BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 2. 156–165
Advance Access publication September 28, 2007
doi:10.1093/bib/bbm046

Computational methods for the comparative quantification of proteins in label-free LC^n-MS experiments

Jason W. H. Wong, Matthew J. Sullivan and Gerard Cagney

Submitted: 28th June 2007; Received (in revised form): 7th September 2007

Abstract

Liquid chromatography (LC) coupled to electrospray mass spectrometry (MS) is well established in high-throughput proteomics. The technology enables rapid identification of large numbers of proteins in a relatively short time. Comparative quantification of identified proteins from different samples is often regarded as the next step in proteomics experiments enabling the comparison of protein expression in different proteomes. Differential labeling of samples using stable isotope incorporation or conjugation is commonly used to compare protein levels between samples but these procedures are difficult to carry out in the laboratory and for large numbers of samples. Recently, comparative quantification of label-free LC^n-MS proteomics data has emerged as an alternative approach. In this review, we discuss different computational approaches for extracting comparative quantitative information from label-free LC^n-MS proteomics data. The procedure for computationally recovering the quantitative information is described. Furthermore, statistical tests used to evaluate the relevance of results will also be discussed.

Keywords: *comparative quantification; mass spectrometry-based proteomics; label-free quantification; spectral counting; ion chromatogram extraction*

INTRODUCTION

For the best part of the past two decades, chromatographic separation of peptides coupled to mass spectrometry (MS) has been extensively used to study the proteome [1]. Due to the complexity of the proteome, liquid chromatography (LC) is used to fractionate proteolytic peptides so that mixtures of lower complexity can be introduced to the instrument over time, thus increasing the efficiency of detection and identification by tandem mass spectrometry (MS/MS). The introduction of orthogonal peptide separation techniques coupled to the mass spectrometer, such as multidimensional protein identification technology (MudPIT) [2–4] has further increased the potential throughput of MS/MS experiments, enabling the identification of 100s or 1000s of proteins from a single sample.

The identification of proteins remains the primary application of LC^n-MS/MS proteomics experiments. However, protein identification is often only the first step in proteomics studies. The ability to quantify levels of proteins present provides an extra dimension of information. One potential drawback of MS is that the data generated is not directly quantitative. The efficiency of the ionization process is dependent on the molecular composition of each molecule. For instance, the apparent ion intensity of different peptides of the same concentration

Corresponding author. Jason W. H. Wong, Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland. Tel: +35317166945; Fax: +35317166962; E-mail: jason.wong@ucd.ie

Jason W. H. Wong is a Postdoctoral research fellow at the Conway Institute of Biomolecular and Biomedical Research at University College Dublin. He received his DPhil from the University of Oxford in 2006. His research interests include bioinformatics, chemical proteomics and mass spectrometry.

Matthew J. Sullivan is an informatics specialist at the Conway Institute for Biomolecular and Biomedical Research at University College Dublin. His background encompasses telecommunications and bioinformatics, and is currently involved in mass spectrometry proteomics data analysis research.

Gerard Cagney is a principle investigator in proteomics at the Conway Institute of Biomolecular and Biomedical Research at University College Dublin. His research interests include mass spectrometry-based proteomics, protein interaction networks and bioinformatics.

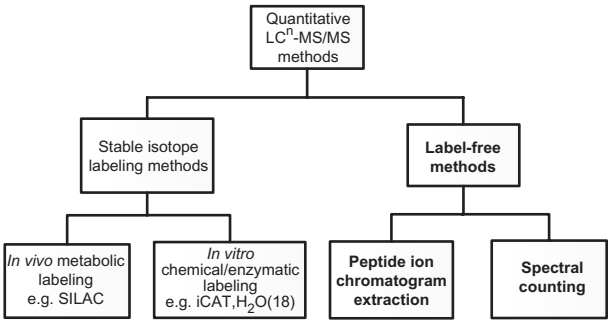

Figure 1: Strategies for comparative proteomics by LC^n-MS. The label-free methods in bold are the subject of this review.

is typically different due to amino-acid composition differences. As a result, the height or area of peaks from two different ions cannot be directly compared without taking into consideration the composition of the ion that has been the subject of some recent research [5–7]. Other quantification issues such as inter-LC^n-MS/MS experimental variations and ion suppression effects [8] can also confound attempts to accurately quantify protein levels. Nevertheless, while the mass spectrometer does not provide true quantitative data, with careful experimental design and data analysis, comparative quantification is an option for researchers.

There are currently two main approaches to comparative quantification by LC^n-MS/MS (Figure 1). One approach is the incorporation of stable isotopes into one or more of the samples being studied [9]. This may be carried out *in vivo* by stable isotope-containing amino acids introduced in the cell culture media (SILAC) [10, 11], for example carbon-13 substituted arginine [12]. Alternatively, the stable isotopes can be incorporated *in vitro*, chemically [9, 13, 14] or enzymatically such as using oxygen-18 water when performing proteolysis with trypsin [15]. Peptides from the sample containing the stable isotope will then be 'heavier' when analyzed simultaneously with the control sample in the mass spectrometer, thereby allowing them to be distinguished. Because the molecular composition of the 'heavier' ion does not change, the ionization efficiently will remain the same and therefore the quantities of identical peptides can be directly compared. Disadvantages of using stable isotope labeling include requirement that cells be culturable (in the case of SILAC), and while it is possible to differentially label up to eight biologically different samples using the iTRAQ® Reagent-8Plex kit, the high cost renders routine application prohibitive. In terms of MS data acquisition, isotope labeling is also more challenging as the number of peptides co-eluting will increase, hence possibly reducing the overall peptide coverage. Subsequent computational analysis will further require specific tools for recovery of differentially labeled peptides.

The second approach to comparative quantification by LC^n-MS/MS is to take a 'label-free' approach. The basis of such approaches is to make the assumption that under well-controlled conditions with sufficient data redundancy, identical peptides across different LC^n-MS/MS experiments can be compared directly. This has been made possible through technical advances in high-performance (HP) LC systems, mass spectrometers with higher resolution and scanning rates, as well as the use of robots for sample preparation. As such, studies have shown that peptide ion counts across control experiments can be highly reproducible [16–19] and results are comparable to stable-isotope labeling approaches [20]. Label free comparative quantification studies have gained popularity in recent years [21–28]. The major advantages of the label-free approach are that it typically does not require any extra steps in experimental procedures and furthermore, comparative quantification can be performed across many samples simultaneously. Generally, the major challenge lies in the computational and statistical analysis of the results.

In this review, two popular computational approaches, extraction of peptide ion intensities [16–19, 29] and spectral counting [30, 31] for performing comparative quantitative analysis of LC-MS proteomics experiments are described (Figure 2). The necessary computational algorithms and tools available for acquiring the comparative data will be discussed. Statistical tests for evaluating the significance of the comparative results will also be covered in this review. Finally, studies comparing the two techniques will be discussed with emphasis on the strength and weaknesses of each technique. A list of publicly available software relevant to this review can be found in Table 1.

COMPARATIVE QUANTIFICATION BY PEPTIDE ION INTENSITY

The height or area of a peak at a particular mass-to-charge ratio (m/z) from a mass spectrum is a measurement of the number of ions detected

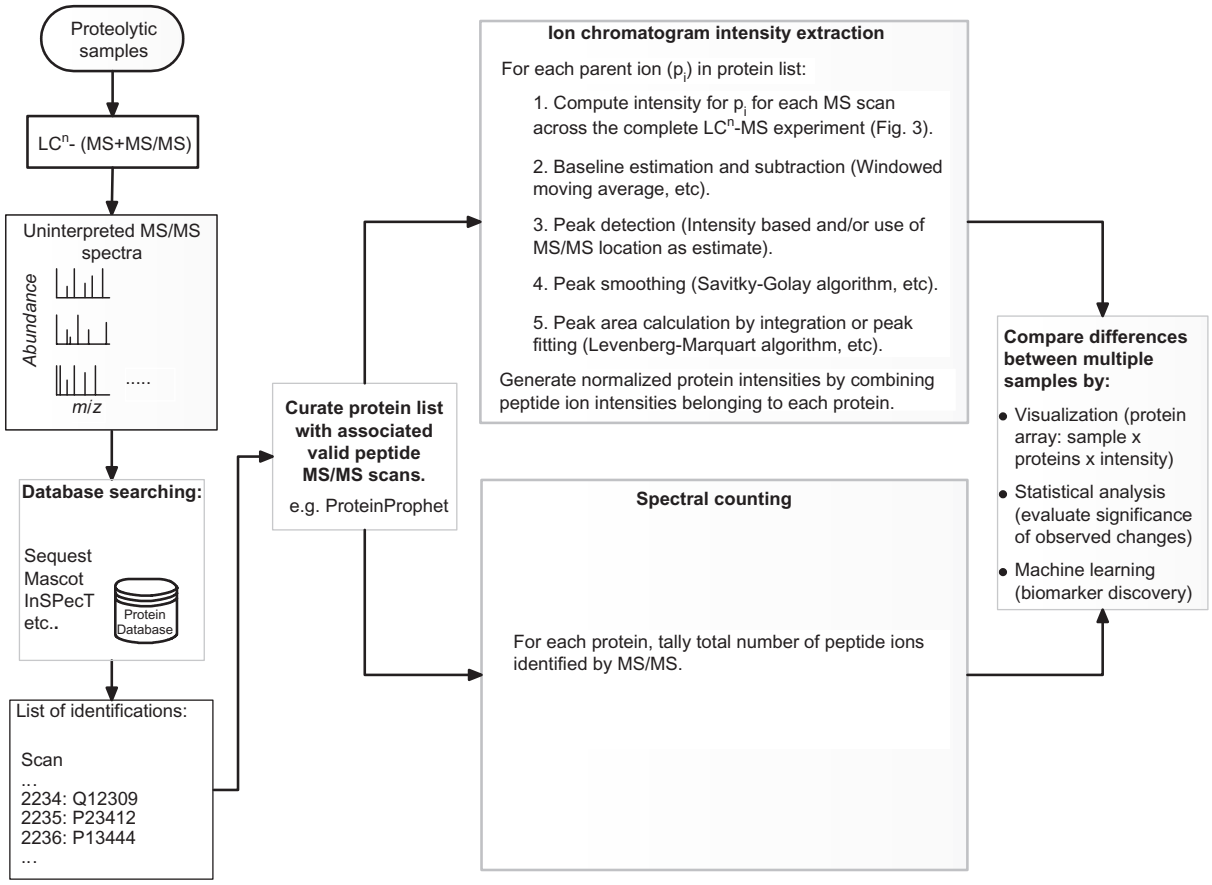

Figure 2: Schematic diagram for label-free quantitative proteomics by LC^n-MS/MS. The computational steps are headed in bold.

Table I: List of publicly available or popular commercial software relevant to label-free protein quantification by LC^n-MS/MS

Software	Website	Availability	Platform/language
LC-MS Imagining based quantification software			
MSight	www.expasy.org/Msight	Freeware	Windows
msInspect	proteomics.fhcrc.org/CPL/msinspect.html	Open source	Java
OpenMS	open-ms.sourceforge.net	Open source	C++
SpecArray	tools.proteomecenter.org/SpecArray.php	Open source	C++
XCMS	masspec.scripps.edu/xcms	Open source	R
Ion chromatogram extraction related software			
ASAPratio	tools.proteomecenter.org/ASAPRatio.php	Open source	C++
MSQuant	msquant.sourceforge.net	Open source	Windows (.NET)
RelEx	fields.scripps.edu/relex	Freeware	Windows
XPRESS	tools.proteomecenter.org/XPRESS.php	Open source	C++
Spectral counting related software			
NoDupe	fields.scripps.edu/nodupe	Freeware	Java
PeptideProphet	tools.proteomecenter.org/PeptideProphet.php	Open source	C++
ProteinProphet	tools.proteomecenter.org/ProteinProphet.php	Open source	C++
Database searching software			
GutenTag	fields.scripps.edu/GutenTag	Freeware	Java
InsPecT	peptide.ucsd.edu/inspect.html	Open source	C++
MASCOT	www.matrixscience.com	Commercial	All
Sequest	fields.scripps.edu/sequest	Commercial	All
X!Tandem	www.thegpm.org	Open source	C++

by the mass spectrometer at any given time. This is typically known as the ion abundance. In a LC-MS experiment of a complex digested protein mixture, peptide ions are separated by a chromatographic gradient followed by mass analysis. The result can be visualized as a two dimensional (2D) image map with retention time and *m/z* on the *x* and *y* axis, respectively and ion count as the intensity [19, 32]. This type of 2D image is similar to those of 2D gel electrophoresis (2DE) of complex protein mixtures [33] and as a result similar methods can be applied for comparative quantification [34, 35]. In fact, some tools such as MSight [36], which implements the Melaine gel image analysis system [37], were originally designed for 2DE for spot detection and quantification. A number of other tools such as XCMS [38], SpecArray [29], msInspect [39] and OpenMS [40] do not rely on image analysis but rather automatically detect potential peptide features directly from the raw data and extract the corresponding quantitative information. To measure the total ion abundance for any peptide ion within a LC-MS experiment, the ion intensity is integrated over time. This process is computationally referred to as ion extraction resulting in an extracted ion chromatogram (Figure 3). If a particular peptide concentration lies within the dynamic range of the instrument for the experiment, a peak will be present in the extracted chromatogram that rises above background assuming that the peptide is readily ionized and not suppressed by other ions. The area of this peak represents the total ion abundance for the peptide. For comparative quantification, the extracted ion chromatogram is computed for each peptide across all samples.

LC-MS and 2DE are complementary for quantitative purposes, although LC-MS offers the advantage of enabling MS/MS data to be acquired in an automated manner. The majority of tools previously mentioned are generally designed to extract peptide features without knowledge of the peptide identity. In many cases, quantitative information is captured for peak features that may represent peptides that cannot be identified (due to absence of the protein sequence in a database or because the peptide contains a posttranslational modification). However, by first identifying peptides of interest by database search tools such as Sequest [41], Mascot [42] or InsPect [43], quantitative information for those peptides can be extracted from the raw data. Peak selection is facilitated

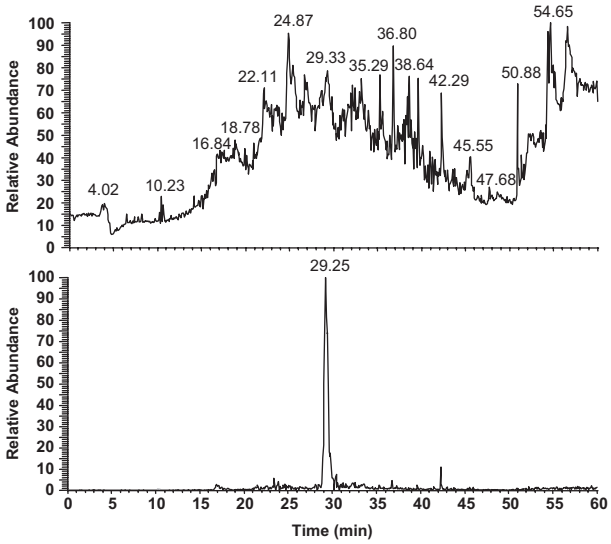

Figure 3: Example chromatogram of a typical LC^n-MS/MS analysis of a tryptically digested proteome. Peptides were separated on a C18 reverse phase column followed by MS and data dependent MS/MS analysis using a ThermoFinnigan LTQ mass spectrometer. The top shows the total ion chromatogram for the run, while the bottom is an extracted ion chromatogram for a particular peptide showing a significant peak. The area of this peak represents the total ion intensity of the peptide.

in these cases by the aligning of scan numbers yielding peptide identification to the elution profile. Furthermore, combining the ion currents or peak areas of different peptides originating from the same protein allows measurement error to be estimated and provides greater confidence when performing comparative quantification across samples.

A number of tools can be used to extract peptide ion intensities following identification such as MSQuant [44] (originally designed for quantifying stable isotope datasets) and Serac PeakExtractor [45]. Alternatively, publicly available tools such as ASAPratio [46], XPRESS [47] and RelEx [48] that have been designed specifically for the comparative quantification of stable-isotope labeled data using the extracted ion chromatogram method may also be modified and integrated into pipelines to compute intensities for specific peptide ions. One problem typical in MS/MS experiments is that the parent ion scans that are used to determine peak areas are interrupted by the MS/MS events, resulting in a serrated profile that causes peak finding algorithms to perform poorly. This means that a manual confirmation step must be used to check

each area assignment. RelEx overcomes some of these problems by using a least squares correlation approach based on the slope of the eluting peak that is largely independent of MS/MS events [48].

The choice of workflow is likely to depend on the type of experiment being performed and the instrument setup. For biomarker discovery studies with multiple replicates, it may be desirable to perform peptide feature extraction prior to identification. However, for datasets where the constituents of the samples are relatively well known or where a specific protein or class of protein is being targeted, annotation of the peptides first may allow comparative quantification with greater specificity. In either case, once a list of proteins with quantitative data has been drawn up for each sample, the lists are clustered to form a matrix of intensities analogous to a protein/peptide array [49] to enable comparative quantification [29], where the sample label and the identified proteins form the axes of the matrix.

COMPUTATIONAL AND STATISTICAL ISSUES RELATING TO ION INTENSITY COMPARATIVE QUANTIFICATION

The first stage of computing the extracted intensity for a particular ion involves the generation of the extracted ion chromatogram by computing the intensity of the given parent ion for each mass spectrum acquired through a LC-MS run. Once the chromatogram has been formed, a series of spectral processing methods such as baseline subtraction, peak picking and smoothing are preformed before the computation of the ion intensity by peak integration or peak fitting (Figure 2).

There are a number of parameters that need to be considered when extracting peptide quantitative data. Generally, this will include determining the *m/z* window over which an ion is defined, the retention time window where the ion is eluted and also the baseline intensity across all ions. The choice of *m/z* window is typically dependent on the resolution of the instrument. The retention time window and the baseline are generally estimated using peak fitting algorithms [29, 38, 46, 47].

When combining the quantitative data across a set of samples, the process is reasonably straightforward if the identity of the peptides or proteins are known, because they can be mapped between samples using scan numbers or elution times. But where the peptide identity is not known, the process can be computationally more challenging. Algorithms may be necessary to correct for shifts in retention time across samples prior to peptide ion intensity extraction [38, 39, 50–52]. Although not a computational issue, it is also worth mentioning that the use of mass spectrometers equipped with high resolution mass analyzer such as the Fourier transform ion cyclotron resonance analyzer [53] or the OrbiTrap analyzer [54] increase the reliability of mapping peptides across samples because of the more narrow *m/z* determinations [32].

When peptides have been identified, it is possible to apply the Student's *t*-test [55] by taking the mean and SD of ratios between the ion intensity of peptides found for each protein. This will result in a two-tailed *P*-value for each protein that can be used to evaluate the significance in change. In the case where there a very few peptides identified for a protein, the Mann–Whitney U-test [56] may be more appropriate as it is less likely to be affected by outliers. In either case, it is possible to apply the Dixon's Q-test to remove outlier peptide ion ratios prior to the test [45, 48].

An alternative to the application of statistical methods to directly evaluate the significance of changes in the intensity of each protein is the use of machine learning methods to discover proteins that can be used to classify samples into distinct classes. Machine learning is a powerful technique that enables the automated discovery to single or multiple proteins in combination that distinguish two or more classes of samples in many types of biomedical data [57]. Where the sample type is not known or there are too few samples to train the machine-learning algorithm, unsupervised clustering techniques such as *k*-means may be applied. Supervised learning methods such as neural networks and support vector machines are generally more powerful but require more samples. The major limitation of machine learning is that for it to be applied successfully, a significant number of replicate samples (relative to the number of proteins identified for each sample) are generally required to obtain reliable results [58]. With insufficient samples, unsupervised learning methods may form incorrect clusters while supervised methods will be prone to over-learning. To this end, since current differential proteomics experiments typically involve a relatively low number of samples, the use of machine learning in conjunction with comparative LC^n-MS/MS experiments remains limited.

COMPARATIVE QUANTIFICATION BY SPECTRAL COUNTING

The spectral count for a protein refers to the number of MS/MS spectra acquired from proteolytic peptide ions for that protein during a LC-MS/MS run. The premise of the method is that the more abundant the peptide, the more likely it will be selected for MS/MS analysis. Software that directs the MS instrument to acquire spectra in a data dependent manner will obviously influence the spectral counts. However, Liu *et al*. [31] showed that spectral counting is highly reproducible and is sensitive to protein abundance changes. Furthermore, in controlled experiments, it was found that the correlation of protein abundance with spectral count is superior to that of protein sequence coverage or peptide count [31, 59].

To facilitate spectral count, all MS/MS spectra are first interpreted using database search programs and then all spectra belonging to the same peptide ion or protein is tallied. Due to its ease of implementation, no specific tools or algorithms have been developed specially for spectral counting but tools such as PeptideProphet [60] and ProteinProphet [61] will automatically report peptide/protein counts as part of their output.

An alternative approach could be to first cluster uninterpreted MS/MS spectra based on their similarity using spectral comparison algorithms applied in tools such as NoDupe [62]. This approach will reduce the number of MS/MS spectra to be searched, but at the same time peptide and subsequently spectral counts can still be obtained. We are currently not aware of any tools or analysis pipelines that calculate spectral counting in this way, since NoDupe was originally designed primarily to reduce searching of similar spectra and not for spectral counting. Nevertheless, the adaptation of analysis pipeline workflows to incorporate spectral similarity comparison algorithms should not be difficult and will likely yield increased pipeline throughput.

COMPUTATIONAL AND STATISTICAL ISSUES RELATING TO SPECTRAL COUNTING COMPARATIVE QUANTIFICATION

It is important to note that when converting raw files to files in DTA format for Sequest analysis, tools such as extractMSn from ThermoFinnigan's XCalibur package pools MS/MS spectra that are deemed to be similar due to parent mass, thereby invalidating spectral counting. However, if files have already been converted to mzXML format alternate conversion tools such as mzXML2other from the Institute of Systems Biology, Seattle [63] are commonly used. The mzXML2other will convert files from mzXML format to the DTA format, generating at least one DTA file for each MS/MS scan (two DTA files may be generated where the parent ion is multiple charged). As a result, the number of identifications can be used directly as the spectral count.

When using spectral counts, the significance of hypothesized abundance changes for proteins in different samples should be verified statistically. Old *et al*. [45] adapted a statistic originally devised for serial analysis of gene expression data [64] to take into account variances in the depth of analysis between different LC^n-MS/MS runs as shown below:

$$R_{SC} = \log_2\left[\frac{(n_2 + f)}{(n_1 + f)}\right] + \log_2\left[\frac{(t_1 - n_1 + f)}{(t_2 - n_2 + f)}\right] \quad (1)$$

where, R_{SC} is the $\log_2$ ratio of abundance between samples 1 and 2, n is spectral count, t is the total number of spectra and f is a correction factor predetermined to be optimal at 1.25. The advantage of using R_{SC} is that it avoids the problem of discontinuous spectral count values necessary for statistical tests such as the Student's *t*-test.

Zhang *et al*. [59] performed a comparison of five different statistical tests for evaluating the significance of comparative quantification by spectral counts. The goodness-of-fit test (G-test) [65], Fisher's exact test [66] and AC test [67] were performed on nonreplicated spectral count data when the data is expressed as a two-way table (Table 2). The Student's *t*-test and Local-Pooled-Error (LPE) test [68] can only be performed on spectral count experiments with replicates such that a mean value is available

Table 2: Display of spectral count data in a two-way table for statistical testing

	Sample 1	Sample 2	Total counts
Protein *x*	x_1	x_2	x
Other proteins	n_1	n_2	n
Total	t_1	t_2	t

for testing. Since the acquisition of MS/MS data typically employ data-dependent programming with dynamic exclusion of detected peptides for a period of time, the assumption of random sampling in statistical tests are generally violated. Nevertheless, Zhang *et al*. [59] found that the sampling of peptides for MS/MS analysis is sufficiently random for the tests to be applicable. They conclude that the Student's *t*-test was best when three or more replicates are available, while the G-test, Fisher's exact test and AC test are all-applicable when no replicates are available. However, out of the latter three tests, the G-test is the best choice due to its computational simplicity and the ability for the test to be generalized for comparing sample abundances that belong to more than two classes.

Most recently, a method for absolute quantification of protein levels based on spectral counting has been proposed [5]. The basis of the method is to form a relationship between the number of observed spectral counts for a particular protein and the expected number of spectral counts for all observed proteins within a sample. The absolute protein expression index (*APEX*) for protein *i*, is expressed as:

$$APEX_i = \frac{n_i p_i}{\#\text{observed}} \times \frac{C}{O_i \sum_{k=1}^{\text{proteins}} \frac{n_k p_k}{O_k}} \quad (2)$$

where, n is the spectral counts, p is the probability that the protein is correctly identified (typically computed by ProteinProphet [61]), O is the expected number of unique peptides that will be observed and C is the estimated total concentration of protein molecules in the sample. The most challenging part of the formula is to estimate the expected number of unique peptides that will be observed for each protein. Machine learning approaches have been applied to predict the likelihood that a peptide will be detected on a variety of instruments [5, 6]. Using these predictors, Lu *et al*. [5] are able to show that the estimated protein concentrations correlates well with the actual values and will generally fall within the correct order of magnitude. The major advantage of absolute quantification is that it will enable samples acquired from different instruments to be compared directly.

DISCUSSION

Old *et al*. [45] directly compared the use of peptide ion intensity against peptide spectral count. They reported that both methods were able to distinguish protein abundance changes of approximately 2.5-fold. Spectral counting was found to have a greater effective dynamic range, meaning that a larger of number of peptides detected showed statistically significant changes when compared using spectral counts. Yet, low spectral counts between 0 and 4 may overestimate protein ratios.

Wienkoop *et al*. [22] applied both quantitative methods in their comparative proteomics study and found that the results from the two methods are generally in good agreement. The only difference in the two methods is that for proteins of very low abundance, where peptide spectral count was 0, a signal for peak integration could still be obtained giving statistically more accurate results. In a further comparative study of the two methods, Xia *et al*. [69] found that spectral counting had the greatest precision when the observed protein ratios are correlated with the true ratios. They also noted that pooling individual peptide spectral counts or intensities into their respective proteins prior to comparative quantification provided greater sensitivity than if each peptide was compared directly. While these studies suggest that spectral counting is advantageous, both methods have been successfully applied in large quantitative studies requiring label-free quantification [21–28].

In terms of implementation, the main disadvantage of intensity-based quantification is that the computational process of extracting peptide quantitative information is significantly more complicated compared to spectral counting. Processes such as chromatographic alignment, smoothing peak integration may be sub-optimal and without manual verification can even introduce spectral preprocessing artifacts into quantitative data. Furthermore, without the use of stable isotope labeling, there is currently no established method to estimate absolution protein levels using intensity-based data.

The most significant drawback of spectral counting compared to intensity-based quantification is that the former is more likely to be influenced by the acquisition program of the mass spectrometer. High abundance proteins can mask low abundance proteins if the data dependent MS/MS acquisition exclusion list is not large enough. On the other hand, if the exclusion list is too large, the spectral count can become rapidly saturated, resulting in reduced sensitivity. The optimal setting is likely to vary between different mass spectrometers and LC setups

meaning that optimization may need to be performed on an individual basis.

CONCLUSION

The development of computational and statistical methods and advances in LC^n-MS/MS systems has meant that label-free quantitative proteomics is now being widely adopted. While, stable-isotope methods may have better dynamic range and sensitivity for low-abundance peptides, the convenience of label-free quantification will enable it to be performed without special experimental consideration.

The use of peptide ion intensity and spectral counting are two distinct approaches that enable comparative quantification of label-free LC^n-MS/MS data. Studies have shown that both methods are complementary and in recent years, both have been successfully applied on real-life proteomics samples. With the increasing popularity of the use of proteomics pipelines for analysis of LC^n-MS/MS data and with a wide variety of publicly available software for acquiring peptide ion intensities and spectral counts, the simultaneous use of both methods could be implemented for greater confidence.

Key Points

- Label-free comparative strategies have significantly increased the accessibility of MS-based quantitative proteomics.
- Peptide ion intensity and spectral counts are the two methods used to infer protein/peptide quantitative information. Both methods have been widely adopted in recent years.
- Spectral counting represents the simplest method for comparative quantitative proteomic analysis. Evidence suggests that it outperforms ion intensity in terms of dynamic range.

Acknowledgement

The funding for this article was provided by Science Foundation of Ireland (02/IN.1/B117 to G.C. and M.J.S.); Irish Research Council for Science, Engineering and Technology (postdoctoral fellowship to J.W.H.W).

References

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;**422**:198–207.
2. Link AJ, Eng J, Schieltz DM, *et al*. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* 1999;**17**: 676–82.
3. Washburn MP, Wolters D, Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 2001;**19**:242–7.
4. Wolters DA, Washburn MP, Yates JR, 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* 2001;**73**:5683–90.
5. Lu P, Vogel C, Wang R, *et al*. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 2007;**25**: 117–24.
6. Mallick P, Schirle M, Chen S, *et al*. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* 2007;**25**:125–31.
7. Tang H, Arnold RJ, Alves P, *et al*. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* 2006:e481–8.
8. Annesley TM. Ion suppression in mass spectrometry. *Clin Chem* 2003;**49**:1041–4.
9. Gygi SP, Rist B, Gerber SA, *et al*. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999;**17**:994–9.
10. Ong SE, Blagoev B, Kratchmarova I, *et al*. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 2002;**1**:376–86.
11. Ong SE, Mittler G, Mann M. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nat Meth* 2004;**1**:119–26.
12. Ong SE, Kratchmarova I, Mann M. Properties of ^{13}C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *J Proteome Res* 2003;**2**: 173–81.
13. Cagney G, Emili A. De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat Biotechnol* 2002;**20**: 163–70.
14. Ross PL, Huang Y, Marchese J, *et al*. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 2004; **3**:1154–69.
15. Mirgorodskaya O, Kozmin Y, Titov M, *et al*. Quantitation of peptides and proteins by matrix-assisted laser desorption/ ionization mass spectrometry using (18)O-labeled internal standards. *Rapid Commun Mass Sp* 2000;**14**:1226–32.
16. Bondarenko P, Chelius D, Shaler T. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem* 2002;**74**:4741–9.
17. Chelius D, Bondarenko P. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* 2002;**1**:317–23.
18. Chelius D, Zhang T, Wang G, *et al*. Global protein identification and quantification technology using two-dimensional liquid chromatography nanospray mass spectrometry. *Anal Chem* 2003;**75**:6658–65.
19. Wang W, Zhou H, Lin H, *et al*. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem* 2003;**75**: 4818–26.
20. Zybailov B, Coleman MK, Florens L, *et al*. Correlation of relative abundance ratios derived from peptide ion

chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal Chem* 2005;**77**:6218–24.

21. Gravett MG, Thomas A, Schneider KA, *et al*. Proteomic analysis of cervical-vaginal fluid: identification of novel biomarkers for detection of intra-amniotic infection. *J Proteome Res* 2007;**6**:89–96.
22. Wienkoop S, Larrainzar E, Niemann M, *et al*. Stable isotope-free quantitative shotgun proteomics combined with sample pattern recognition for rapid diagnositics. *J Sep Sci* 2006;**29**:2793–801.
23. Wang HX, Qian WJ, Chin MH, *et al*. Characterization of the mouse brain proteome using global proteomic analysis complemented with cysteinyl-peptide enrichment. *J Proteome Res* 2006;**5**:361–9.
24. Ruth MC, Old WM, Emrick MA, *et al*. Analysis of membrane proteins from human chronic myelogenous leukemia cells: comparison of extraction methods for multidimensional LC-MS/MS. *J Proteome Res* 2006;**5**: 709–19.
25. Le Bihan T, Goh T, Stewart II, *et al*. Differential analysis of membrane proteins in mouse fore- and hindbrain using a label-free approach. *J Proteome Res* 2006; **5**:2701–10.
26. Kislinger T, Cox B, Kannan A, *et al*. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* 2006;**125**: 173–86.
27. Fang RH, Elias DA, Monroe ME, *et al*. Differential label-free quantitative proteomic analysis of *Shewanella oneidensis* cultured under aerobic and suboxic conditions by accurate mass and time tag approach. *Mol Cell Proteomics* 2006;**5**:714–25.
28. Cao R, Li XW, Liu Z, *et al*. Integration of a two-phase partition method into proteomics research on rat liver plasma membrane proteins. *J Proteome Res* 2006;**5**: 634–42.
29. Li XJ, Yi EC, Kemp CJ, *et al*. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol Cell Proteomics* 2005;**4**:1328–40.
30. Gao J, Opiteck GJ, Friedrichs MS, *et al*. Changes in the protein expression of yeast as a function of carbon source. *J Proteome Res* 2003;**2**:643–9.
31. Liu HB, Sadygov RG, Yates JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 2004;**76**: 4193–201.
32. Pasa-Tolic L, Masselon C, Barry RC, *et al*. Proteomic analyses using an accurate mass and time tag strategy. *Biotechniques* 2004;**37**:621–39.
33. Görg A, Obermaier C, Boguth G, *et al*. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* 2000;**21**:1037–53.
34. Berger S, Lee S-W, Anderson G, *et al*. High-throughput global peptide proteomic analysis by combining stable isotope amino acid labeling and data-dependent multiplexed-MS/MS. *Anal Chem* 2002;**74**:4994–5000.
35. Palmblad M, Ramstrom M, Markides K, *et al*. Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Anal Chem* 2002;**74**:5826–30.
36. Palagi P, Walther D, Quadroni M, *et al*. MSight: an image analysis software for liquid chromatography-mass spectrometry. *Proteomics* 2005;**5**:2381–4.
37. Appel RD, Hochstrasser D, Funk M, *et al*. The MELANIE project: from a biopsy to automatic protein map interpretation by computer. *Electrophoresis* 1991;**12**: 722–35.
38. Smith C, Want E, Tong G, *et al*. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* 2006;**78**:779–87.
39. Bellew M, Coram M, Fitzgibbon M, *et al*. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* 2006;**22**:1902–9.
40. Kohlbacher O, Reinert K, Gröpl C, *et al*. TOPP – the openMS proteomics pipeline. *Bioinformatics* 2007;**23**: e191–7.
41. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectra data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;**5**:976.
42. Perkins DN, Pappin DJ, Creasy DM, *et al*. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;**20**:3551–67.
43. Tanner S, Shu HJ, Frank A, *et al*. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* 2005;**77**:4626–39.
44. MSQuant (2004). http://msquant.sourceforge.net/.
45. Old W, Meyer-Arendt K, Aveline-Wolf L, *et al*. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* 2005;**4**: 1487–502.
46. Li XJ, Zhang H, Ranish JA, *et al*. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal Chem* 2003;**75**:6648–57.
47. Han D, Eng J, Zhou H, *et al*. Quantitative profiling of differentiation-induced microsomal proteins using isotope- coded affinity tags and mass spectrometry. *Nat Biotechnol* 2001;**19**:946–51.
48. MacCoss MJ, Wu CC, Liu HB, *et al*. A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem* 2003;**75**:6912–21.
49. Cutler P. Protein arrays: the current state-of-the-art. *Proteomics* 2003;**3**:3–18.
50. Katajamaa M, Miettinen J, Orešič M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 2006;**22**: 634–6.
51. Wong JWH, Durante C, Cartwright HM. Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal Chem* 2005;**77**:5655–61.
52. Zhang X, Asara J, Adamec J, *et al*. Data pre-processing in liquid chromatography–mass spectrometry-based proteomics. *Bioinformatics* 2005;**21**:4054–9.
53. Marshall AG, Guan S. Advantages of high magnetic field for fourier transform ion cyclotron resonance mass spectrometry. *Rapid Commun Mass Sp* 1996;**10**:1819–23.

54. Hu Q, Noll R, Li H, *et al*. The Orbitrap: a new mass spectrometer. *J Mass Spectrom* 2005;**40**:430–43.
55. Student. On the probable error of the mean. *Biometrika* 1908;**6**:1–25.
56. Mann H, Whitney D. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947;**18**:50–60.
57. Sajda P. Machine learning for detection and diagnosis of disease. *Annu Rev Biomed Eng* 2006;**8**:537–65.
58. Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* 2003:1484–91.
59. Zhang B, VerBerkmoes N, Langston M, *et al*. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Prot Res* 2006;**5**:2909–18.
60. Keller A, Nesvizhskii AI, Kolker E, *et al*. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002;**74**: 5383–92.
61. Nesvizhskii AI, Keller A, Kolker E, *et al*. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 2003;**75**:4646–58.
62. Tabb DL, MacCoss MJ, Wu CC, *et al*. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem* 2003;**75**:2470–7.
63. Sashimi Project. http://sashimi.sourceforge.net.
64. Beissbarth T, Hyde L, Smyth G, *et al*. Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics* 2004;**20**: I31–9.
65. Sokal R, Rohlf F. *Biometry: the Principles and Practice of Statistics in Biological Research*. New York: Freeman, 1994.
66. Fisher R. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J Roy Stat Soc* 1922;**85**: 87–94.
67. Audic S, Claverie J. The significance of digital gene expression profiles. *Genome Res* 1997;**7**:986–95.
68. Jain N, Thatte J, Braciale T, *et al*. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* 2003;**19**: 1945–51.
69. Xia QW, Wang TS, Park Y, *et al*. Differential quantitative proteomics of Porphyromonas gingivalis by linear ion trap mass spectrometry: non-label methods comparison, q-values and LOWESS curve fitting. *Int J Mass Spectrom* 2007;**259**:105–16.