PEAKS Q: Software for MS-based quantification of stable isotope labeled peptides

Weijie Yang[1], Weiwu Chen[1], Iain Rogers[1], Sean Bendall[2], Derek Smith[2], Bin Ma[3], Gilles Lajoie[3]
[1]Bioinformatics Solutions Inc., Waterloo, ON., [2]Genome BC Proteomics Centre, Victoria, BC., [3]University of Western Ontario, London, ON.

Bioinformatics Solutions Inc.

Introduction

Several mass spectrometry-based stable isotope-labeling technologies have been developed for global proteome profiling. These include methods for in vivo labeling, such as $^{14}N/^{15}N$ and SILAC (Stable Isotope Labeling with Amino Acids in Cell Culture), and in vitro isotope labeling of target peptides at their N/C terminal or at specific residues. In this work we describe a new software, PEAKS-Q, designed to automatically identify and quantify proteins from these isotope labeling experiments. The software is written in Java and includes an intuitive graphical user interface.

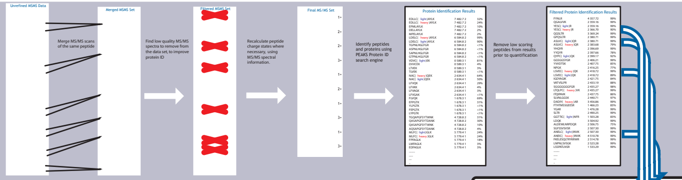
Methods

The following functions are necessary:

1) Data Preprocessing: Sophisticated algorithms are used to merge spectra with similar retention time and m/z values, determine charge state when necessary, remove poor quality MS/MS scans, and remove noise, centroid and deconvolute data within MS/MS scans.

2) Protein identification: Peptides are identified from MS/MS data with the PEAKS protein identification algorithms.

3) Protein quantification: the ratio of each identified labeled peptide is calculated from the intensities of MS peaks that differ in mass by the mass of the label. The abundance of a peptide is obtained by averaging ratios from all observed charge state species of that peptide. Statistical methods are then applied to calculate the relative protein abundance and its associated standard deviation.



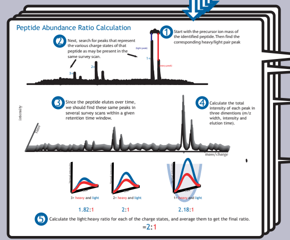
Removal of Deficient MS/MS spectra prior to analysis

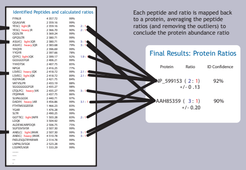
Electrical noise, poor detection and contaminants scanned by the MS mean that only a small portion of data are quality MS/MS spectra representing peptides. Database search engines and de novo sequencing tools are adequate in discarding the bad spectra; nevertheless, false positives abound, and plenty of time is wasted. Hence a filter that eliminates poor spectra before the analysis can significantly improve throughput and robustness in a quantification software.

The algorithm included in the software had 99.61% accuracy in finding spectra of deficient quality and took only 130 seconds to reduce a

Peptide charge determination

Usually, we examine the initial MS survey scan of a peptide to determine the peptide precursor charge state. But we cannot use this method with low resolution data, as from most ion-trap instruments. If we let a protein identification tool decide the charge, we increase the risk of false positive matches, and triple the search time. As such, an algorithm is built into the software to find precursor charges with confidece, using low resolution tandem mass spectra data alone. The algorithm took less than four seconds to correctly assign charge on 313 spectra, with 92% accuracy[2].





Protein quantification

The ratio of each identified labeled peptide is calculated from the intensities of MS peaks that differ in mass by the mass of the heavy and light labels. The abundance of a peptide is obtained by averaging the ratios from all the observed charge state of that peptide. Dixon's test algorithm is used to remove extreme values (outliers) from a continuous data set. Statistical methods are then applied to calculate the relative protein abundance and its associated log deviation.

Let I denote peak intensity where I^L and I^H denote light and heavy peak intensities respectively. Φ^{+z} denotes all isotopic peak groups where "z" is the charge of the peptide. Assume a peptide is eluted between a range of retention time range (T_1 to T_n) where $1 \le k \le n$, so that I_k^L and I_k^H can be written as follows:

$$I_k^H = \sum_{z \in all\_possible\_charge} \sum_{j \in \Phi^{+z}} I_{z,j}^H \qquad I_k^L = \sum_{z \in all\_possible\_charge} \sum_{j \in \Phi^{+z}} I_{z,j}^L$$

$$r = \frac{\sum_{k=2}^{n} (I_k^L + I_{k-1}^L)(T_k - T_{k-1})}{\sum_{k=2}^{n} (I_k^H + I_{k-1}^H)(T_k - T_{k-1})}$$

And the ratios (r) of peptide can be calculated by the following equation:

$$\log r_p = \frac{1}{|S(p)|} \sum_{i \in S(p)} \log r_i$$

$$LD_p = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (\log r_i - \log r_p)^2}$$

Isotope Distribution

The software contains a new algorithm to predict the isotope distribution of a given peptide. The isotope abundance distribution is useful for comparing observed mass spectrometry data in helping to predict the number of atoms of a given element in the formula, and to distinguish signal peaks from chemical noise. The abundance distribution D is calculated using the observed natural abundance of each element in the formula, and convoluting these natural occurring abundances to predict the experimental isotopic abundances.

$$D^k[i] = \sum_{j}^{\min(i, r \cdot k/2)} D^{k/2}[i-j] \cdot D^{k/2}[j] \qquad i \in [0, m]$$

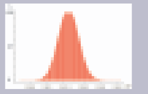
Improved time complexity (for abundance distribution calculation):

Removal of outliers

If we can assume that the abundance ratio of any peptide belonging to a protein is representative of that protein's abundance ratio, then in theory it follows that all peptides belonging to a protein should have the same abundance ratio. An outlier is a relatively small or large data point within a data set where these values are statistically different from the main body of the data. Outliers should be removed when calculating the protein abundance.

To find and remove outliers, the software computes the ratio between the difference of the minimum [or maximum value] with its neighbor value and the difference of the maximum and minimum values. This ratio should follow a certain distribution. The outlying minimum [or maximum] value is removed from the data set If does not follow the supported distribution[23].

of two which are less than n, then combine them together to reproduce n's abundance.



The method is very fast, the result is very accurate, and the advantage is memory-efficiency when calculating distributions for extremely large proteins/peptides. For protein gi|7460236|, whose molecular mass is 1184139.008Da, and has 13288 amino acids, the calculation time was less than seven seconds.

Experimental Results: BSA with cleavable ICAT

BSA samples were labeled with either light or heavy cleavable ICAT reagent and digested with trypsin. The light and heavy-labeled samples were then mixed together with ratios of approximately 1:1, 1:2, 1:4, 2:1 and 4:1. These samples were analyzed by LC MS and MS/MS on a Waters QTOF instrument. The software successfully identified 4 of ICAT derived peptides that differ exactly by 9 Da as light/heavy pairs. The correct abundance ratio for each sample was determined, which indicates that the software can accurately determine abundance ratios over the dynamic range provided for this labeling experiment.

Protein	Peptide sequence	ratio	Protein ratio ± SD
ALBU_BOVIN (1:1)			0.83±0.03
	C(+)ASLQK	0.57(outlier, removed)	
	GAC(-)LLPK	0.78	
	GAC(+)LLPK	0.78	
	YLC(+)DNQDTLSSK	0.82	
	YLC(-)DNQDTLSSK	0.83	
	EAC(+)FAVEGPK	0.84	
	QNC(-)DQFEK	0.85	
	QNC(+)DQFEK	0.85	
	C(-)C(-)TESLVNR	17.39(outlier, removed)	
ALBU_BOVIN (2:1)			1.13±0.1
	SHC(-)LAEVEK	0.98	
	SHC(+)LAEVEK	0.98	
	SLHTLFGDELC(+)K	1.04	
	C(-)ASLQK	1.05	
	C(+)ASLQK	1.05	
	SLHTLFGDELC(-)K	1.08	
	GAC(-)LLPK	1.09	
	DDPHAC(-)YSTVFDK	1.16	
	YLC(-)DNQDTLSSK	1.17	
	YLC(+)DNQDTLSSK	1.17	
	LC(-)VLHEK	1.19	
	QNC(-)DQFEK	1.24	
	QNC(+)DQFEK	1.24	
	LKPDPNTLC(+)DEFK	1.26	

Protein	Peptide sequence	ratio	Protein ratio ± SD
ALBU_BOVIN (1:2)			0.34±0.03
	SHC(+)LAEVEK	0.24(outlier, removed)	
	GAC(-)LLPK	0.29	
	GAC(+)LLPK	0.29	
	YLC(+)DNQDTLSSK	0.33	
	C(+)ASLQK	0.35	
	QNC(+)DQFEK	0.35	
	LC(+)VLHEK	0.37	
	LKPDPNTLC(+)DEFK	0.45(outlier, removed)	
	TC(-)VADESHAGC(-)EK	4.51(outlier, removed)	
ALBU_BOVIN (4:1)			2.53±0.50
	SHC(-)LAEVEK	2.13	
	GAC(+)LLPK	2.84	
	C(-)C(-)TESLVNR	14.46(outlier, removed)	
ALBU_BOVIN (1:4)			0.18±0.02
	QNC(-)DQFEK	0.16	
	QNC(+)DQFEK	0.16	
	LC(+)VLHEK	0.16	
	EAC(+)FAVEGPK	0.17	
	EAC(-)FAVEGPK	0.17	
	GAC(+)LLPK	0.19	
	C(+)ASLQK	0.21	
	C(+)C(+)TESLVNR	0.35(outlier, removed)	
	C(+)C (+)TKPESER	0.93(outlier, removed)	

Experimental Results: Genome BC Proteomics Centre; *E. Coli* Complex mixture, ICAT

Proteins from *E. Coli* bacteria grown at the Genome BC Proteomics Centre were labeled with the isotopically light and heavy ICAT reagent and were analyzed by LC MS/MS on an Applied Biosystems Q-STAR instrument. The proteins were identified by PEAKS Studio database search against the NCBI NR protein database with the taxonomy defined as ***E. coli***. Sixty five (65) proteins were identified from 1055 tandem mass spectra of ICAT labeled peptide pairs. Of all identified peptides, 96 peptides had acceptable chromatographic peaks in both the light and the heavy isotopic forms. The range of ratios determined by the software was between 7.73 and 0.06. About 87% of the peptide abundance ratios were within one log deviation. Sixteen (16) proteins had significant abundance changes, as measured by statistically derived p values. By evaluating the p values to specify the significance of protein abundance changes, the software is clearly capable of focusing and quantifying proteins of interest from a very large background.

Experimental Results: Genome BC Proteomics Centre; 5 to 1 BSA ICAT

Two samples, both containing TRFE_HUMAN, ALBU_BOVIN, and LCA_BOSMU were labelled -- one with isotopically light ICAT reagent and the other with heavy ICAT -- at the Genome BC Proteomics Centre. The two samples were combined together to create one mixture with 5:1 light to heavy ratio. The mixture was subsequently analyzed by LC MS/MS on an ABI QSTAR (ESI-QTOF), and the resulting data run through PEAKS Q for quantification. All three proteins were confidently identified by a PEAKS Protein ID search against the UniProt database. Table 1 summarizes the results of identification and

Accession	Peptide	Source	ID Score(%)	Ratio(light/heavy)	StdDev
+ TRFE_HUMAN	(P02787) Serotransferrin precursor (Transferrin) (Siderophi...		7.4	4.51	0.47
+ ALBU_BOVIN	(P02769) Serum albumin precursor (Allergen Bos d 6)...		12.8	5.27	0.52
	E[C-][C-]DKPLLEK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1319]	52.85	4.44	
	[C-][C-]TESLVNR	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1423]	80.13	4.56	
	[C-][C-]TESLVNR	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1422,1426]	99	4.62	
	L[C-]VLHEK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1238,1241,1246]	78	4.66	
	QN[C-]DQFEK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1303]	87.75	4.89	
	RP[C-]FSALTPDETYVPK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1171,1175]	63.86	5.04	
	YL[C-]DNQDTLSSK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1483,1490]	55.45	5.07	
	RP[C-]FSALTPDETYVPK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1118,1121,1123]	73.68	5.11	
	SH[C-]LAEVEK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1621]	99	5.13	
	YL[C-]DNQDTLSSK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1462,1471,1475]	44.59	5.44	
	SLHTLFGDEL[C-]K	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1361,1365,1369]	99	5.49	
	T[C-]VADESHAG[C-]EK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1350]	99	5.56	
	[C-][C-]AADDKEA[C-]FAVEGPK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1482,1486]	67.29	5.66	
	LKE[C-][C-]DKPLLEK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1355,1358,1363]	95.51	5.95	
	LKE[C+][C+]DKPLLEK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1254,1258]	82.6	5.95	
	YL[C-]DNQDTLSSK	ICAT 5_1 test Mix Dec 16_Dec 30 2003 [1607,1611]	98.98	5.99	
+ LCA_BOSMU	(Q9TSR4) Alpha-lactalbumin precursor (Lactose synthase B pr...		3.37	3.3	0.22

Table 1: Summary of quantification results on TRFE_HUMAN, ALBU_BOVIN, and LCA_BOMSU. Peptide details for ALBU_BOVIN are shown.

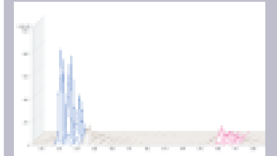


Figure 1a: 3D view of survey scans for peptide LKE[C-][C-]DKPLLEK

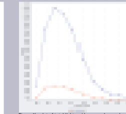


Figure 1b: plot of total light and heavy peak areas by survey scan for peptide LKE[C-][C-]DKPLLEK

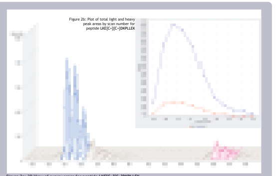


Figure 2b: Plot of total light and heavy peak areas by scan number for peptide LKE[C+][C+]DKPLLEK

Figure 2a: 3D View of survey scans for peptide LKE[C+][C+]DKPLLEK

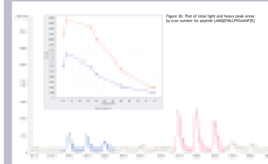


Figure 3b: Plot of total light and heavy peak areas by scan number for peptide LAMQEFMLLPVGAANF[R]

Figure 3a: 3D View of survey scans for peptide LAMQEFMLLPVGAANF[R]

Experimental Results: SILAC

HeLa cells were passaged 2 time (1:10) for one week in media containing 20% dialyzed FBS (Gibco) and custom DMEM containing either standard or heavy ($^{13}C6$) arginine (Specialty media / Sigma). 90-95% labeling was confirmed after the 2nd passage by manual MS analysis. Cells were then harvested and mixed at 1:1, 1:4, and 4:1 ratios according to total cell count. Actual protein ratios may have varied from the intended ratios due to cell counting error (+/- 15%) and varying levels of protein in each cell. A fourth, control sample, was prepared with $^{13}C6$ Arg cells alone. Cells were lysed in 8M urea with 100mM ammonium bicarbonate, reduced with 10mM DTT, and alkylated with 40mM iodoacetamide prior to digest with trypsin (promega) overnight. Digests were concentrated and desalted on a C18 solid phase extraction cartridge (Waters) and a small fraction of each digest was analyzed by nano-LC-MS/MS in a 1 hour data dependant acquisition on a Q-TOF Global (Micromass/Waters) separating on a 15cmX75um C18 column (LC Pa demonstrate the analysis of two peptides.

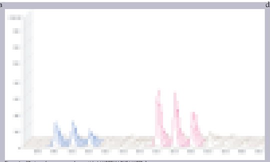


Figure 1a: 3D view of survey scans for peptide LAMQEFMLLPVGAANF[R+]

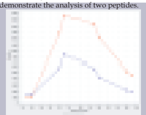


Figure 1b: plot of total light and heavy peak areas by survey scan for peptide LAMQEFMLLPVGAANF[R+]

Accession	Peptide	Source	ID Score (%)	Ratio(light/heavy)	StdDev
+ ENOA_HU	>ENOA_HUMAN (P06733) Alpha-enolase (EC 4.2.1.11) (2-phospho-D-glycerate hydro-lyase) (Non-neural eno		2.61	0.5	0.07
	LAMQEFMLLPVGAANF[R]	sb_130506_SILAC_P2_4-1 [505,506]	64.89	0.42	
	LAMQEFMLLPVGAANF[R+]	sb_130506_SILAC_P2_4-1 [511]	96.67	0.47	
	AAVPSGASTGLYEALEL[R+]	sb_130506_SILAC_P2_4-1 [495]	99	0.56	
+ ACT1_OR	>ACT1_ORYLA (Q98972) Actin, muscle-type (OlMA1);;		1.95	0.8	0
+ TBB1_CR	>TBB1_CRIGR (P69893) Tubulin beta-1 chain (Beta-tubulin isotype I) (Class I beta tubulin);;		0.7	0.6	0
+ KPYM_HU	>KPYM_HUMAN (P14618) Pyruvate kinase, isozymes M1/M2 (EC 2.7.1.40) (Pyruvate kinase muscle isozyme)		0.99	0.69	0
+ ANXA2_R	>ANXA2_RAT (Q07936) Annexin A2 (Annexin II) (Lipocortin II) (Calpactin I heavy chain) (Chromobindin-		1.96	0.56	0
+ HS90B_M	>HS90B_MOUSE (P11499) Heat shock protein HSP 90-beta (HSP 84) (Tumor-specific transplantation 84 kDa		0.98	0.64	0
+ G3P2_HU	>G3P2_HUMAN (P04406) Glyceraldehyde-3-phosphate dehydrogenase, liver (EC 1.2.1.12) (GAPDH);;		1.82	0.74	0.05
+ ACTN1_C	>ACTN1_CHICK (P05094) Alpha-actinin 1 (Alpha-actinin cytoskeletal isoform) (Non-muscle alpha-actinin		1.98	0.84	0
+ HSP7C_B	>HSP7C_BOVIN (P19120) Heat shock cognate 71 kDa protein (Heat shock 70 kDa protein 8);;		0.95	0.57	0
+ LDHA_HU	>LDHA_HUMAN (P00338) L-lactate dehydrogenase A chain (EC 1.1.1.27) (LDH-A) (LDH muscle subunit) (LDH		1.98	0.91	0.14
+ LDHB_HU	>LDHB_HUMAN (P07195) L-lactate dehydrogenase B chain (EC 1.1.1.27) (LDH-B) (LDH heart subunit) (LDH-		0.99	1.02	0
+ H2A_SIP	>H2A_SIPNU (P02270) Histone H2A;;		1.86	0.7	0.08
+ GRP75_H	>GRP75_HUMAN (P38646) Stress-70 protein, mitochondrial precursor (75 kDa glucose regulated protein)		0.99	0.66	0
+ NPM_HUM	>NPM_HUMAN (P06748) Nucleophosmin (NPM) (Nucleolar phosphoprotein B23) (Numatrin) (Nucleolar protein		1.81	0.69	0

Table 2: Summary of quantification results on HeLa cells.

Results and conclusions

PEAKS Q is demonstrated to be robust, easy to use software for quantification of peptides from mass spectrometry. Its protein quantification accruacy is well within experimental error, and significance of abundance changes between samples is easy to see.

References

1. Chen, C., Rogers, I., Filtering out MS/MS spectra of insufficient quality before database searching (ASMS 2006 poster presentation).
2. Chen, C., Rogers, I. Morey, J., Intact Peptide Charge Determination from Ion Trap MS/MS (ASMS 2006 poster presentation).
3. Steen, H., Kuster, B., Fernandez, M., Pandey, A., and Mann, M. (2002) J Biol Chem 277, 1031-1039.
4. Pandey, A., Andersen, J. S., and Mann, M. (2000) Sci STKE 2000, PL1.
5. Steen, H., and Pandey, A. (2002) Trends Biotechnol 20, 361-364.
6. Patton, W. F. (2002) J Chromatogr B Analyt Technol Biomed Life Sci 771, 3-31.
7. Hamdan, M., and Righetti, P. G. (2002) Mass Spectrom Rev 21, 287-302.
8. Summers, S. A., Yin, V. P., Whiteman, E. L., Garza, L. A., Cho, H., Tuttle, R. L., and Birnbaum, M. J. (1999) Ann N Y Acad Sci 892, 169-186.
9. Saltiel, A. R., and Pessin, J. E. (2002) Trends Cell Biol 12, 65-71.
10. Dupont, J., and LeRoith, D. (2001) Horm Res 55 Suppl 2, 22-26.
11. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Electrophoresis 20, 3551-3567.
12. Wang, Q., Bilan, P. J., and Klip, A. (1998) Mol Biol Cell 9, 3057-3069.
13. Holgado-Madruga, M., Emlet, D. R., Moscatello, D. K., Godwin, A. K., and Wong, A. J. (1996) Nature 379, 560-564.
14. White, M. F. (1998) Mol Cell Biochem 182, 3-11.
15. Cobb, M. H., Boulton, T. G., and Robbins, D. J. (1991) Cell Regul 2, 965-978.
16. de Vries-Smits, A. M., Burgering, B. M., Leevers, S. J., Marshall, C. J., and Bos, J. L. (1992) Nature 357, 602-604.
17. Avruch, J. (1998) Mol Cell Biochem 182, 31-48.
18. Hasegawa, T., Takeuchi, A., Miyaishi, O., Xiao, H., Mao, J., and Isobe, K. (2000) Biochem J 347 Pt 1, 55-59.
19. Mann, M., Ong, S. E., Gronborg, M., Steen, H., Jensen, O. N., and Pandey, A. (2002) Trends Biotechnol 20, 261-268.
20. Anderson, N. G., Maller, J. L., Tonks, N. K., and Sturgill, T. W. (1990) Nature 343, 651-653.
21. Payne, D. M., Rossomando, A. J., Martino, P., Erickson, A. K., Her, J. H., Shabanowitz, J., Hunt, D. F., Weber, M. J., and Sturgill, T. W. (1991) Embo J 10, 885-892.
22. Robbins, D. J., and Cobb, M. H. (1992) Mol Biol Cell 3, 299-308.
23. David B. Rorabacher, Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon's "Q" Parameter and Related Subrange Ratios at the 95% Confidence Level, Department of Chemistry, Wayne State University, Detroit, Michigan 4820.

The complete list of references is far too extensive for inclusion on this poster presentation. The authors will be happy to provide a complete list upon request.