. Another parameter, the redundancy metric, is also available in PhosSA algorithm. Redundancy is found to be extremely effective for classification [22], and is set as a constant value of 7 in the PhosSA algorithm. The redundancy metric along with dCn is used for classification of the peptides. We recommend a value of $dC_n = 0.99$ to be used for most data sets. The results are reported in a text file and a brief explanation of the format of the results is discussed in the next section.

3 Results Report

A	B	C	D	E
JH_1_HCD.1799.1799.3	K.HPEAPDEES*DHDYQNHI.-	0.92626024	1	Ambiguous
JH_1_HCD.1806.1806.3	R.LGNRKS*VVFTSAR.A	0.92318013	1	Ambiguous
JH_1_HCD.1810.1810.3	K.EELEQQT*DGDCDEEDDDKDGEMPK.S	1	2	Passed
JH_1_HCD.1814.1814.2	K.IGGHGGEYGEEALQR.M	1	1	Passed
JH_1_HCD.1815.1815.2	R.VRPASSAAS*VYAGAGGSGSR.I	0.18312842	2	Ambiguous
JH_1_HCD.1818.1818.3	R.VRPASSAAS*VYAGAGGSGSR.I	0.05426612	2	Ambiguous
JH_1_HCD.1820.1820.3	R.RRQS*VELHS*PQSLPR.G	0.50481409	108	Passed
JH_1_HCD.1821.1821.2	K.EELEQQT*DGDCDEEDDDKDGEMPK.S	1	2	Passed
JH_1_HCD.1845.1845.2	K.LSSPATLNSR.V	1	2	Passed
JH_1_HCD.1867.1867.3	K.AKPS*PAPS*PTISAPDASGPQKR.S	0.04270222	1	Ambiguous
JH_1_HCD.1869.1869.2	R.RRQSVELHS*PQS*LPR.G	0.22501839	1	Ambiguous
JH_1_HCD.1870.1870.3	R.RRQS*VELHSPQSLPR.G	0.58898423	3	Ambiguous
JH_1_HCD.1873.1873.3	R.RRQS*VELHS*PQSLPR.G	0.36523256	108	Passed
JH_1_HCD.1877.1877.4	R.RRQS*VELHSPQSLPR.G	0.8396499	3	Ambiguous
JH_1_HCD.1887.1887.4	R.RRQS*VELHS*PQSLPR.G	0.6618544	108	Passed
JH_1_HCD.1892.1892.2	K.HTGPNS*PDTANDGFVR.L	0.39708935	1	Ambiguous
JH_1_HCD.1905.1905.3	K.GVPM#KARM#IHSLS*GKK.S	0.12197665	2	Ambiguous
JH_1_HCD.1922.1922.2	R.RRQS*VELHS*PQSLPR.G	0.38376987	108	Passed
JH_1_HCD.1925.1925.2	K.YNEVLTQCCTESDK.A	1	1	Passed
JH_1_HCD.1929.1929.3	R.RRQS*VELHS*PQSLPR.G	0.35837882	108	Passed
JH_1_HCD.1936.1936.4	R.RRQS*VELHS*PQSLPR.G	0.69158719	108	Passed
JH_1_HCD.1940.1940.3	R.RRQS*VELHSPQSLPR.G	0.52142149	3	Ambiguous

Figure 3: The output format in a text file is shown.

Figure 3 shows standard formatted PhosSA output. As can be seen in the figure the output consists of 5 columns each of which is distinctly color coded in the figure for ease of presentation. The first column consists of the name of the .out and .dta file. Note that the file format shown and used in PhosSA is same as used in Sequest output format i.e. Name-of-sample.start-scan.end-scan.charge.[out/dta]. The second column reports a peptide with site assigned. Note that each of the reported peptide has a amino acid at the start followed by a dot and each of the peptide ends by a dot followed by an amino acid. Each of these start and stop sites are used as sentinel values for efficient processing in the computer program. Therefore these amino acids are to be ignored when doing further processing i.e. start of the *real* peptide are after the dot and end of the real peptide is before the second dot. The third column reports the dC_n value for the reported peptide and scan number.

The closer this number is to 1, the more confident the site assignment. The fourth column reports the redundancy metric that is calculated for the peptide. The details about the implementation of the redundancy metric have been previously described[22]. In general, a higher redundancy values indicate a more confident site assignment. The last column reports two states, "passed" or "ambiguous", for each peptide. This is the final verdict given by PhosSA algorithm.

3.1 Ambiguous Site Assignment

PhosSA is designed for site assignment at the peptide level. The standard PhosSA output consists of two types of assignments, "passed" and "ambiguous". Peptides that are labeled "passed" possess phosphorylation sites that have been assigned with high confidence based on both the dCn value and redundancy filter. Peptides that are labeled "ambiguous" possess phosphorylation sites that have not been assigned with confidence and may require further investigation. There are multiple reason why the algorithm would assign a phosphopeptide as "ambiguous" e.g. poor spectral quality, the presence of composite spectra (i.e. spectra containing more than one distinct peptide species), and instances where multiple phosphorylatable residues are in close proximity (resulting in a low dCn value). In cases where the phosphopeptide is reported "ambiguous", we suggest users may wish to try an additional program CPhos [27], to break the tie between assignment(s). CPhos utilizes an information-theory based algorithm to assess the conservation of phosphorylation sites among species. We assert that the site (even if it is reported ambiguous by PhosSA) is likely to be correctly assigned if it is well conserved across multiple species, as conserved phosphorylation sites are more likely to play functional roles than non-conserved sites. Ba and Moses, 2010, Landry 2009, Malik 2008.

4 Discussion

In this report we have given a short introduction to a graphical user interface that we have developed for phosphorylation site assignment. The site assignment algorithm called PhosSA has been implemented using JAVA programming language and is machine and operating system(OS) independant. The implemented GUI has been tested on windows, mac and Linux operating systems and is available free of charge to any one for non-commercial purposes. This report is intended as a guide for phosphorylation site assignment using PhosSA algorithm for proteomics practitioners. In this report we discussed the format/layout of the GUI and explained in some detail the various parts and function of the user interface. We also discussed the different format conversion that are available in the interface and some aspects of parameters that are used in the algorithm. The output format of the algorithm is also discussed in some detail to aid the users in understanding the results from PhosSA. The developed interface is easy to use, requires minimal configuration and is portable on various platforms. We expect a wide variety of proteomics and mass spectrometry users to benefit from our easy to use graphical user interface for site assignment.

Bibliography

[1] J. Hoffert, T. Pisitkun, G. Wang, R. Shen, and M. Knepper, "Quantitative phosphoproteomics of vasopressin-sensitive renal cells: Regulation of aquaporin-2 phosphorylation at two sites.," *Proc Natl Acad Sci U S A*, 2006.

[2] X. Li, S. A. Gerber, A. D. Rudner, S. A. Beausoleil, W. Haas, J. E. Elias, and S. P. Gygi, "Large-scale phosphorylation analysis of alpha-factor-arrested saccharomyces cerevisiae.," *J Proteome Res*, vol. 6, no. 3, pp. 1190–7, 2007.

[3] A. Gruhler, J. V. Olsen, S. Mohammed, P. Mortensen, N. J. FÃrgeman, M. Mann, and O. N. Jensen, "Quantitative Phosphoproteomics Applied to the Yeast Pheromone Signaling Pathway," *Molecular & Cellular Proteomics*, vol. 4, pp. 310–327, March 2005.

[4] S. Tanner, H. Shu, A. Frank, L. C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna, "InsPecT: identification of post translationally modified peptides from tandem mass spectra.," *Analytical Chemistry*, vol. 77, pp. 4626–4639, July 2005.

[5] J. K. Eng, B. Fischer, J. Grossmann, and M. J. Maccoss, "A Fast SEQUEST Cross Correlation Algorithm," *J. Proteome Res.*, September 2008.

[6] N. Musbacher, T. B. Schreiber, and H. Daub, "Glycoprotein Capture and Quantitative Phosphoproteomics Indicate Coordinated Regulation of Cell Migration upon Lysophosphatidic Acid Stimulation," *Molecular & Cellular Proteomics*, vol. 9, no. 11, pp. 2337–2353, 2010.

[7] M. F. Moran, J. Tong, P. Taylor, and R. M. Ewing, "Emerging applications for phosphoproteomics in cancer molecular therapeutics," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1766, no. 2, pp. 230 – 241, 2006. Genomics and Predicting Drug Sensitivity.

[8] D. B. Solit and I. K. Mellinghoff, "Tracing cancer networks with phosphoproteomics," *Nat Biotech*, vol. 28, pp. 1028–1029, 10 2010.

[9] M. H., "Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, p. 2199, 2007.

[10] G. A., "Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway," *Mol. Cell. Proteomics*, vol. 4, p. 310, 2005.

[11] W.-Y. A., "Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, p. 5860, 2007.

[12] C. G. T., "Iii quantitative phosphoproteomic analysis of the tumor necrosis factor pathway," *J. Proteome Res.*, vol. 5, p. 127, 2006.

[13] B. S. A., "Large-scale characterization of hela cell nuclear phosphoproteins," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, p. 12130, 2004.

[14] O. J. V., "Global, in vivo, and site-specific phosphorylation dynamics in signaling networks," *Cell*, vol. 127, p. 635, 2006.

[15] H. J. D., "Quantitative phosphoproteomics of vasopressin-sensitive renal cells: regulation of aquaporin-2 phosphorylation at two sites," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, p. 7159, 2006.

[16] S. A. Beausoleil, J. Villen, S. A. Gerber, J. Rush, and S. P. Gygi, "A probability-based approach for high-throughput protein phosphorylation analysis and site localization," *Nature Biotechnology*, vol. 24, pp. 1285–1292, September 2006.

[17] D. MacLean, M. Burrell, D. Studholme, and A. Jones, "Phoscalc: A tool for evaluating the sites of peptide phosphorylation from mass spectrometer data," *BMC Research Notes*, vol. 1, no. 1, p. 30, 2008.

[18] J. Cox and M. Mann, "Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification," *Nat Biotech*, vol. 26, pp. 1367–1372, 12 2008.

[19] D. Li, Y. Fu, R. Sun, C. X. Ling, Y. Wei, H. Zhou, R. Zeng, Q. Yang, S. He, and W. Gao, "pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry," *Bioinformatics*, vol. 21, no. 13, pp. 3049–3050.

[20] B. E. Ruttenberg, T. Pisitkun, M. A. Knepper, and J. D. Hoffert, "PhosphoScore: an open-source phosphorylation site assignment tool for MSn data.," *Journal of proteome research*, vol. 7, pp. 3054–3059, July 2008.

[21] D. L. Swaney, C. D. Wenger, J. A. Thomson, and J. J. Coon, "Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry," *Proceedings of the National Academy of Sciences*, vol. 106, no. 4, pp. 995–1000, 2009.

[22] F. Saeed, T. Pisitkun, J. D. Hoffert, G. Wang, M. Gucek, and M. A. Knepper, "An efficient dynamic programming algorithm for phosphorylation site assignment of large-scale mass spectrometry data," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*, pp. 618–625, IEEE, 2012.

[23] J. Hoffert, T. Pisitkun, F. Saeed, J. Song, C. Chou, and M. Knepper, "Dynamics of the g protein-coupled vasopressin v2 receptor signaling network revealed by quantitative phosphoproteomics," *Molecular & Cellular Proteomics*, vol. 11, no. 2, 2012.

[24] S. Bolger, P. Hurtado, J. Hoffert, F. Saeed, T. Pisitkun, and M. Knepper, "Quantitative phosphoproteomics in nuclei of vasopressin-sensitive renal collecting duct cells," *American Journal of Physiology-Cell Physiology*, 2012.

[25] J. K. Eng, A. L. McCormack, and J. R. Y. III, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 11, pp. 976 – 989, 1994.

[26] S. Tanner, P. A. Pevzner, and V. Bafna, "Unrestrictive identification of post-translational modifications through peptide mass spectrometry," *Nat. Protocols*, vol. 1, pp. 67–72, 06 2006.

[27] B. Zhao, T. Pisitkun, J. D. Hoffert, M. A. Knepper, and F. Saeed, "Cphos: A program to calculate and visualize evolutionarily conserved functional phosphorylation sites," *Proteomics*, vol. 12, no. 22, pp. 3299–3303, 2012.