New Quantitation Software Package Based on PEAKS Identification

Lei Xin[1]; Gilles Lajoie[2]; Chris Hughes[2]; Bin Ma[2]; Derek Smith[3]
[1]Bioinformatics Solutions Inc, Waterloo, ON
[2]University of Western Ontario, London, ON
[3]University of Victoria-Genome BC Proteomics Centre, Victoria, BC

Bioinformatics Solutions Inc.

Introduction

Isotopic labeling for protein expression analysis has become routine for quantitative proteomics studies. Reagents such as iTRAQ, ExacTag and ICAT are common tools used in this area. Label-free techniques can also be used in cases where isotopic labeling is impractical to perform. As a subsequent step to protein identification, some search engines provide modules for quantitation analysis based on these techniques. Here, we present a new software package designed to automatically quantify proteins from experiments using isotopic labeling or label-free techniques based on PEAKS [1] protein identification results.

Methods

Flexible Label System

PEAKS Studio 5.0 provides a quantitation package that supports quantification data generated using three different techniques.

Isotopic labeling techniques where relative abundances are measured in the MS/MS scan (iTRAQ and ExacTag, etc.)

After digestion, peptides in a sample are specifically labeled. When the peptides are fragmented, part of the label falls off, producing reporter ions. PEAKS Quantitation allows the user to specify any number of reporter ions generated by these reagents and their specific masses. The software will automatically collect the peak information of reporter ions and compute the correspondent ratios.

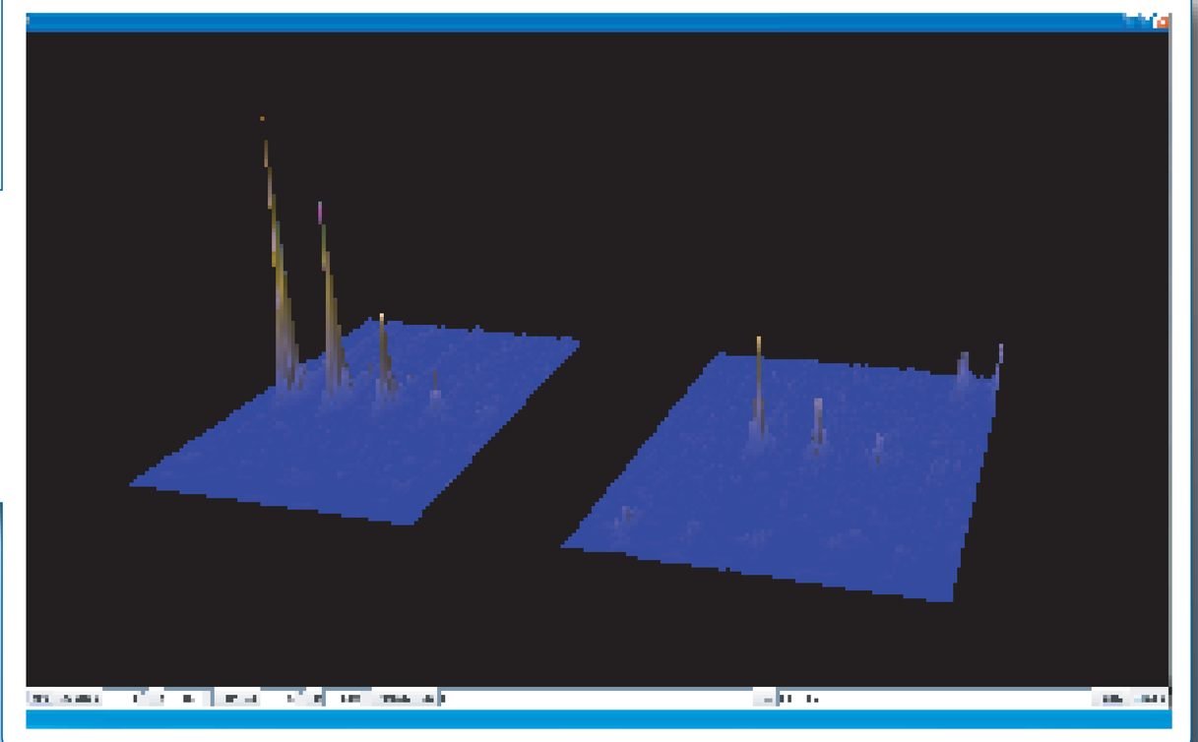
Isotopic labeling techniques where relative abundances are measured in the MS scan (ICAT and SILAC, etc.)

Cells growing in culture are labeled with heavy amino acids and incorporate these amino acids into their proteins. The same peptide from different samples will be labeled with amino acids with different masses and will therefore produce different reporter ions in the MS scan. PEAKS Quantitation allows any number of samples and any number of modifications in each sample. The software will automatically find the peaks corresponding to different samples in the MS scan and compute the correspondent ratios.

Label-free techniques

The third category deals with the data without labels. Protein quantification without isotopic labeling determines protein abundance by comparing peptide signal intensities in sequential MS spectra. The method uses peptide mass and its corresponding elution time to uniquely identify peptide pairs that have significant intensity differences between two LC/MS runs. The peaks of interest are selected and analyzed by MS/MS in order to identify the peptide. As the instrumentation for LC/MS is becoming more reproducible, it is anticipated that label-free techniques will gain in popularity.

Label Free Quantitation: Comparison of a single peptide in different runs with different intensities



General Process

All the three types of quantitation analysis follow the process as below:

1) Data Preprocessing: the data preprocessing for quantitation is quite different from that used in protein identification. New statistical models and algorithms are used a) to merge spectra with similar retention time and m/z values, b) determine charge state when necessary, c) remove poor quality MS/MS scans, d) remove noise, e) centroid and f) deconvolute data within MS/MS scans.

2) Parsing the PEAKS protein identification results: supporting peptides are extracted from the PEAKS database search results and mapped back to the corresponding spectra for each identified protein

3) Protein quantitation: 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 (for ICAT, SILAC) or from the intensities of MS/MS peaks of special mass reagents (for iTRAQ, ExacTag). Statistical methods are then applied to calculate the relative protein abundance and the associated standard deviation.

Removal of Poor Quality MS/MS Spectra Prior to Analysis

Because of chemical and electronic noise as well as very weak signals from peptides, the MS scans contain only a small portion of MS/MS spectra corresponding to identifiable peptides. This typically results in a high rate of false positive identification. 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 quantitation software. The algorithm included in the software had 99.61% accuracy in finding spectra of poor quality [2].

Peptide Charge State Determination

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 that obtained from most ion-trap instruments. If we let a protein identification tool decide the charge, there is an in creased risk of false positive matches, which triples the search time. In PEAKS, an algorithm is built into the software to find precursor charges with high confidence, using low resolution tandem mass spectra data alone. The algorithm takes less than four seconds to correctly assign charge on 313 spectra, with 92% accuracy [3].

Peptide Quantitation

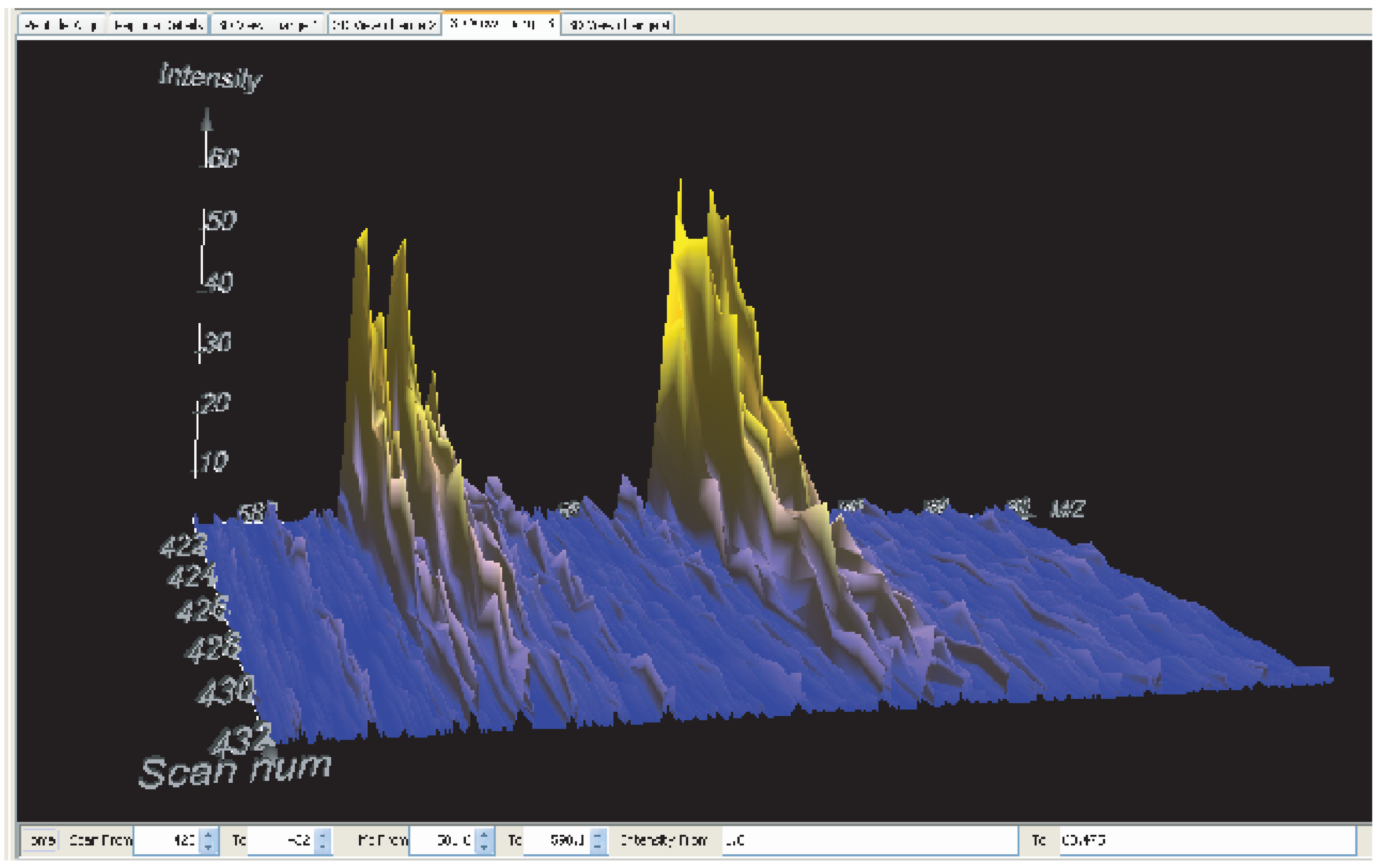
ICAT/SILAC

For ICAT/SILAC labeling, 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 states 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 [4].

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 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.

ICAT Quantitation
Ratio 1:1

	Heavy:Light	SD
P02769 | ALBU_BOVIN	1.146962	0.089292355
C[-]C[-]TKPESER	0.03	Outlier
C[-]C[-]TESLVNR	0.26	Outlier
EAC[+]FAVEGPK	1.0103922	
LKPDPNTLC[+]DEFK	1.056511	
SHC[+]LAEVEK	1.0819472	
YLC[+]DNQDTLSSK	1.1508412	
YLC[-]DNQDTLSSK	1.1607366	
QNC[-]DQFEK	1.2042723	
QNC[+]DQFEK	1.2077675	
GAC[+]LLPK	1.2794976	
C[+]ASLQK	2.21	Outlier
C[-]ASLQK	2.21	Outlier

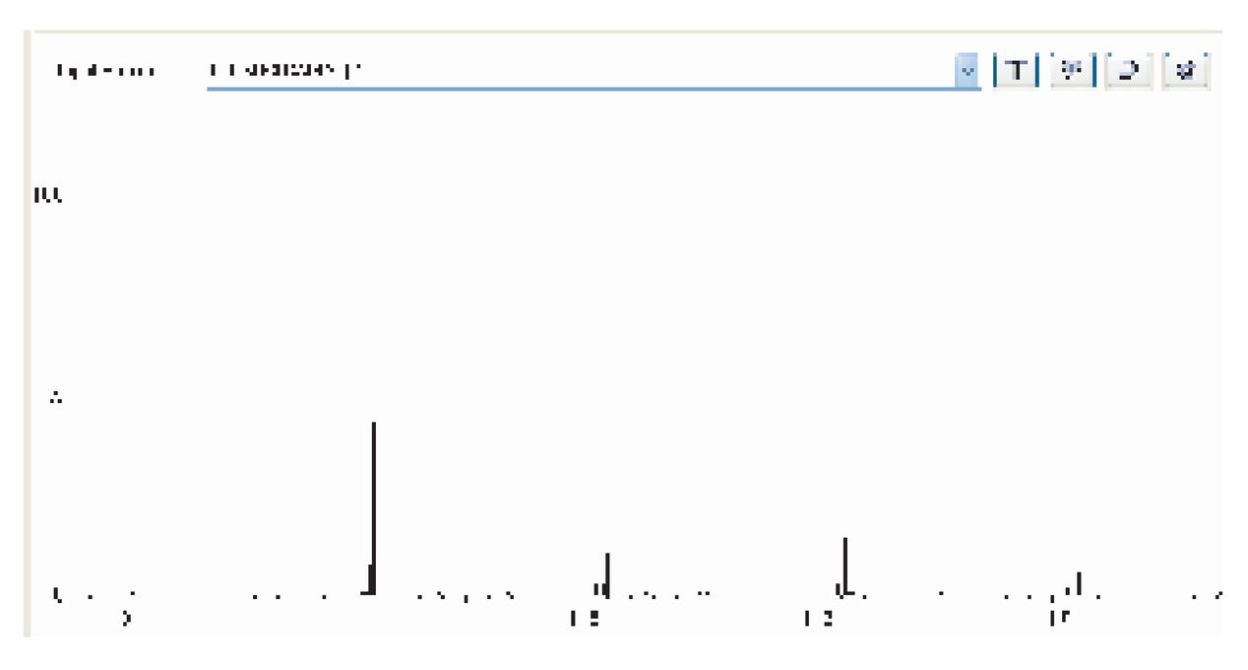


iTRAQ

For iTRAQ labeling, the ratio of labeled peptides is determined at the MS/MS level. The first stage for iTRAQ quantitation analysis is to associate each reporter ion with a group of peaks with specific m/z values in the MS/MS scan. In the second stage, a statistical algorithm is used to decide significant peaks in each group. Then all the significant peaks in each group are centralized separately. The abundance of a peptide is obtained from the ratios of centralized peak intensities for reporter ions.

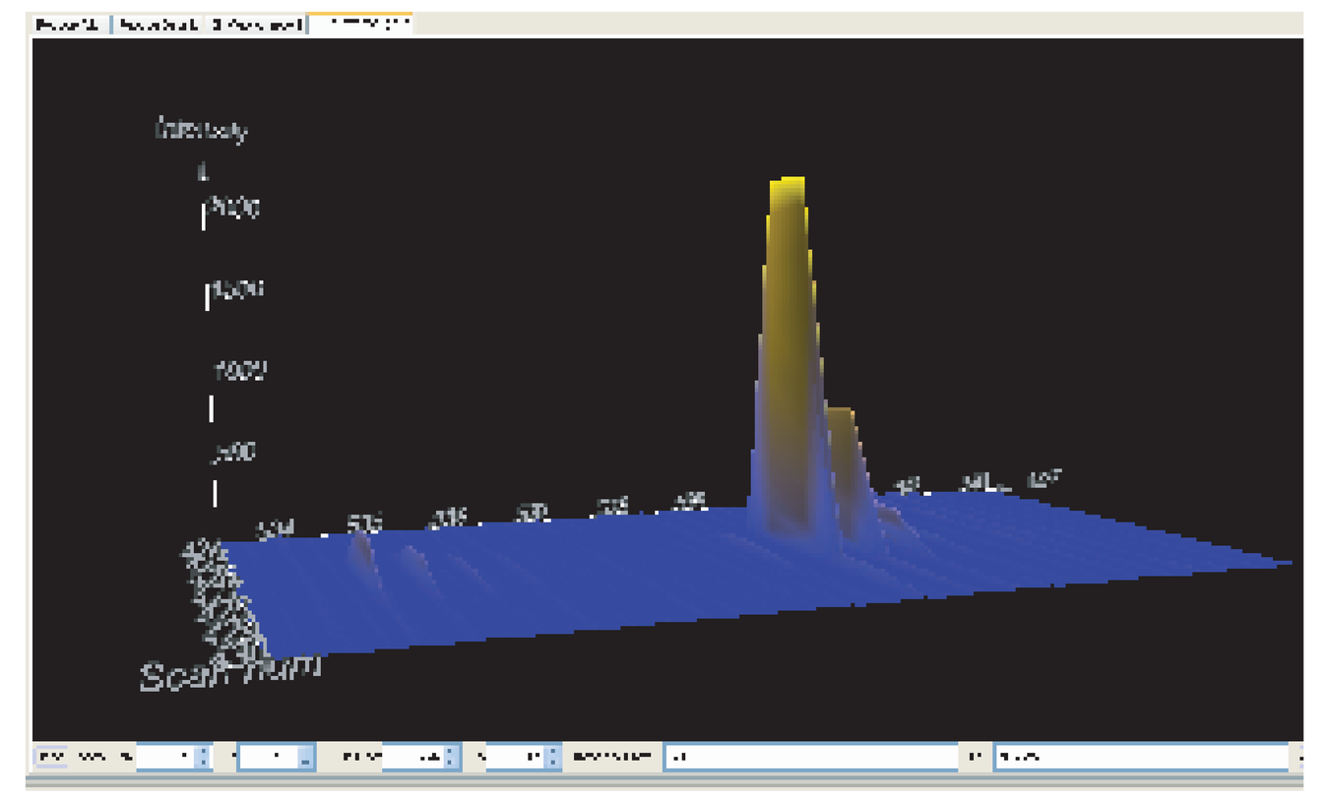
Removal of Outliers

If we assume that the abundance ratio of any peptide belonging to a protein is representative of the abundance ratio of that protein, then in theory it follows that all peptides belonging to a protein should have the same abundance ratio. An outlier is a data point within a data set where these values are statistically different from the main body of the data. Outliers have to 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 underlying distribution.



ICAT Quantitation
Ratio 4:1

	Heavy:Light	SD
P02769 | ALBU_BOVIN	1.146962	0.089292355
YNGVFQEC[+]C[+]QAEDK	0.18	Outlier
C[-]C[-]TKPESER	0.99	Outlier
C[+]C[+]TKPESER	0.99	Outlier
C[-]C[-]TESLVNR	2.75	Outlier
YLC[-]DNQDTLSSK	3.55	Outlier
LKPDPNTLC[-]DEFK	3.92	
MPC[-]TEDYLSLLLNR	4.04	
RPC[-]FSALTPDETYVPK	4.65	
SHC[-]LAEVEK	5.72	
C[-]ASLQK	5.76	
EAC[+]FAVEGPK	5.87	
EAC[-]FAVEGPK	5.93	
DDPHAC[-]YSTVFDK	6.06	
QNC[-]DQFEK	6.16	
LC[-]VLHEK	6.39	



Visual Verification

No matter what kind of technique is used, protein quantification analysis is still a difficult task. Improper parameter settings will often cause high deviation from the actual ratios. In this manner, visual inspection of raw spectra becomes very important. It allows researchers the chance to adjust the parameter setting to gain much more accurate ratios. The PEAKS quantification package provides a 2D view of the MS/MS spectrum as well as a 3D view of the parent scan for each identified peptide. Thus the actual ratios can be easily verified by inspecting the correspondent raw spectra.

iTRAQ Quantitation

Proteins from E. coli BL21 cells were digested in trypsin at a protein to enzyme ratio of 10:1. The digests were labeled with ITRAQ reagent according to the manufacturer's protocol at a protein to reagent ratio of 1:5 and 1:10. These samples were then separated by SCX high performance liquid chromatography (HPLC) and analyzed by nano-ESI MS/MS using a Tempo nanoflow MDLC system coupled to a QSTAR Elite Qq-TOF system equipped with a NanoSpray source and heated interface.

Protein	115:114	SD	116:114	SD	117:114	SD
Actual ratio	1:10		1:1		1:10	
P0A6N3 | EFTU_ECO57	0.22672254	0.049005963	0.82194555	0.1736066	0.14183252	0.028204044
Peptides	115:114		116:114		117:114	
G[1]LTLNTSHVEYDTPTR		0.24		0.93		0.26
G[1]LTLNTSHVEYDTPTR		0.19		1.08		0.19
G[1]LTLNTSHVEYDTPTR		0.2		1.01		0.18
G[1]LTLNTSHVEYDTPTR		0.26		1.2		0.21
G[1]LTLNTSHVEYDTPTR		0.15		0.98		0.15
G[1]LTLNTSHVEYDTPTR		0.23		1.17		0.03
G[1]LTLNTSHVEYDTPTR		0.2		0.95		0.15
G[1]LTLNTSHVEYDTPTR		0.27		0.86		0.16
G[1]LTLNTSHVEYDTPTR		0.17		1.04		0.12
G[1]LTLNTSHVEYDTPTR		0.2		0.77		0.11
G[1]LTLNTSHVEYDTPTR		0.27		0.65		0.08
A[1]LEGDAEWEAK[2] 0.28		0.93		0.11		
E[1]LLSQY[3]DFPGDDTPLVR		0.19		0.65		0.03
E[1]LLSQY[3]DFPGDDTPLVR		0.24		0.57		0.09
T[1]YGGAAR		0.18		1.07		0.15
E[1]LLSQYDFPGDDTPLVR		0.33		0.74		0.3
M[1]VVTLLHPLAMDDGLR		0.16		0.96		0.14
T[1]TLTAALTTVLAK[2]		0.2		0.95		0.1
T[1]TLTAALTTVLAK[2]TYGGAAR		0.08		1.5		0
T[1]TLTAALTTVLAK[2]TYGGAAR		0.07		0.89		0.04
A[1]LDK[2]PFLLPLEDVFSLSGR		0.31		0.92		0.15
A[1]LDK[2]PFLLPLEDVFSLSGR		0.28		0.69		0.13
A[1]LDK[2]PFLLPLEDVFSLSGR		0.29		0.74		0.12
A[1]LDK[2]PFLLPLEDVFSLSGR		0.29		0.52		0.01
A[1]LDK[2]PFLLPLEDVFSLSGR		0.16		0.44		0.01
A[1]LDK[2]PFLLPLEDVFSLSGR		0.2		0.73		0.16
A[1]LDK[2]PFLLPLEDVFSLSGR		0.22		0.73		0.02
A[1]LDK[2]PFLLPLEDVFSLSGR		0.1		0.27		0.13
A[1]LDK[2]PFLLPLEDVFSLSGR		0.09		0.51		0.09
A[1]LDK[2]PFLLPLEDVFSLSGR		0.17		0.54		0.08
A[1]LDK[2]PFLLPLEDVFSLSGR		0.18		0.61		0.14
A[1]LDK[2]PFLLPLEDVFSLSGR		0.29		0.86		0.02
A[1]LDK[2]PFLLPLEDVFSLSGR		0.08		0.92		0.15

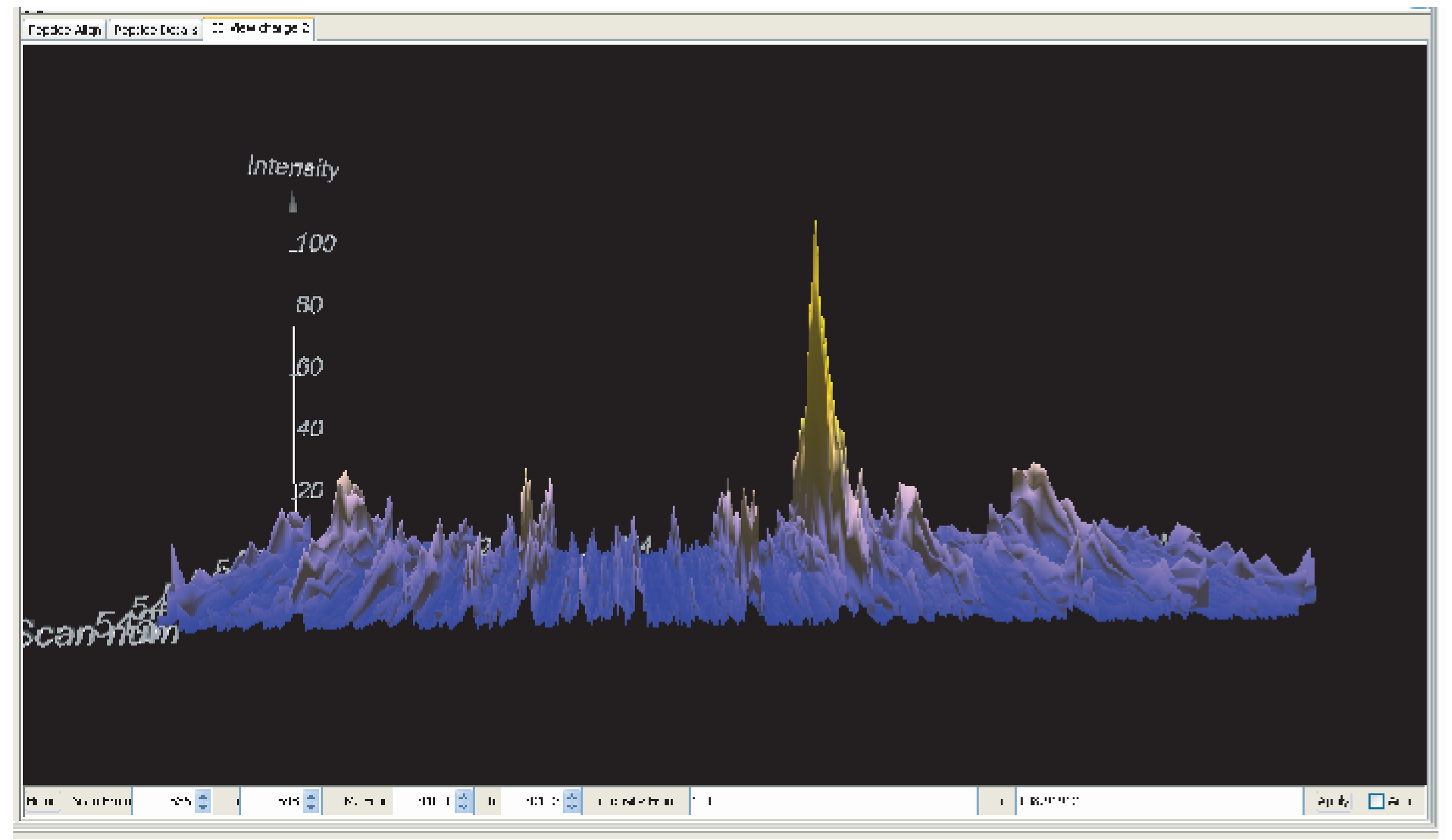
SILAC Quantitation

E14K mouse embryonic stem cells (mESCs) were cultured in SILAC media. The light and heavy labelled isotopic samples were mixed with ratio of 4:1 based on total protein concentration prior to digestion. The mixture was digested with trypsin. The resulting peptides were extreacted using solid phase extraction cartridge and then analyzed by LC MS and MS/MS on a Waters Q-TOF Ultima.

Protein	Protein Ratio (H:L)	SD
P62806 | H4_MOUSE	5.0706797	1.3086568
Peptide	Heavy:Light	isOutlier
LSGLLYEETRGVLK	0.1	Outlier
VLGAR	0.11	Outlier
DNLQGLTK[2]PALR[1]R[1]	2.71	
TVTAMDVVYALK[2]R[1]	3.02	
DAVTYTEHAK[2]	3.74	
LSGLLYEETR[1]GVLK[2]	3.74	
TVTAMDVVYALK[2]	4.21	
VFLENVLR[1]	5.1	
VLLDLQDNLNLHLLAR[1]	5.32	
DNLQGLTK[2]PALR[1]	5.83	
VFLENVLR[1]DAVTYTEHAK[2]	6	
LSGLLYEETR[1]GVLK[2]	7.21	
TVTAMDVVYALK[2]R[1]	11.65	Outlier

Protein	Protein Ratio (H:L)	SD
Q8CGP6 | H2A1H_MOUSE	3.4973903	0.9250561
Peptide	Heavy:Light	isOutlier
AGLQFPVGR	0.32	Outlier
LLR[1]K[2]GNYSER[1]	2.59	
HLQLALRNDEELNKLLGR	2.6	
HLQLALR[1]NDEELNK[2]LLGR[1]	3.04	
NDEELNKLLGR	3.71	
HLQLALR[1]NDEELNK[2]LLGR[1]	4.3	
NDEELNK[2]LLGR[1]	4.83	

Protein	Protein Ratio (H:L)	SD
P58252 | EF2_MOUSE	3.842803	0.98382316
Peptide	Heavy:Light	isOutlier
LLEK[2]LDLK[2]	2.4	
GPLMMYLSK[2]	2.79	
K[2]GLK[2]EGLPALDNFLDK[2]L	3.78	
TLLMMGR[1]	3.79	
GHVFEESQVAGTPMFVVK[2]	4.85	
K[2]GLK[2]	5.16	
EGLPALDNFLDK[2]L	6	
YLAEK[2]	10.53	Outlier
AR[1]YLAEK[2]YEWDVAEAR[1]	38.54	Outlier



Results and Conclusions

PEAKS quantitation is demonstrated to be robust, easy to use software for quantitation of peptides from mass spectrometry. Its protein quantitation accuracy is well within experimental error, and significance of abundance changes between samples is easy to see. This is available as part of the PEAKS Studio 5.0 proteomic softwaer.

Reference

1. 1. Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, Gilles Lajoie. PEAKS: Powerful Software for Peptide De Novo Sequencing by MS/MS. Rapid Communications in Mass Spectrometry, 17(20):2337-2342. 2003. Early version appeared in 50th ASMS Conference 2002.
2. Clark Chen, Iain Rogers, Filtering out MS/MS spectra of insufficient quality before database searching (ASMS 2006 poster presentation).
3. Clark Chen, Iain Rogers, Intact Peptide Charge Determination from Ion Trap MS/MS(ASMS 2006 poster presentation).
4. Weijie Yang, Clark Chen, Iain Rogers, Bin Ma, Gilles Lajoie, Derek Smith, PEAKS Q: Software for MS-based quantification of stable isotope labeled peptides (Bioinformatics Solutions Inc., Genome BC Proteomics Centre, University of Western Ontario) ASMS 2006 poster WP531

Protein	115:114	SD	116:114	SD	117:114	SD
Actual ratio	5:1		1:1		5:1	
Q8FBM1 | METE_ECOL6	4.6909356	1.1420326	1.1536548	0.24317375	4.6618843	0.9109644
Peptides	115:114	116:114	117:114			
G[1]WPETR		7.22	1.56	8.22		
G[1]W[4]PETR		3.97	1.03	3.49		
L[1]DFK[2]		1.66	1.05	1.62		
V[1]HNPAVEK[2]R		9.33	2.33	9.33		
V[1]HNPAVEK[2]R		4.33	1.39	3.78		
V[1]HNPAVEK[2]R		1.75	0.63	0.63		
V[1]HNPAVEK[2]R		14	4	8.67		
V[1]HNPAVEK[2]R		10	2	0.5		
V[1]HNPAVEK[2]R		1	1	2		
V[1]K[2]PVLLGPLTY[3]LW[4]-LGK[2]VK[2]		4	1.23	3.77		
V[1]HNPAVEK[2]		4.17	1.71	3.38		
L[1]GLDVLVHGEAER		4.83	0.83	4.92		
L[1]GLDVLVHGEAER		3.73	1.07	3.9		
L[1]GLDVLVHGEAER		6.32	2.26	4.84		
L[1]GLDVLVHGEAER		8.5	2.1	6.1		
L[1]GLDVLVHGEAER		3.7	1.22	3.87		
L[1]GLDVLVHGEAER		7.86	2	7.57		
L[1]GLDVLVHGEAER		40	11	52		
L[1]GLDVLVHGEAER		9.4	1.2	9		
L[1]GLDVLVHGEAER		6.09	1.13	5.26		
L[1]GLDVLVHGEAER		3.11	1.44	3.56		
L[1]GLDVLVHGEAER		5.8	0.8	6		
L[1]GLDVLVHGEAER		7.75	0.75	5		
L[1]GLDVLVHGEAER		7.33	3.67	7		
L[1]GLDVLVHGEAER		4.8	1	6.2		
L[1]GLDVLVHGEAER		5.54	1.29	5.06		
L[1]GLDVLVHGEAER		3.5	1	4.33		
L[1]GLDVLVHGEAER		9.8	3.6	8.6		
L[1]GLDVLVHGEAER		4.25	1.75	6		
L[1]GLDVLVHGEAER		6.6	1	4.2		
L[1]GLDVLVHGEAER		3.15	1.15	4.08		
L[1]GLDVLVHGEAER		4	1	7.5		
L[1]GLDVLVHGEAER		3.6	1	4		
L[1]GLDVLVHGEAER		3.75	1.5	6.5		
L[1]GLDVLVHGEAER		6	1	4.33		
L[1]GLDVLVHGEAER		10	9	12		
L[1]GLDVLVHGEAER		4	1	4		
L[1]GLDVLVHGEAER		2.33	2.33	2.33		
W[1]FNTNYHYM[4]VPEFVK[2]		9.53	3.6	7.4		
W[1]FNTNYHYMVPEFVK[2]		18	5.78	13.56		
H[1]STR		4.5	1	1.75		
V[1]K[2]		0	0	0		
V[1]K[2]		0	0	0