3

Precision Quantitative Proteomics with Fourier-Transform Mass Spectrometry

Qingbo Li
University of Illinois at Chicago
U.S.A.

1. Introduction

The Fourier-transform mass spectrometry (FTMS) instrument offers a mass resolution higher than most of other mass spectrometers. This high resolution is in part due to the better stability offered by a superconductor magnet in FTMS than a radio frequency voltage utilized in many other mass spectrometers. The extremely high resolution of FTMS has very important application in biomedical proteomics research. The high resolution not only dramatically improves the reliability of protein identification but also the accuracy of protein quantitation.

In this chapter, we present several examples of proteomics study that takes advantage of the high resolution offered by FTMS. Particularly, we describe examples of proteome dynamic study with isotopomer analysis, and precise peptide and protein label-free quantitation with rigorous statistical assessments.

In protein dynamic studies, FTMS readily resolves all of the isotopomer peaks for peptides. The well-resolved isotopomer peaks allow a direct integration of the intensities of different isotopologue envelopes without the need of a deconvolution algorithm. With the high-resolution data, we were able to show that protein turnover measurement revealed more subtle changes in the dynamics of a proteome.

The high resolution of FTMS helps to reduce the interference of contaminant peaks (**Fig. 1**). The ability to resolve the targeted peptide isotopomer peaks from interfering ones greatly facilitates the implementation of a label-free quantitative proteomics method that relies on peptide cross reference between liquid chromatography runs and the integration of extracted ion chromatographic intensities (Lipton et al., 2002). With both high confidence in peptide identification and quantitation, we showed that the major source of variability lies more in sample preparation than in liquid chromatography/mass spectrometry analysis. Such a result has a direct consequence in the statistical approaches utilized to assign significance in label-free quantitative proteomics.

Two sections are presented in this chapter. The first one briefly discusses the general aspect of a protein turnover analysis followed by examples of proteome dynamics study in acid stressed and iron limited mycobacterial cells. The second one describes the utilization of high-resolution FTMS data for label-free quantitation of proteins with a rigorous statistical assessment of significance in differential protein abundances.

The spectra shown in the panels a to e of Fig. 1 are for a tryptic peptide from a superoxide dismutase in *M. smegmatis* (MSMEG_6427; Mn) with a sequence of AFWNVVNWDDVQNR.

The panels a and b are the ion-trap mass spectrometry (ITMS) MS^2-scan spectra for a +2 precursor ion of the peptide in the nanoLC/LTQ-FT and the nanoLC/LTQ respectively. Shown atop of panel a is the peptide sequence labeled with the detected y-series ions in the MS^2-scan spectra. For clarity, only y-series ion fragments are shown with labels. The panels c to e show the MS-scan spectra obtained with a FTMS scan (c), an ITMS zoom MS-scan (d), and an ITMS full MS-scan (e), respectively. Only an m/z range of 880 to 885 is shown for the +2 peptide charge state. The three short arrows in the panels c, d, and e indicate the first three isotopic peaks of the +2 peptide charge state. The asterisks in panel c indicate the isotopic peaks probably not related to the peptide. The theoretical molecular weight of the peptide is shown atop of panel c. The long dashed arrows point to the respective MS-scans from which a precursor ion is selected for the MS^2 scans. Adapted from (Li, 2010a).

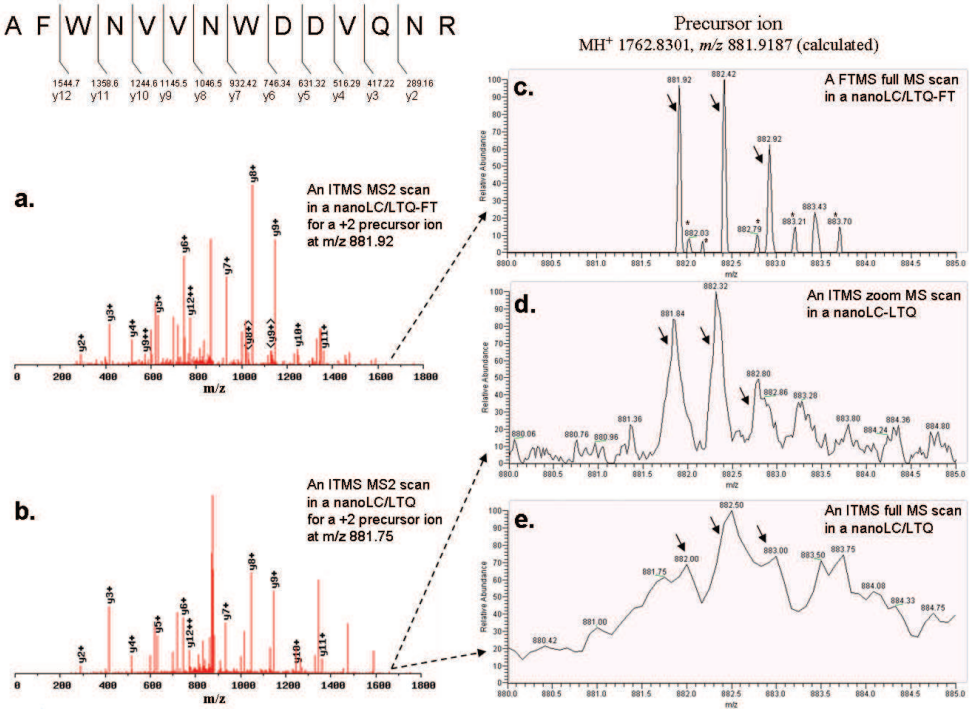

Fig. 1. A resolution and a signal-to-noise ratio in MS-scan and MS^2-scan spectra are compared between a nanoLC/LTQ-FT and a nanoLC/LTQ mass spectrometry systems

2. Protein turnover analysis with FTMS

2.1 Protein turnover

Protein turnover is a fundamental cellular process in all cell types having important implications in many aspects of biological science (Larrabee et al., 1980; Wilkinson, 2005). Advancement of high resolution proteomic technologies has provided the possibility to study protein turnover for multiple proteins simultaneously in complex cellular protein extracts (Beynon, 2005; Cargile et al., 2004; Pratt et al., 2002; Rao et al., 2008a; Rao et al.,

2008b; Vogt et al., 2005). We showed that a combination of protein abundance and turnover data provides a highly interesting insight into the dynamic process of and interconnection among protein synthesis, degradation, and secretion (Rao et al., 2008a).

Protein turnover shares an equally important role with gene transcription and protein translation. Synthesis of new proteins and degradation of old ones form a dynamic process in an organism. Turnover does not only help to clear old proteins but also aid in a fast adaptation to a new condition or environment by adjusting the rate of protein synthesis and degradation (Goldberg & Dice, 1974). Apart from this, turnover also brings new proteins into action with reduced strain on the resources of an organism because preexisting cellular materials are reused. Some early global protein turnover studies identified different *E. coli* proteins that might have different turnover rates. One of those earlier studies showed that a dynamic state for individual proteins existed in non-growing as well as growing cells (Larrabee et al., 1980). Studies of turnover offer a dynamic view of the abundances of proteins. When being applied to a larger scale of the proteome, protein turnover analysis allows one to study the dynamic nature of the entire proteome (Li, 2010a).

Mass spectrometry continues to serve as a major approach for a protein turnover study due to its wide availability and flexibility to analyze both single-cell cultures and multi-cellular organisms. Except for the required administration of stable isotope-labeled metabolites, amino acids, or water to the study subjects, mass spectrometry-based approaches do not require any genetic manipulation of the study subjects. The avoidance of genetic manipulation of the study subjects helps to minimize any unwanted perturbation to a biological system and delivers the most physiologically relevant results. A range of methodologies has been established to measure protein turnover based on stable isotope labeling and mass spectrometry (Doherty & Beynon, 2006).

2.2 High-resolution mass spectrometry for protein turnover analysis

The advent of highly automated high-resolution mass spectrometry technology promises to bring about in-depth insight into the dynamic nature of a proteome at the global level. The work done by Pratt et al. (Pratt et al., 2002) demonstrated the determination of protein degradation rate constants in a steady state population of yeast grown in a chemostat. The authors used isotope labeling along with 2D gel analysis to study protein turnover and advocated that protein turnover is 'a missing dimension in proteomics.' Another study done by Cargile et al. (Cargile et al., 2004) labeled *E. coli* cells with ^{13}C to study the relative synthesis over degradation ratio (S/D). These earlier works demonstrated the global analysis of protein turnover with individual protein identifications but their data were not correlated with abundance values.

With one-dimensional SDS/PAGE fractionation and subsequent nanoLC/LTQ-FTMS analysis, Rao et al. determined the global turnover profiles of *Mycobacterium smegmatis*, a non-pathogenic surrogate of *Mycobacterium tuberculosis*, under acid-shock and iron-limitation conditions (Rao et al., 2008b). A dynamic range of 3-orders of magnitude was demonstrated for relative turnover measurements. The study provided direct evidence that relative turnover in growing mycobacterium cells, with or without stress, was highly heterogeneous. The results obtained in that study addressed the long-standing question whether a 'dynamic state' exists in growing bacteria (Borek et al., 1958), and illustrated the benefits and needs to study protein turnover at the global level with the most advanced mass spectrometry technology.

In the work by Rao et al. (Rao et al., 2008b), the cells were grown with two different methodologies for both different types of stresses. For the pH stress the cells were initially grown in [^{14}N]-containing media at pH 7.0. Once the cells are in the initial log phase, the cells were divided into two flasks and the media was doped with 50% [^{15}N] and the pH was reduced to 5.0 in one of the flasks. The cells were harvested after one doubling and analyzed for protein turnover using LC/LTQ-FTMS. For low iron analysis, the cells were first grown in [^{15}N]-containing media until mid-log phase. The cells were then collected by centrifugation and the media was then exchanged with [^{14}N]-containing media. The cells were then allowed to grow to one doubling and harvested to be analyzed by LC/LTQ-FTMS.

In either the complete isotope swapping (iron-limitation experiments) or the partial isotope labeling conditions, the isotopologue envelopes and the individual isotopomer peaks of a peptide are clearly resolved. The complete resolution of the isotopomer peaks and the isotopologue envelops for the old proteins and the de novo synthesized proteins facilitates simple calculation of the abundances of the new and old fractions of a peptide (**Fig. 2**).

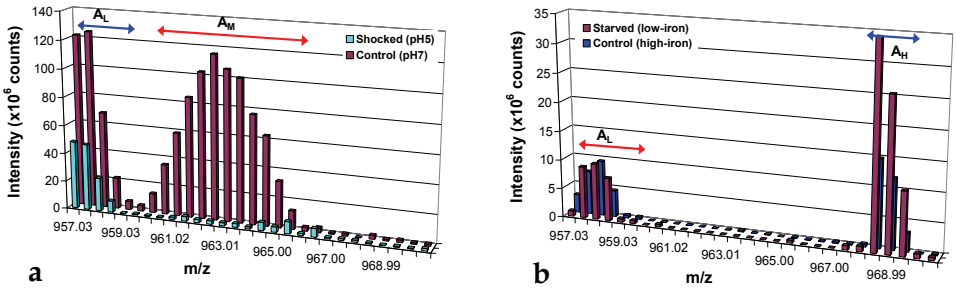

Fig. 2. Calculation of isotopologue intensities for a representative tryptic peptide ANLLGLSAPEMTTLVGGLR (MH_2^{2+}, 23 N atoms) of protein KatG (MSMEG6346) in shocked (pH5) and control (pH7) cultures (panel a), and in starved (low-iron) and control (high-iron) cultures (panel b). A_L, A_H, and A_M represent the isotopologue intensities with light label (99.6At%[^{14}N]), heavy label (99At%[^{15}N]) and medium label (50At%[^{15}N]) respectively. Red arrows and text labels indicate *de novo* synthesized proteins. Blue arrows and text labels indicate old proteins. Adapted from (Rao et al., 2008b) with permission

Turnover analysis of *M. smegmatis* under both stressful conditions revealed two different patterns (Rao et al., 2008b). In the low pH condition, many proteins had increased turnover at pH 5.0 as compared to pH 7.0. It was an obvious reaction since the bacteria has to readjust its proteome in order to counter the stress posed by increased proton concentrations. The correlation coefficient for the low pH shock cells was small which indicated that the proteins in the cells exposed to pH 5.0 underwent extensive readjustment in different directions. In the low iron stress the correlation coefficient being high suggested that either there was not much rearrangement of turnover values or all the proteins had changes in a similar direction. KatG and Tpx, which are important for protection of mycobacterial cells against oxidative stress, had low protein turnover values in both low iron as well as low pH conditions. A study on *M. tuberculosis* Tpx suggested that it might be an important protein

against oxidative stress because Tpx mutants were unable to survive in the macrophages in an infected mouse model. However, it would be interesting to analyze how the low turnover of Tpx correlates with the survival of mycobacteria in the cell.

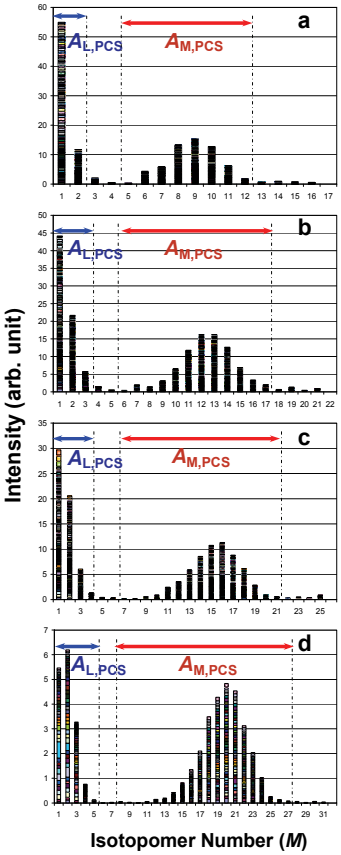

Fig. 3. Average isotopomer profiles and selected isotopomer ranges to calculate the abundances of peptides in *M. tuberculosis*. $A_{L,PCS}$ and $A_{M,PCS}$ are peptide abundance for old and new proteins respectively. Each panel was the stacked column graph of the normalized isotopomer profiles of the detected peptide charge states (PCSs) having the same number of N atoms (n). Profiles are shown for n equal to 11 (a), 16 (b), 20 (c), and 26 (d) respectively. The blue and red arrows indicate the M ranges for calculating $A_{L,PCS}$ and $A_{M,PCS}$ respectively

An open question is how the protein turnover values correlate with protein abundances. To investigate the correlation between protein abundance and protein turnover values in *M. tuberculosis*, Rao et al. analyzed *M. tuberculosis* cells in an iron replete and iron depleted condition using the high resolution LC/LTQ-FTMS instrument (Rao et al., 2008a). The approach employed many large-scale quantitative proteomics techniques to make it readily accessible for protein turnover studies at the global level. The concomitant measurement of protein turnover and abundance was previously shown by Gerner et al. with 2D gel

electrophoresis for separation and fluorography and autoradiography for quantitation of individual gel spots without any protein quantification (Gerner et al., 2002). With the advanced nanoLC/LTQ-FTMS system and a label-free quantitation approach, Rao et al. demonstrated that protein abundance and turnover could both be measured with a dynamic range of at least 3-orders of magnitude in an automated fashion (Rao et al., 2008a).

The study compared the sensitivity of relative turnover and relative abundance measurements to detect a dynamic response of *M. tuberculosis* when the cell culture was shifted from an iron-starved to iron-sufficient condition. An unlabeled iron-depleted *M. tuberculosis* culture was grown to late-log phase and diluted with a fresh iron-replete medium that was labeled with [^{15}N]-labeled nitrogen source (Rao et al., 2008a). The incorporation of the [^{15}N]-labeled nitrogen source into the newly synthesized proteins resulted in a complete separation between the old and the newly synthesized peptide isotopologue profiles (**Fig. 3**). Similar to that shown in Fig. 2 for *M. smegmatis*, the complete separation of the isotopologue profiles and the isotopomers allows the quantitation of the abundances of old proteins, newly synthesized proteins, and the total proteins.

In this work, we are able to obtain both the protein abundance and turnover values to more comprehensively assess the dynamic response of the H37Rv cells when they were shifted from iron-starved stationary-phase to fresh low- and high-iron media. This is achieved by applying both the isotope chasing and a label-free quantitation method (Rao et al., 2008a).

The results indicated that a relative turnover measurement was much more sensitive to monitor the dynamic response of the *M. tuberculosis* cells. Meanwhile, a combination of turnover and abundance measurements provided insight into the correlation of protein synthesis, degradation, and secretion. A further principal component analysis of the *M. tuberculosis* proteome dynamics reveals that protein relative turnover properties are orthogonal to protein relative abundance properties (Rao & Li, 2009a). Thus, a study of protein turnover at the global level would likely bring forward new findings that can be missed with a protein abundance analysis alone.

The data obtained from the comparison of protein abundance and protein turnover values showed that protein turnover is a much more sensitive measurement to discern the changes in the proteome than abundance measurements. Upon the transfer of late-log phase cells from a low iron to a high iron media, protein abundance measurements showed that out of the 104 proteins that we identified, only 5 proteins were upregulated and 16 proteins were downregulated in the HI media. Relative abundance of KatG was upregulated in cells grown in the high iron media.

Protein turnover analysis of the proteins compared between cells grown in a low iron and a high iron media showed that more proteins had increased synthetic activity in the high iron grown cells. The S/D had increased for 24 proteins in the cells grown in the low iron media. Eight proteins had decreased turnover. However, for cells grown in the high iron media, 56 proteins had increased S/D and 5 proteins had decreased S/D. A comparison of protein abundance measurements to protein turnover measurements clearly suggests that protein turnover does give more information to uncover the dynamic response of the proteome (**Fig. 4**).

In addition to providing the information about synthesis and degradation of proteins, the protein turnover analysis can also provide information regarding whether a protein has been secreted when the protein turnover values are analyzed together with the protein abundance measurements. In our study of proteome dynamics, we found that some proteins had low changes in relative abundances even though their synthesis had increased

significantly. As stated before the relative abundance of a protein in the cell can be affected not only by synthesis or degradation but also by a secretion process. In our turnover analysis, we found discrepancies between protein abundance values of certain proteins and their turnover values. A previous proteomic study of *M. tuberculosis* culture filtrates showed many of those proteins to be secreted into the culture filtrate. Proteins such as FbpC2, KatG, and the mammalian cell entrance protein Rv0172 were also predicted to be secreted (Malen et al., 2007).

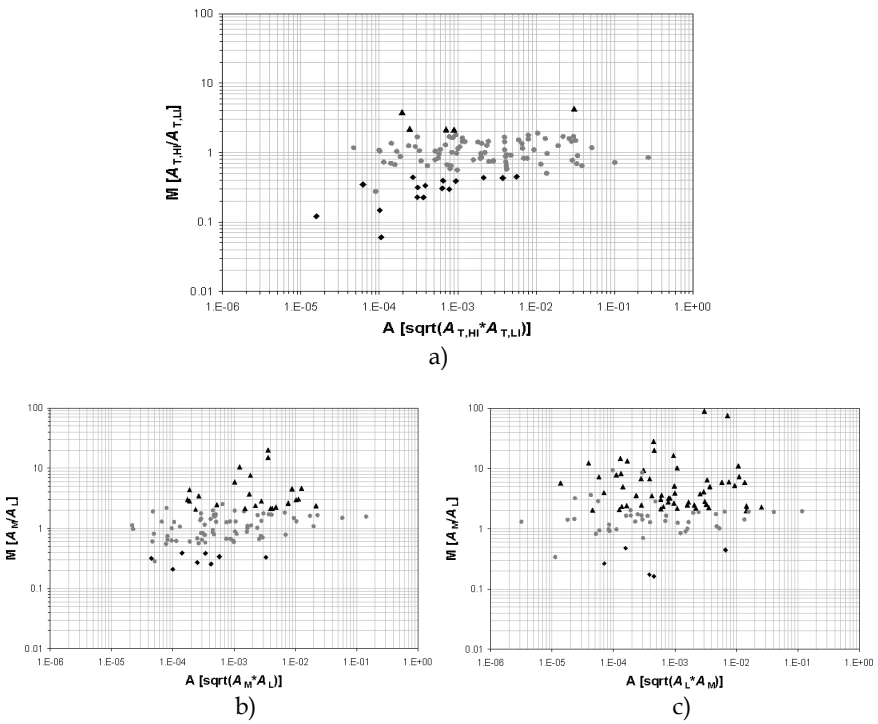

Fig. 4. M-A plots representing the total protein abundances in high iron (HI) versus low-iron (LI) cells (panel a), the newly synthesized protein abundances versus the old protein abundances in LI cells (panel b) and in HI cells (panel c) respectively. The proteins with 2-fold significant (p <.05) change in relative abundance are marked with black triangles and diamonds. The M–axis represents the relative abundance values and the A-axis represents the average abundance values. Adapted from (Rao et al., 2008a) with permission

These results support that protein turnover in combination with abundance analysis could predict the secretion of proteins and reveal the interconnected roles of protein synthesis, degradation, and secretion in determining the protein abundances in cells. These analyses illustrate that protein turnover can divulge information that classical proteomics does not provide. The integration of data from transcriptome studies, abundance measurements and turnover analyses will likely provide a more complete picture of the dynamics associated with the proteome. To some extent, it will probably reconcile the discordances between transcriptome and proteome analyses.

2.3 Implication of proteome turnover studies in mycobacteria

With the advent of high-precision and automated mass spectrometry instrumentation to support large-scale proteomic studies, protein turnover analysis at the global level potentially has an increasing importance for biomedical research. One example is the application of protein turnover analysis to study *M. tuberculosis*, especially at its non-replicating or dormant state. Whereas over a hundred research articles have been published on mycobacterial proteomes, only a few dealt with non-replicating *M. tuberculosis* (Cho et al., 2006; Rosenkrands et al., 2002). Information about protein turnover in non-replicating and dormant *M. tuberculosis* is scarce in the literature. With the potential importance of proteome dynamics in bacterial cell sporulation or dormancy (Bernlohr, 1967; Bernlohr, 1972; Mandelstam, 1958; Spudich & Kornberg, 1968), a study of mycobacterial protein turnover at the global level (Rao et al., 2008a; Rao et al., 2008b) will likely help to advance our understanding of the molecular basis of *M. tuberculosis* persistence (Rao & Li, 2009b).

Over the last century, a variety of control and eradication measures have been implemented against tuberculosis such as vaccination, aggressive chemotherapy, and public health surveillance. But tuberculosis still remains a major global health problem to continue to cause nearly 2 million deaths and 9 million new infection cases per year. *Ca.* one third of the world population is infected with *M. tuberculosis*. A majority of the infected individuals remain asymptomatic whereas they carry a lifetime risk to develop an active disease i.e., the latent tuberculosis infection. The state of latency represents the greatest obstacle to eradicate the tuberculosis disease.

The metabolic requirement of *M. tuberculosis* in latency is unclear and difficult to study because the bacilli presumably remain dormant in a granuloma (Pagan-Ramos et al., 2006). The anti-tuberculosis drugs used today have their maximum effect against the growing but not the dormant bacilli. The long therapeutic regime required to treat latent tuberculosis infection is probably explained by the lack of a direct target that is specific to dormant *M. tuberculosis*. There are new drug treatment regimes and several new anti-tuberculosis drugs in a development pipeline that aim to shorten the treatment period and to overcome multi-drug resistant strains (Murphy & Brown, 2008). Most of these new drugs are still based on existing classes of antimicrobial compounds to target the conventional pathways and molecular machinery that are critical for the growth of *M. tuberculosis*. These drugs could still be countered by drug resistant strains that emerge from the non-adherence of a prolonged regime against latent tuberculosis infection. Thus, the need to discover novel drug targets, especially those against dormant bacteria, is urgent.

A 'simple but nonetheless vexing problem' in target discovery against non-replicating *M. tuberculosis* is that many methods rely on a growth-inhibition measurement to assess the effect of drug treatment (Murphy & Brown, 2008). Rao et al. showed that a protein-turnover measurement was much more sensitive than a protein relative abundance measurement alone to uncover protein synthesis activities in *M. tuberculosis* (Rao et al., 2008a); those data suggest that protein dynamics analysis with turnover and abundance measurements could potentially add a valuable alternative to the drug target discovery problem for non-replicating *M. tuberculosis*. The sensitive protein turnover analysis could also be useful to detect and validate drug treatment effects at an early phase, during which the most relevant drug effect can be isolated from other non-specific cell stress responses.

3. Label-free quantitation of proteins

There is an increased use of LC/MS instrumentation for proteomics studies at a large scale. The depth of a proteomic analysis, i.e. the number of protein species that can be precisely identified and characterized in an experiment, depends on the precision and sensitivity of a mass spectrometry instrument. The impact of the precision and sensitivity of an LC/MS instrument on a proteomic study is manifested in many areas of proteomic studies. For example, in the analysis of intracellular bacterial proteomes, FTMS-based approach clearly identified more proteins from scarce and complex intracellular bacterial samples compared to many other proteomic methods.

The precise retrieval of biological information from a large LC/MS dataset critically depends on algorithms for data interpretation, which remains a current bottleneck in the rapid advance of proteomics technology (Mortensen et al., 2010). The quantitation of differentially regulated proteins represents a major type of proteomics application in biological studies. Protein quantitation with LC/MS data includes three conceptually different methods i.e., spectral counting, differential stable isotope labelling, and label-free LC/MS measurements by using extracted ion chromatographic intensities (Mueller et al., 2008). Due to the increased time and complexity of sample preparation in stable isotope labelling, cost of labelling reagents, and requirement of higher starting sample amount, however, researchers are increasingly using label-free proteomics for faster and simpler protein quantitation (Zhu et al., 2010).

Most of proteomics studies infer proteins with ≥2 identified peptides as reliable protein identifications and usually disregard proteins with a single-peptide hit as unreliable for quantitation. This "two-peptide" rule was recently challenged with the evidence that it reduced protein identifications more in a target database than in a decoy database and thus increased false discovery rates in protein identification (Gupta & Pevzner, 2009). Indeed, it was shown that proteins with a single-peptide hit could represent 30% of the proteins identified with ≥2 MS^2 spectrum matches at p <.01 (Li & Roxas, 2009). Because those single-peptide proteins had ≥2 MS^2 spectrum matches (p <.01) in multiple LC/MS analyses under the same condition, they had an adequate level of statistical confidence to be included for quantitation.

But the inclusion of single-peptide proteins in a differential quantitative proteomics analysis raises two issues. The first is that a conventional statistical test such as a t-test can not be applied toward these single-peptide proteins when the t-test relies on multiple quantified peptides as replicates to calculate the t-statistic for the protein relative abundance (Li & Roxas, 2009). The second is that many single-peptide proteins are at a lower abundance and thus noisy. More stringent thresholds are needed to control the false discovery rate when these single-peptide proteins are included for the selection of differentially regulated proteins.

Pavelka et al. applied a power law global error model (PLGEM) and the signal-to-noise ratio (STN) statistic (Pavelka et al., 2004) to select differentially regulated proteins based on a spectral counting quantitation method (Pavelka et al., 2008). The PLGEM-STN statistic utilized a re-sampling approach to estimate the null distribution from replicates of a sample. After the error model was calculated from a pool of re-sampling statistics that constituted the null distribution, a set of STN thresholds was applied at a specified confidence level toward samples with any level of replicates. The PLGEM-STN method is attractive in that it could be applied toward samples with no replicates if several replicates for one sample are

provided to estimate the null distribution. It is also applicable to proteins with any number of identified peptides. The PLGEM-STN method, however, has not been demonstrated for label-free quantitation with extracted ion chromatographic intensities.

In the work described in the following, the PLGEM-STN statistic was applied toward a LC/MS dataset obtained with a high-resolution mass spectrometer (Roxas & Li, 2009). The peptide and protein abundances were quantified with a label-free approach based on extracted ion chromatographic intensities. The false discovery rate was estimated at different confidence levels of the PLGEM-STN statistics.

The PLGEM-STN statistic alone did not provide a desired level of false discovery rate control. Insufficient stringency in false discovery rate control was similar to the situation when a t-test statistic was used alone (Li & Roxas, 2009). With the combination of a t-test and the rule of minimum number of permuted significant pairings (MPSP), however, the false discovery rate was significantly reduced in that study.

The combination of MPSP and PLGEM-STN was further tested to control the false discovery rate. PLGEM-STN does not require that a protein have to have at least two detected peptides for an assessment of statistical significance. Thus, the combination of MPSP and PLGEM-STN has the potential to extend the selection of differentially regulated proteins to those with lower fold-changes and to those with single-peptide hits. Similarly, a fold-change threshold can be applied toward proteins with any number of detected peptides. With a control, the statistical significance of a differentially regulated protein can also be assessed based on a fold-change threshold. Therefore, the combination of MPSP and fold-change thresholds was also tested and compared with the PLGEM-STN-MPSP approach.

It is important to use the high-resolution FTMS instrument to acquire data for the label-free quantitation so that proteins with a single peptide hit can be quantified based on peptide cross reference and extracted ion chromatographic intensity. Proteins with fewer than three peptide hits are typically difficult to quantify by the spectral count method.

3.1 Purpose of the study

There are two purposes to investigate the combination of the PLGEM-STN statistic or a fold-change threshold with the MPSP method to identify differentially regulated proteins. One was to extend the selection of differentially regulated proteins to those that had single-peptide hits. The other was to select differentially regulated proteins at smaller fold-changes and at a false discovery rate ≤.05. The approaches to achieve this two-fold purpose were investigated under a scenario where the number of sample replicates was small. When the number of sample replicates is small, other typical statistics such as a t-test might not perform well to provide the necessary specificity in the label-free quantitation of differentially regulated proteins. Therefore, it was found necessary to insert an additional measure, such as the MPSP rule to compensate for the lack of sample replicates (Li & Roxas, 2009). Even when more sample replicates are available, the additional use of MPSP might still further increase the specificity although this possibility will need to be tested.

3.2 Results

With a null distribution built from the labelled control sample to establish thresholds, different approaches were experimented with to select differentially regulated proteins by using the combination of MPSP, PLGEM-STN, and fold-change methods. Differentially regulated proteins were selected from the unlabeled sample pair S_P and R_P.

The following two subsections of Results are described. Subsection 3.2.1 analyzes the source of variability in the peptide and protein quantitation processes. Subsection 3.2.2 performs multi-step extended selection of differentially regulated proteins.

3.2.1 Source of variability

An observed differential abundance of a PCS or protein between samples arose not only from the difference in biological samples but also from measurement noise that included the variability from multiple steps that involve LC/MS injection replicates, sample preparation replicates, biological replicates, or the data processing method. The multi-step experimental procedures are summarized in **Fig. 5**.

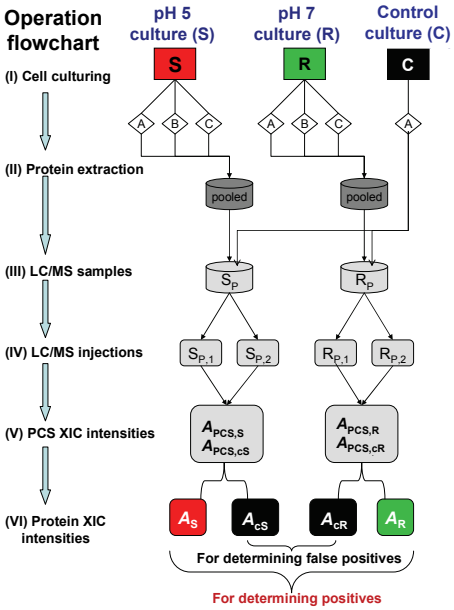

Fig. 5. Experimental outline of the label-free protein quantitation approach to assess the acid stress response between the unlabeled stressed culture (S) and the unlabeled reference culture (R) with the [^{15}N]-labeled culture as control (C)

The biological sample model used in the study was the proteome response of an acid stressed *M. smegmatis* culture (S) in reference to a neutral pH culture (R) (Roxas & Li, 2009) (Fig. 5). Both S and R cultures were unlabeled. The proteins from a [^{15}N]-labelled control culture (C) were used as an internal standard to mix with the proteins from the unlabeled cultures. Because the proteins from the control culture were analyzed repeatedly with two other unlabeled samples, the repeated analyses of the labelled control provided replicates to construct a null distribution.

In the null distribution, there were no true differentially regulated proteins. The null distribution was thus useful to model the noise in the experiment. The error model was derived from the null distribution that consists of at least two replicates of sample preparation. The error model was preferred not to derive from the unlabeled protein

samples S_P and R_P, because each of these two unlabeled samples had only LC/MS run replicates but no sample preparation replicates. The use of only LC/MS replicates to model the noise is likely to underestimate the noise level in the experiment.

The experimental procedures were divided into six stages (I-VI). Briefly, equal amounts of protein extract from the S culture triplicates were pooled. Equal amounts of protein extract from the R culture triplicates were also pooled. Into these two pooled unlabeled protein samples, an equal amount of protein extract from the C culture was added. This resulted in the two pooled samples i.e., S_P and R_P. The proteins differentially expressed between the S and R cultures were determined based on comparison of the abundances of the unlabeled proteins i.e., A_S and A_R, between samples S_P and R_P. For the purpose of false discovery rate assessment, the abundances of the [^{15}N]-labeled proteins i.e., A_{cS} and A_{cR}, were quantified and compared between S_P and R_P in the same way as between A_S and A_R. The proteins found differentially expressed between A_S and A_R were considered positives, because they reflected the difference between the S and R cultures. The proteins found differentially expressed between A_{cS} and A_{cR} in the labeled form were false positives, because difference was not expected from the identical C sample that was run concurrently with two unlabeled samples in separate runs.

To assist in the assessment of the source of variability in the label-free quantitation of the LC/MS data, another three samples were used in addition to S_P and R_P. The three additional samples were the biological replicates of the S culture sample, namely S_A, S_B, and S_C. S_P was generated by pooling S_A, S_B, and S_C.

The 3rd of the five fractions of an SDS/PAGE gel lane was processed for LC/MS analysis for the protein samples S_A, S_B, S_C, S_P, and R_P with duplicate injections for each sample (Li & Roxas, 2009). The five samples with two LC/MS injections per sample resulted in 10 LC/MS runs. These 10 LC/MS runs of the 3rd fraction allowed the quantitation of 349 proteins for the 3rd fraction (Li & Roxas, 2009). Because a protein was quantified in both the unlabeled form (for culture S or R) and the labelled form (for culture C), there were 20 quantitation categories for each protein (**Table 1**). Thus, these 349 proteins and the 20 quantitation categories formed a 349 x 20 matrix. The 349 x 20 matrix was examined by a clustering analysis (Eisen et al., 1998). The clustering analysis provides an overview of the correlation among the protein samples and LC/MS injections, thus reveals the major source of variability.

From the dendrogram of the 20 quantitation categories shown in **Fig. 6**, it could be seen that the distance between each pair of duplicate LC/MS injections was the shortest compared to those between any other sample pairings. The closest distance of the duplicate LC/MS injections for a sample indicated that the variability between LC/MS injections was the smallest, which also indicated that the label-free data analysis methodology (Li & Roxas, 2009) did not introduce a more significant variability.

	Unlabeled protein samples from culture S or R					[^{15}N]-labeled protein samples from control culture C				
	S_P	R_P	S_A	S_B	S_C	cS_P	cR_P	cS_A	cS_B	cS_C
Quantitation category	$S_{P,1}$ $S_{P,2}$	$R_{P,1}$ $R_{P,2}$	$S_{A,1}$ $S_{A,2}$	$S_{B,1}$ $S_{B,2}$	$S_{C,1}$ $S_{C,2}$	$cS_{P,1}$ $cS_{P,2}$	$cR_{P,1}$ $cR_{P,2}$	$cS_{A,1}$ $cS_{A,2}$	$cS_{B,1}$ $cS_{B,2}$	$cS_{C,1}$ $cS_{C,2}$

Table 1. Twenty quantitation categories arising from the duplicate LC/MS analyses of the 3rd gel fraction for samples S_A, S_B, S_C, S_P, and R_P along with the labeled control in them

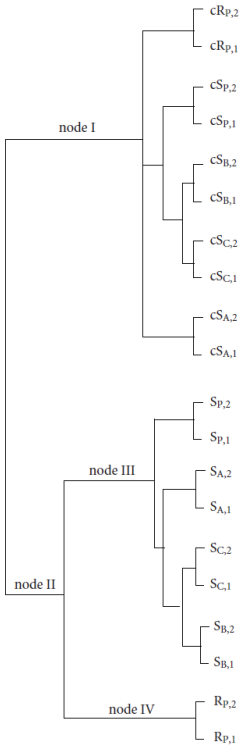

Fig. 6. Clustering of the 20 quantitation categories based on the 349 proteins quantified from the 3rd gel fraction for the five protein samples S_P, R_P, S_A, S_B, and S_C (Li & Roxas, 2009)

In Fig. 6, it was also apparent that the unlabeled and labelled quantitation categories were separated into two distinct branches represented by nodes I and II, respectively. The separation of the unlabeled and labelled quantitation categories into the two distinct clusters indicated that the difference between cultures C and S or C and R was larger than the difference between S and R. From the tree branch under node II, it could be seen that the distance between the unlabeled protein samples S_P and R_P was larger than the distance among the S culture replicates i.e., S_A, S_B, S_C. The result indicated that the difference between cultures S and R exceeded the difference among the S culture replicates, suggesting that the variability in biological sample replicates was less than the actual difference between the biological samples treated with different conditions.

Therefore, the clustering result in Fig. 6 indicated that the variability increased in the order of LC/MS injections < sample preparation replicates (under node I) ~ biological replicates (under node III) < biological samples (between nodes III and IV). Because these differences were evaluated based on the proteomic quantitation data, a variability observed among biological replicates also included the variability introduced during sample preparation for LC/MS analysis. The similarity between the variability observed among the sample preparation replicates and the variability observed among the biological replicates suggested that the variability among biological replicates was not larger than the variability among sample preparation replicates.

3.2.2 Extended selection of differentially regulated proteins

This subsection describes the multiple steps leading to the extended selection of differentially regulated proteins from all quantified proteins including those with only a single-peptide hit. The proteins with a single-peptide hit represent 1/3 of the identified proteins (Li & Roxas, 2009). Therefore, it is desirable to have a procedure to select regulated proteins from all of the proteins including those with a single-peptide hit to maximize the potential of the global protein expression profiling.

Establishing a null distribution

Based on the evaluation with the clustering analysis (Fig. 6), the variability among sample preparation replicates appeared to be comparable with that among biological replicates. Samples S_P and R_P represented the average of triplicate biological replicates for cultures S and R respectively, because each of them was the pooled sample of three biological replicates. The pooling process further reduced the biological variability between S_P and R_P. Therefore, the [^{15}N]-labelled control sample replicates (Table 1) were adequate to represent a null distribution in which there was no differentially regulated protein.

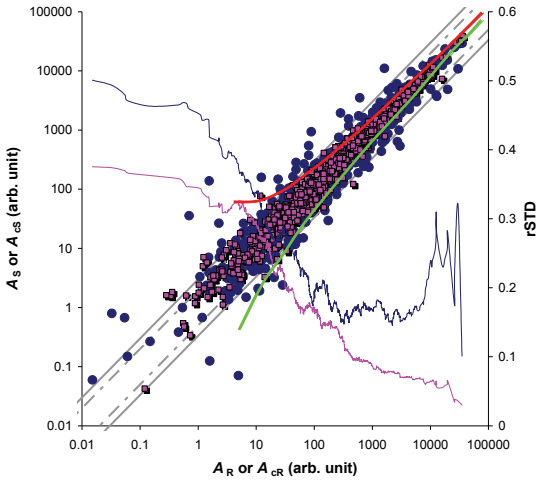

Fig. 7. A_{PRO} scatter plots, local variability, and thresholds for selecting differentially regulated proteins. The blue dots represent the A_{PRO} scatter plot of A_S vs. A_R corresponding to the unlabeled proteins in sample S_P vs R_P. A_S is the average of $A_{S,1}$ and $A_{S,2}$. A_R is the average of $A_{R,1}$ and $A_{R,2}$. The red dots represent the A_{PRO} scatter plot of A_{cS} vs. A_{cR} corresponding to the labeled proteins in control sample replicate cS_P vs cR_P. A_{cS} is the average of $A_{cS,1}$ and $A_{cS,2}$. A_{cR} is the average of $A_{cR,1}$ and $A_{cR,2}$. $A_{S,1}$, $A_{S,2}$, $A_{R,1}$, $A_{R,2}$, $A_{cS,1}$, $A_{cS,2}$, $A_{cR,1}$, and $A_{cR,2}$ were the A_{PRO} values for the eight quantitation categories defined in Table 1. To evaluate the local noise of A_{PRO} measurement, the relative standard deviation (rSTD) for each protein was calculated from its four unlabeled A_{PRO} values $A_{S,1}$, $A_{S,2}$, $A_{R,1}$, and $A_{R,2}$ (the blue trace) or its four labeled A_{PRO} values $A_{cS,1}$, $A_{cS,2}$, $A_{cR,1}$, and $A_{cR,2}$ (the pink trace). The rSTD- A_{PRO} traces were smoothed with a 100-point moving box. The grey straight lines indicated a 3-fold (solid line) and a 2-fold (dashed line) change threshold. The solid red and green curves represent the fold-change thresholds established with the PLGEM-STN statistics based on the local variance in the null distribution (the pink-dot scatter plot)

The null distribution afforded an estimation of measurement noise. The determined measurement noise was then used to estimate the false discovery rate for the selected differentially regulated proteins between samples S_P and R_P. The null distribution provided a reference for setting thresholds to maximize the selection of differentially regulated proteins (positives) while minimizing false positives. In **Fig. 7**, such a null distribution was illustrated with the scatter plot represented by the pink dots.

To investigate the relationship between measurement variability and protein abundance A_{PRO}, relative standard deviation (rSTD) was plotted against the mean A_{PRO} value for each protein in the unlabeled protein samples (blue trace) or the labelled control protein samples (pink trace) (Fig. 7). The rSTD-A_{PRO} trace in pink reflected the local noise of the null distribution. The local noise of the null distribution was mainly due to the variability that was introduced during sample preparation (Fig. 6). The rSTD-A_{PRO} trace in pink clearly suggested that the A_{PRO} measurement noise had a reciprocal dependence on the A_{PRO} amplitude. The rSTD-A_{PRO} trace in blue reflected both sample preparation variability and biological sample difference between cultures S and R. Thus, the blue trace had higher rSTD values than the pink trace throughout the A_{PRO} range.

Modelling local noise in the null distribution

Because of the reciprocal dependence of A_{PRO} rSTD on the A_{PRO} value, a universal 3-fold-change cut-off missed some positives at higher A_{PRO} values where a <3-fold change was already significantly different from the local noise. Missed positives at higher A_{PRO} values could be observed in Fig. 7 by examining the spread of the two scatter plots in the high A_{PRO} ranges. At A_{PRO} >1000, the rSTD was a few times smaller than that at A_{PRO} of ~100. From the figure, it could be seen that it was possible to detect a < 2-fold change for the proteins with A_{PRO} >1000. To the contrary, at A_{PRO} <10, a 3-fold change threshold was not sufficient to eliminate many false positives. Therefore, a criterion adaptive to the dependence of A_{PRO} noise on A_{PRO} values would uncover more differentially regulated proteins. This extended selection of differentially regulated proteins could be achieved by penalizing proteins with higher A_{PRO} values less than proteins with lower A_{PRO} values. Such an adaptive criterion, however, requires a systematic modelling of the noise to establish the thresholds according to local variability (Pavelka et al., 2004).

In this study, the PLGEM-STN statistic was experimented with for the selection of differentially regulated proteins quantified with label-free proteomics based on protein extracted ion chromatographic intensities. There were two reasons for the choice of the PLGEM-STN method.

First, the PLGEM-STN method allowed statistical analyses of the proteins quantified with a single PCS because the PLGEM-STN statistic did not rely on multiple PCSs of a protein like a t-test (Li & Roxas, 2009). Because single-peptide proteins constituted a third of the quantified proteins, being able to quantify these single-peptide proteins was important to maximize the potential value of the data. Second, the PLGEM-STN method took into account the dependence of A_{PRO} noise on A_{PRO} levels. A threshold adjustable to the local dependence of A_{PRO} noise on A_{PRO} levels allowed the selection of differentially regulated proteins with a smaller fold-change threshold at a higher A_{PRO} level.

Therefore, the PLGEM-STN method potentially could select more differentially regulated proteins by applying a smaller fold-change threshold in the higher A_{PRO} range where the variability was smaller. This possibility was tested as described below.

Selecting differentially regulated proteins with PLGEM-STN

PLGEM-STN confidence level	FP, P, and FDR	PLGEM-STN					PLGEM-STN-MPSP
		Permuted sample pairings				Average	
		I	II	III	IV		
0.01	FP (cS_P/cR_P)	31	68	22	46	42	13
	P (S_P/R_P)	141	155	134	148	145	101
	FDR	0.22	0.44	0.16	0.31	0.29	0.13
0.002	FP (cS_P/cR_P)	6	15	3	9	8	2
	P (S_P/R_P)	47	50	46	51	49	44
	FDR	0.13	0.30	0.07	0.18	0.16	0.05

Table 2. Numbers of differentially regulated proteins selected with PLGEM-STN alone or in combination with MPSP

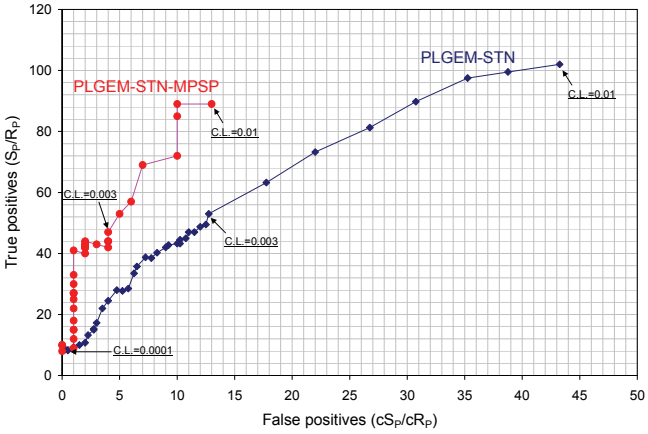

Fig. 8. Receiver operating characteristic analysis of the PLGEM-STN approach with (red curve) or without (blue curve) the combination with MPSP. *Positives* are the differentially regulated proteins selected from the comparison of protein abundances between samples S_P and R_P. *False positives* are the differentially regulated proteins selected from the comparison of proteins abundances between samples cS_P and cR_P. *True positives* are estimated by subtracting false positives from the positives. For each approach, i.e. PLGEM-STN-MPSP or PLGEM-STN, 37 data points at different confidence levels (C.L.) are plotted in this figure, starting from C.L.=0.0001 up to C.L.=0.01. The increment is 0.001 between C.L. of 0.0001 and 0.003 (30 data points). Between C.L. of 0.003 and 0.01, the increment is 0.01 (7 data points)

Table 2 shows the result of the PLGEM-STN analysis for the unlabeled samples S_P and R_P and the labelled sample replicates cS_P and cR_P. cS_P and cR_P were the labelled control samples analyzed concurrently with S_P and R_P, respectively. The differentially regulated proteins

found between S_P and R_P were positives (P), and those found between cS_P and cR_P were false positives (FP). Because each protein sample was analyzed with duplicate LC/MS injections, permutation of the four LC/MS injections for a sample pair resulted in four permuted sample pairings (Li & Roxas, 2009). These four permuted sample pairings were numbered as I to IV in Table 2. In each column for a permuted sample pairing in Table 3, the numbers of false positives and positives and the false discovery rate (FDR) were listed. The false positives were determined as the differentially regulated proteins for the sample pair cS_P/cR_P. The positives were determined as the differentially expressed proteins for the sample pair S_P/R_P.

In Table 2, the positives and false positives were selected with the PLGEM-STN method at the confidence level of 0.01 and 0.002, respectively. The results indicate that the numbers of positives or false positives were not the same among the four permuted sample pairings. To estimate an average false discovery rate, the numbers of positives and false positives were respectively averaged among the four permuted sample pairings. The false discovery rate was then calculated as the ratio of the average number of false positives divided by the average number of positives. The false discovery rate was determined at two different PLGEM-STN confidence levels (Table 2). With a receiver operating characteristic analysis, the PLGEM-STN approach is examined over a broader confidence level range (**Fig. 8**) and will be compared with another approach that is to be described below.

Incorporating the MPSP rule

Initially, the PLGEM-STN approach was carried out by comparing the duplicate LC/MS injections from the two samples R and S without permutation pairings. But the false discovery rate stayed high unless the sensitivity was severely compromised to reduce the false discovery rate. For example, at a confidence level of 0.0001, only 16 differentially regulated proteins were selected at 6% false discovery rate (data not shown). With all of the permutation pairs and a combination of PLGEM-STN and MPSP, 44 differentially regulated proteins were selected at a false discovery rate of 5% (Table 2). Therefore, a high sensitivity is achieved to uncover differentially regulated proteins by utilizing all possible permutation pairs with a combination of PLGEM-STN and MPSP.

Because of the variable numbers of positives and false positives among the four permuted sample pairings, it was necessary to determine a consensus list of differentially regulated proteins from the four permuted sample pairings. Previously, the rule of MPSP was applied to determine the consensus list of differentially regulated proteins from four permuted sample pairings (Li & Roxas, 2009). The MPSP rule required that only those proteins that were found differentially regulated in a certain number of permuted sample pairings were counted as positives (for S_P/R_P) or false positives (for cS_P/cR_P). When a sample pair such as S_P/R_P had no sample replicates but had duplicate LC/MS injections, MPSP was found to be optimum at four (Li & Roxas, 2009). Setting MPSP at four meant that a differentially regulated protein had to be found differentially regulated in all of the four permuted sample pairings.

Selecting differentially regulated proteins with a PLGEM-STN-MPSP approach

The application of the MPSP rule towards the PLGEM-STN results decreased both false positives and positives (Table 2). But the false discovery rate was also decreased relative to that when only the PLGEM-STN statistic was applied. From Table 2, it could be seen that the number of true positives, which was estimated from the difference between the numbers of

positives and false positives, remained about the same. Therefore, the combination of the MPSP rule with the PLEGM-STN method reduced the false discovery rate by 2-3 times without compromising the sensitivity.

As summarized in Fig. 8, the receiver operating characteristic analysis clearly shows that the PLGEM-STN-MPSP approach significantly reduces false positives to improve the specificity without significantly affecting the sensitivity. Compared to the use of the PLGEM-STN statistic alone, the combination of PLGEM-STN and MPSP performs better in controlling false discovery rates without compromising the sensitivity to select differentially regulated proteins.

Selecting differentially regulated proteins with a fold-change-MPSP approach

Fold change	FP, P, and FDR	Fold-change					Fold-change-MPSP
		Permuted sample pairings				Average	
		I	II	III	IV		
2	FP (cS_P/cR_P)	68	77	118	45	77	22
	P (S_P/R_P)	171	154	186	147	165	104
	FDR	0.40	0.50	0.63	0.31	0.47	0.21
3	FP (cS_P/cR_P)	30	33	47	20	33	9
	P (S_P/R_P)	66	70	85	60	70	42
	FDR	0.45	0.47	0.55	0.33	0.47	0.21
4	FP (cS_P/cR_P)	17	24	32	10	21	1
	P (S_P/R_P)	42	50	53	35	45	26
	FDR	0.40	0.48	0.60	0.29	0.47	0.04

Table 3. Number of differentially regulated proteins selected with a fold-change threshold alone or in combination with MPSP

The use of MPSP with fold-change criteria was also examined (**Table 3**). With fold-change criteria alone, the false discovery rate did not drop below 46% at 2- to 4-fold changes. With the combination of MPSP and the fold-change criteria, the false discovery rate was reduced from 46% to 21% at 2- and 3-fold changes. At a 4-fold change, the false discovery rate was reduced to 4%. Compared to the combination of PLGEM-STN and MPSP, however, the combination of fold-change and MPSP reduced more true positives at the similar false discovery rate of 4-5%. Therefore, the application of MPSP with the fold-change criteria reduced sensitivity. The reduced sensitivity was due to the increase in the fold-change threshold.

With the 4-fold-change-MPSP and the PLGEM-STN-MPSP approaches, 26 and 44 proteins were respectively selected as differentially regulated at a false discovery rate of 4% or 5% (Tables 2 and 3). Among these 26 and 44 proteins, there were 55 unique proteins (Li, 2010b). These 55 unique proteins included all of the 20 high-confidence differentially regulated proteins identified previously with an empirical fold-change and abundance level cut-off approach (Roxas & Li, 2009).

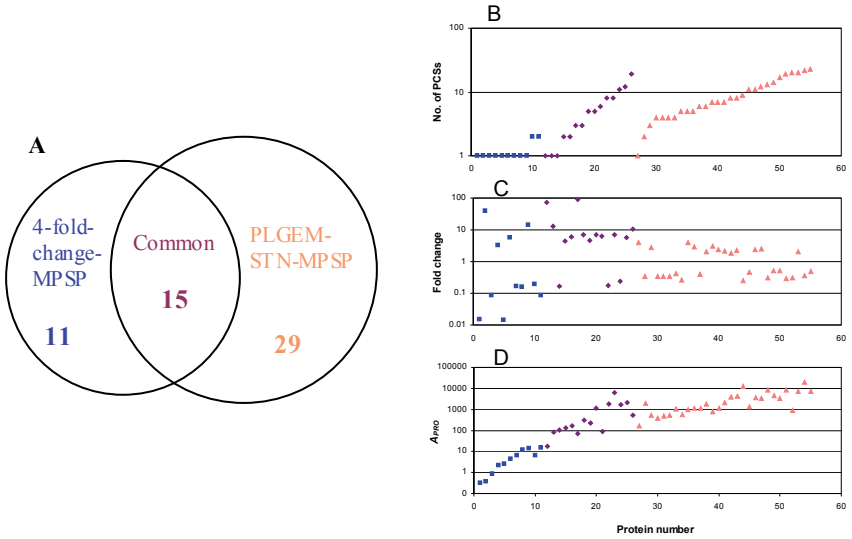

Fig. 9. Comparison of the 26 and 44 differentially regulated proteins respectively selected by the 4-fold-change-MPSP and PLGEM-STN-MPSP approaches at a 5% false discovery rate. (A) Overlap of the two sets of differentially regulated proteins. Panels B-D show the distributions of (B) the number of detected PCSs, (C) the fold changes, and (D) the abundances of the quantified proteins. The blue square, the purple diamond, and the tan triangle markers represent the differentially regulated proteins selected by 4-fold-change-MPSP only, by both, and by PLGEM-STN-MPSP respectively. The protein number was from 1 to 55 on the x-axis representing the 55 unique proteins ranked according to their A_{PRO} in each of the three groups (blue, purple, or tan)

Comparing the PLGEM-STN-MPSP and fold-change-MPSP approaches

Only 15 proteins were common between the two sets of differentially regulated proteins selected with the 4-fold-change-MPSP and the PLGEM-STN-MPSP approaches (**Fig. 9A**). The 4-fold-change-MPSP approach selected more single-PCS proteins than the PLGEM-STN-MPSP approach (Fig. 9B). The PLGEM-STN-MPSP approach selected proteins with a fold-change as low as 1.8-fold (Fig. 9C). However, these differentially regulated proteins selected with PLGEM-STN-MPSP had a protein abundance higher than most of the differentially regulated proteins selected with the 4-fold-change-MPSP approach (Fig. 9D). Thus, the two approaches complement each other and could be used simultaneously.

3.3 Discussions

3.3.1 Motivation of the extensive label-free quantitative proteomics analysis

Despite the relative complexity in label-free proteomics data analysis and the demand of more stringently controlled LC/MS experimental conditions, there are strong motivations stemming from biological and experimental perspectives to use the label-free approach, as discussed below.

As shown in Fig. 6, the unlabeled and labelled quantitation categories are separated into two distinct clusters. One includes the quantitation categories from the labelled control culture C

(under node I). The other includes the quantitation categories from the two unlabeled cultures S and R (under node II). Thus, there was a larger difference between the labelled (C) and either of the two unlabeled samples (S or R) than between the two unlabeled cultures (S and R). The number of differentially regulated proteins between the labelled culture and either of the unlabeled culture was about three times as many as that between the two unlabeled cultures. Compared to the difference between the two unlabeled cultures, the difference between the labelled culture and either of the unlabeled cultures was larger. This larger difference was probably because the labelled culture was cultured in a synthetic minimal medium while the two unlabeled cultures were grown in a commercial 7H9 broth that was richer in ingredients. Another factor was that the acidic growth condition was a relatively mild stress so that not many proteins were differentially regulated.

The apparent difference in proteome profile for cells cultured in different media is actually a strong motivation for this study. In microbiological works, it is not always convenient to make a [^{15}N]-labelled medium with complex ingredients required to cultivate bacteria under more physiologically relevant conditions. Even some of the stable-isotope-labelled media are technically feasible to make, they often bear a costly price tag. For microbiological works, one might not want to be restricted by the type of medium that can be used because of the stable isotope labelling limitation. For example, some mycobacteria are difficult to cultivate on simple synthetic media and prefer complex media. Thus, unlabeled media are always convenient choices if the down-stream proteomic analysis is established to proceed with the quantitation.

For such reasons, the focus of this study was on the comparison of protein expression profiles between the two unlabeled cultures S and R. The labelled control culture C was used as an internal standard to estimate false discovery rates.

3.3.2 The use of a [^{15}N]-labeled internal standard for null distribution construction

The label-free quantitation scheme presented in this study incorporated a labelled internal control to provide replicates for noise modelling without a requirement of other unlabeled sample replicates. The inclusion of a labelled internal control facilitates the estimation and control of false discovery rates.

Internal standards are commonly used to improve reliability of quantitative proteomics such as to aid in removing outlier data and to detect fluctuation in instrument performance (Mirzaei et al., 2009). Compared to other synthetic peptide internal standards(Mirzaei et al., 2009; Winter et al., 2010), the [^{15}N]-labelled control culture C provides more comprehensive peptide internal standards. For most of the peptides, the extracted ion chromatographic intensities can be matched among the three protein samples originated from the two unlabeled (S and R) cultures and the labelled (C) culture. The C protein sample was mixed and run together with either S or R protein sample, so that the reliability of the internal standards was improved.

To construct the null distribution for the error model in PLGEM-STN, it would be ideal to have the labelled internal standard identical to an unlabeled sample in protein composition. As mentioned above, however, that requirement could restrict the culturing conditions available for biological experiments. Thus, it is acceptable and sometimes necessary to use a labelled protein mixture sample as internal standard, even though the internal standard sample might be somewhat different from the unlabeled samples in protein abundance profiles.

Nevertheless, the null distribution is only utilized to establish the relation between the signal-to-noise ratio and the peptide abundance in the PLGEM-STN method. There is no requirement of direct one-to-one comparison between the labelled and unlabeled version of a protein during this process. Therefore, the difference in proteome composition between the labelled internal standard sample C and the two unlabeled samples S and R is not expected to affect the modelling parameters derived from the null distribution constructed from the labelled C sample.
One could choose to run multiple replicates of an unlabeled sample and use the replicates to construct the null distribution. That approach would require more LC/MS runs as discussed previously (Li & Roxas, 2009).

3.3.3 Label-free data analyses and selection of differentially regulated proteins

The LC/MS data used in this work was acquired with a high-resolution mass spectrometer that resolved peptide peaks from a complex sample mixture to allow the determination of the extracted ion chromatographic intensities of peptides and proteins. Repeated LC/MS injections showed the highest reproducibility among several other types of replicates, indicating that the major variability of the label-free quantitation did not lie within the LC/MS separation and the data analysis method. Rather, sample preparation replicates represented a major source of the variability. With a labelled control sample to run concurrently with each of the unlabeled samples, replicates for the labelled control sample were obtained. The control sample replicates provided data to model the noise in the label-free quantitation with extracted ion chromatographic intensities.
We performed a two-step normalization procedure in which the information about the abundance of a peptide or protein in a sample was preserved (Li, 2010b). The preservation of the information about the abundance of a peptide or protein in the samples is critical for performing the PLGEM-STN analysis. In addition, because protein extracted ion chromatographic intensity was represented by the sum of the PCS extracted ion chromatographic intensities belonging to that protein, the summation weighed the low-intensity PCSs less than the high-intensity PCSs. Such a summation of PCS extracted ion chromatographic intensities probably suppressed noise from lower-intensity PCSs. When a protein abundance ratio is calculated as the average of PCS abundance ratios without weighing, the noise from a lower-intensity PCS would be amplified. We have avoided this potential issue by summing the PCS intensities to represent protein abundances before calculating protein abundance ratios.
Single-peptide proteins made up about 35% of the quantified proteins (Li, 2010b). Selection of differentially regulated proteins from these single-peptide proteins required a significance assessment method that did not rely on multiple-peptide detection to calculate a statistic about the confidence of a protein differential abundance. The use of a statistic that does not rely on the detection of multiple peptides is especially useful when the sample replicates were too low to use a typical statistical test such as a t-test. PLGEM-STN was a method that fits this criterion.
However, PLGEM-STN alone was not strict enough to control the false discovery rate without further diminishing the number of positives (Fig. 8). The lack of stringency by using the PLGEM-STN method alone was similar to that by using the t-test alone (Li & Roxas, 2009). In that prior study, the lack of specificity with a t-test alone was overcome by introducing the rule MPSP. The MPSP rule simply requires that a protein be selected as differentially regulated only when it was repeatedly found so in certain number of

permuted sample pairings. The MPSP rule was introduced to deal with datasets with small replicates where other more sophisticated statistical tests could not be applied (Li & Roxas, 2009). Although the MPSP rule was originally used in combination with a t-test statistic and a fold-change threshold, this study shows that it can be used in combination with other types of statistical tests such as the PLGEM-STN method (Fig. 8).

The combination of the MPSP rule allowed the selection of differentially regulated proteins at a false discovery rate <5%, which would have been impossible for a fold-change method (Table 3). The MPSP rule significantly reduced false positives while keeping the number of true positives relatively constant, thus effectively improving the statistical confidence of the selected differentially regulated proteins by lowering the false discovery rate (Table 3). The results suggest that MPSP is a rule that can be used in combination with different types of statistics to select differentially regulated proteins.

The label-free quantitation simplified cell culturing and sample preparation. Another useful aspect of the label-free quantitation is that peptide cross-reference could be used to increase the number of proteins quantified in all of the samples run under the same condition (Andreev et al., 2007). Lipton et al. introduced the concept of accurate mass and elution time peptide tag for global protein quantitation using high resolution mass spectrometry (Lipton et al., 2002). One advantage of this method over using the spectral counting method is that the large number of identifications that occur in a LC/MS injection can be used as the basis for improved quantitation of another LC/MS injection (Andreev et al., 2007; Fang et al., 2006; Strittmatter et al., 2003). The accurate mass and elution time peptide tag approach uses the extracted ion chromatographic intensities as the quantitative measurement of peptides and proteins. The linear response of peptide extracted ion chromatographic intensities to protein quantities was demonstrated (Hochleitner et al., 2005; Lundrigan et al., 1997; Wang et al., 2006). This method was thus used to improve the comparability of proteins quantified between samples, among LC/MS injections, and for different isotopic forms of a protein (Rao et al., 2008a). The quantitation of 349 proteins from a single gel fraction for several samples clearly demonstrated the power of the peptide cross-reference feature in extracted ion chromatographic intensity-based label-free quantitative proteomics (Li & Roxas, 2009).

One drawback of extracted ion chromatographic intensity-based label-free quantitative proteomics is that the success of an analysis critically depends upon the reproducibility of LC/MS runs that have to be maintained across multiple samples. The reproducibility of LC/MS runs across multiple samples is a prerequisite to reliable peptide cross reference (Andreev et al., 2007). With the advancement in LC/MS instrumentation and the availability of improved LC/MS chromatogram alignment methods (Fischer et al., 2006; Podwojski et al., 2009), the reproducibility of LC/MS runs is unlikely to remain an obstacle for the increasing use of label-free quantitative proteomics.

3.4 Summary

A label-free quantitative proteomics scheme was demonstrated to select differentially regulated proteins with single-peptide hits and <2-fold changes at a 5% false discovery rate. The scheme incorporated a labeled internal control into multiple unlabeled samples to facilitate error modeling when there were no replicates for the unlabeled samples. The error modeling allowed the use of the PLGEM-STN statistic to facilitate the selection of differentially regulated proteins with single-peptide hits. The PLGEM-STN statistic also facilitated the selection of differentially regulated proteins at different fold-change

thresholds according to the local abundance level of the proteins. While the PLGEM-STN statistic uncovered more differentially regulated proteins at higher abundance with smaller fold-changes, the PLGEM error modeling of local variance versus abundance over-penalized the proteins with lower abundance. With a constant fold-change threshold, however, differentially regulated proteins with higher abundance were overlooked. Thus, the results from this study showed that the PLGEM-STN and a constant fold-change threshold were complementary to each other and could be used simultaneously. But, neither the PLGEM-STN nor the 4-fold-change criterion alone was stringent enough for selecting differentially regulated proteins at a 5% false discovery rate.

MPSP was introduced and shown to be a rule that could decrease false discovery rates when being used in combination with the PLGEM-STN statistic or the 4-fold-change threshold. The MPSP rule played a critical role in extending the selection of differentially regulated proteins to those with single-peptide hit or with a lower fold-change in label-free proteomics when sample replicates were limited. Although the approaches were demonstrated for a representative replicate-limited scenario, they potentially can also be applicable to a situation where more sample replicates are available.

4. Conclusion

This chapter presents several examples of proteomic studies utilizing the LTQ-FTMS instrument. The instrument typically provides a high resolution (> 500,000), a large mass range (one order magnitude in a single scan), and high mass accuracy (< 2 ppm) in many experiments.

The protein turnover studies benefit greatly from the high resolution in a large mass range, which allows the resolution of isotopomers and isotopologue profiles that could possibly be generated by almost any degree and type of stable isotope labeling. The well resolved full-range spectra simplify the data processing steps and improve data quality. Because a spectral count method reports the MS^2 events mostly on monoisotopic peaks, it is usually not suitable for protein turnover study. Therefore, a high-resolution mass spectrometer such as an FTMS instrument is highly desired for protein turnover studies.

The high mass accuracy is critical in the implementation of a label-free proteomics approach. The label-free quantitative proteomics relies on a cross reference of peptides between runs and an integration of extracted ion chromatographic intensities. The peptide cross reference allows the quantitation of a peptide in a run in which the peptide is not identified but is identified in another run. It is based on the accurate mass and reproducible liquid chromatographic elution time of the peptide in multiple runs. The cross reference allows more peptides to be quantified. It also improves the comparability of samples because the same peptides, thus the same proteins, are quantified in multiple samples.

The label-free quantitation is useful not only for protein differential expression studies, but also for proteome dynamics studies where the abundances of newly synthesized proteins are to be quantified separate from those of the total proteins and the old proteins.

Thus, the label-free quantitation approach is universally applicable to different proteomic studies. The assessment of statistical significance in such a quantitation approach is challenging especially for the proteins with few peptides identified and when the replicates are limited. A combination of the MPSP rule with a global error modeling method and a fold-change threshold proves to be effective in controlling the false discovery rates in protein differential analysis.

5. References

Andreev, V. P.; Li, L.; Cao, L., et al. (2007). A new algorithm using cross-assignment for label-free quantitation with LC-LTQ-FT MS. *J Proteome Res,* 6, 6, 2186-94.

Bernlohr, R. W. (1967). Changes in amino acid permeation during sporulation. *J Bacteriol,* 93, 3, 1031-44.

Bernlohr, R. W. (1972). 18 Oxygen probes of protein turnover, amino acid transport, and protein synthesis in Bacillus licheniformis. *J Biol Chem,* 247, 15, 4893-9.

Beynon, R. J. (2005). The dynamics of the proteome: strategies for measuring protein turnover on a proteome-wide scale. *Brief Funct Genomic Proteomic,* 3, 4, 382-90.

Borek, E.; Ponticorvo, L. & Rittenberg, D. (1958). Protein Turnover in Micro-Organisms. *Proc Natl Acad Sci U S A,* 44, 5, 369-74.

Cargile, B. J.; Bundy, J. L.; Grunden, A. M. & Stephenson, J. L., Jr. (2004). Synthesis/degradation ratio mass spectrometry for measuring relative dynamic protein turnover. *Anal Chem,* 76, 1, 86-97.

Cho, S. H.; Goodlett, D. & Franzblau, S. (2006). ICAT-based comparative proteomic analysis of non-replicating persistent Mycobacterium tuberculosis. *Tuberculosis (Edinb),* 86, 6, 445-60.

Doherty, M. K. & Beynon, R. J. (2006). Protein turnover on the scale of the proteome. *Expert Rev Proteomics,* 3, 1, 97-110.

Eisen, M. B.; Spellman, P. T.; Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A,* 95, 25, 14863-8.

Fang, R.; Elias, D. A.; Monroe, M. E., et al. (2006). 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,* 5, 4, 714-25.

Fischer, B.; Grossmann, J.; Roth, V., et al. (2006). Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics,* 22, 14, e132-40.

Gerner, C.; Vejda, S.; Gelbmann, D., et al. (2002). Concomitant determination of absolute values of cellular protein amounts, synthesis rates, and turnover rates by quantitative proteome profiling. *Mol Cell Proteomics,* 1, 7, 528-37.

Goldberg, A. L. & Dice, J. F. (1974). Intracellular protein degradation in mammalian and bacterial cells. *Annu Rev Biochem,* 43, 0, 835-69.

Gupta, N. & Pevzner, P. A. (2009). False discovery rates of protein identifications: a strike against the two-peptide rule. *J Proteome Res,* 8, 9, 4173-81.

Hochleitner, E. O.; Kastner, B.; Frohlich, T., et al. (2005). Protein stoichiometry of a multiprotein complex, the human spliceosomal U1 small nuclear ribonucleoprotein: absolute quantification using isotope-coded tags and mass spectrometry. *J Biol Chem,* 280, 4, 2536-42.

Larrabee, K. L.; Phillips, J. O.; Williams, G. J. & Larrabee, A. R. (1980). The relative rates of protein synthesis and degradation in a growing culture of Escherichia coli. *J Biol Chem,* 255, 9, 4125-30.

Li, Q. (2010a). Advances in protein turnover analysis at the global level and biological insights. *Mass Spectrom Rev,* 29, 5, 717-36.

Li, Q. (2010b). Assigning significance in label-free quantitative proteomics to include single-peptide-hit proteins with low replicates. *Int J Proteomics,* 2010, 731582, 15 pages.

Li, Q. & Roxas, B. A. (2009). An assessment of false discovery rates and statistical significance in label-free quantitative proteomics with combined filters. *BMC Bioinformatics,* 10, 43.

Lipton, M. S.; Pasa-Tolic, L.; Anderson, G. A., et al. (2002). Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. *Proc Natl Acad Sci U S A,* 99, 17, 11049-54.

Lundrigan, M. D.; Arceneaux, J. E.; Zhu, W. & Byers, B. R. (1997). Enhanced hydrogen peroxide sensitivity and altered stress protein expression in iron-starved Mycobacterium smegmatis. *Biometals,* 10, 3, 215-25.

Malen, H.; Berven, F. S.; Fladmark, K. E. & Wiker, H. G. (2007). Comprehensive analysis of exported proteins from Mycobacterium tuberculosis H37Rv. *Proteomics,* 7, 10, 1702-18.

Mandelstam, J. (1958). Turnover of protein in growing and non-growing populations of Escherichia coli. *Biochem J,* 69, 1, 110-9.

Mirzaei, H.; Brusniak, M. Y.; Mueller, L. N., et al. (2009). Halogenated peptides as internal standards (H-PINS): introduction of an MS-based internal standard set for liquid chromatography-mass spectrometry. *Mol Cell Proteomics,* 8, 8, 1934-46.

Mortensen, P.; Gouw, J. W.; Olsen, J. V., et al. (2010). MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J Proteome Res,* 9, 1, 393-403.

Mueller, L. N.; Brusniak, M. Y.; Mani, D. R. & Aebersold, R. (2008). An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J Proteome Res,* 7, 1, 51-61.

Murphy, D. J. & Brown, J. R. (2008). Novel drug target strategies against Mycobacterium tuberculosis. *Curr Opin Microbiol,* 11, 5, 422-7.

Pagan-Ramos, E.; Master, S. S.; Pritchett, C. L., et al. (2006). Molecular and physiological effects of mycobacterial oxyR inactivation. *J Bacteriol,* 188, 7, 2674-80.

Pavelka, N.; Fournier, M. L.; Swanson, S. K., et al. (2008). Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol Cell Proteomics,* 7, 4, 631-44.

Pavelka, N.; Pelizzola, M.; Vizzardelli, C., et al. (2004). A power law global error model for the identification of differentially expressed genes in microarray data. *BMC Bioinformatics,* 5, 203.

Podwojski, K.; Fritsch, A.; Chamrad, D. C., et al. (2009). Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics,* 25, 6, 758-64.

Pratt, J. M.; Petty, J.; Riba-Garcia, I., et al. (2002). Dynamics of protein turnover, a missing dimension in proteomics. *Mol Cell Proteomics,* 1, 8, 579-91.

Rao, P. K. & Li, Q. (2009a). Principal Component Analysis of Proteome Dynamics in Iron-Starved Mycobacterium Tuberculosis. *J Proteomics Bioinform,* 2, 19-31.

Rao, P. K. & Li, Q. (2009b). Protein turnover in mycobacterial proteomics. *Molecules*, 14, 9, 3237-58.

Rao, P. K.; Rodriguez, G. M.; Smith, I. & Li, Q. (2008a). Protein dynamics in iron-starved Mycobacterium tuberculosis revealed by turnover and abundance measurement using hybrid-linear ion trap-Fourier transform mass spectrometry. *Anal Chem,* 80, 18, 6860-9.

Rao, P. K.; Roxas, B. A. & Li, Q. (2008b). Determination of global protein turnover in stressed mycobacterium cells using hybrid-linear ion trap-fourier transform mass spectrometry. *Anal Chem,* 80, 2, 396-406.

Rosenkrands, I.; Slayden, R. A.; Crawford, J., et al. (2002). Hypoxic response of Mycobacterium tuberculosis studied by metabolic labeling and proteome analysis of cellular and extracellular proteins. *J Bacteriol,* 184, 13, 3485-91.

Roxas, B. A. & Li, Q. (2009). Acid stress response of a mycobacterial proteome: insight from a gene ontology analysis. *Int J Clin Exp Med,* 2, 309-328.

Spudich, J. A. & Kornberg, A. (1968). Biochemical studies of bacterial sporulation and germaination. VII. Protein turnover during sporulation of Bacillus subtilis. *J Biol Chem,* 243, 17, 4600-5.

Strittmatter, E. F.; Ferguson, P. L.; Tang, K. & Smith, R. D. (2003). Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *J Am Soc Mass Spectrom,* 14, 9, 980-91.

Vogt, J. A.; Hunzinger, C.; Schroer, K., et al. (2005). Determination of fractional synthesis rates of mouse hepatic proteins via metabolic 13C-labeling, MALDI-TOF MS and analysis of relative isotopologue abundances using average masses. *Anal Chem,* 77, 7, 2034-42.

Wang, G.; Wu, W. W.; Zeng, W.; Chou, C. L. & Shen, R. F. (2006). Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes. *J Proteome Res,* 5, 5, 1214-23.

Wilkinson, K. D. (2005). The discovery of ubiquitin-dependent proteolysis. *Proc Natl Acad Sci U S A,* 102, 43, 15280-2.

Winter, D.; Seidler, J.; Kugelstadt, D., et al. (2010). Minimally permutated peptide analogs as internal standards for relative and absolute quantification of peptides and proteins. *Proteomics,* 10, 7, 1510-4.

Zhu, W.; Smith, J. W. & Huang, C. M. (2010). Mass spectrometry-based label-free quantitative proteomics. *J Biomed Biotechnol,* 2010, 840518, 6 pages.

Fourier Transforms - New Analytical Approaches and FTIR Strategies

Edited by Prof. Goran Nikolic

ISBN 978-953-307-232-6
Hard cover, 520 pages
Publisher InTech
Published online 01, April, 2011
Published in print edition April, 2011

New analytical strategies and techniques are necessary to meet requirements of modern technologies and new materials. In this sense, this book provides a thorough review of current analytical approaches, industrial practices, and strategies in Fourier transform application.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Qingbo Li (2011). Precision Quantitative Proteomics with Fourier-Transform Mass Spectrometry, Fourier Transforms - New Analytical Approaches and FTIR Strategies, Prof. Goran Nikolic (Ed.), ISBN: 978-953-307-232-6, InTech, Available from: http://www.intechopen.com/books/fourier-transforms-new-analytical-approaches-and-ftir-strategies/precision-quantitative-proteomics-with-fourier-transform-mass-spectrometry

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821