Detecting Differential and Correlated Protein Expression in Label-Free Shotgun Proteomics

Bing Zhang,*[,†,‡,+] Nathan C. VerBerkmoes,[§] Michael A. Langston,[‡,||] Edward Uberbacher,[‡,⊥] Robert L. Hettich,[§] and Nagiza F. Samatova*[,†,‡]

Computer Science and Mathematics Division, Computational Biology Institute, Chemical Science Division, and Life Science Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, and Department of Computer Science, University of Tennessee, Knoxville, Tennessee 37996

Received January 26, 2006

Recent studies have revealed a relationship between protein abundance and sampling statistics, such as sequence coverage, peptide count, and spectral count, in label-free liquid chromatography−tandem mass spectrometry (LC−MS/MS) shotgun proteomics. The use of sampling statistics offers a promising method of measuring relative protein abundance and detecting differentially expressed or coexpressed proteins. We performed a systematic analysis of various approaches to quantifying differential protein expression in eukaryotic *Saccharomyces cerevisiae* and prokaryotic *Rhodopseudomonas palustris* label-free LC−MS/MS data. First, we showed that, among three sampling statistics, the spectral count has the highest technical reproducibility, followed by the less-reproducible peptide count and relatively nonreproducible sequence coverage. Second, we used spectral count statistics to measure differential protein expression in pairwise experiments using five statistical tests: Fisher's exact test, *G*-test, AC test, *t*-test, and LPE test. Given the *S. cerevisiae* data set with spiked proteins as a benchmark and the false positive rate as a metric, our evaluation suggested that the Fisher's exact test, *G*-test, and AC test can be used when the number of replications is limited (one or two), whereas the *t*-test is useful with three or more replicates available. Third, we generalized the *G*-test to increase the sensitivity of detecting differential protein expression under multiple experimental conditions. Out of 1622 identified *R. palustris* proteins in the LC−MS/MS experiment, the generalized *G*-test detected 1119 differentially expressed proteins under six growth conditions. Finally, we studied correlated expression of these 1119 proteins by analyzing pairwise expression correlations and by delineating protein clusters according to expression patterns. Through pairwise expression correlation analysis, we demonstrated that proteins co-located in the same operon were much more strongly coexpressed than those from different operons. Combining cluster analysis with existing protein functional annotations, we identified six protein clusters with known biological significance. In summary, the proposed generalized *G*-test using spectral count sampling statistics is a viable methodology for robust quantification of relative protein abundance and for sensitive detection of biologically significant differential protein expression under multiple experimental conditions in label-free shotgun proteomics.

Keywords: label-free • LC−MS/MS • shotgun proteomics • differential expression • correlated expression • clustering • *Saccharomyces cerevisiae* • *Rhodopseudomonas palustris*

Introduction

Quantification of protein abundance to detect differential protein expression is a fundamental challenge in proteomics. A mass spectrometry (MS)-based approach is emerging and promising. *Label-free* liquid chromatography−tandem mass spectrometry (LC−MS/MS) using the "shotgun" approach is particularly effective for large-scale protein identification.[1,2] In addition, the relative protein abundance is related to the observed sampling statistics for this methodology.[3−4] Namely, the *spectral count*, or the total number of MS/MS spectra taken on peptides from a given protein in a given LC/LC−MS/MS analysis, is linearly correlated with the protein abundance over a dynamic range of 2 orders of magnitude.[4] Protein ratios determined by spectral count agree well with those determined from peak area intensity measurements, and both are consis-

* To whom correspondence should be addressed. E-mail: bing.zhang@vanderbilt.edu (B.Z.), samatovan@ornl.gov (N.F.S.); Phone: 865-241-4351; Fax: 865-576-5491.

[†] Computer Science and Mathematics Division.

[‡] Computational Biology Institute.

[§] Chemical Science Division.

[||] Department of Computer Science.

[⊥] Life Science Division.

[+] Current address: Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232.

10.1021/pr0600273 CCC: $33.50

Published on Web 10/11/2006

tent with independent measurements based on gel staining intensities.[5] Protein abundance is also correlated with other sampling statistics, such as the *sequence coverage*[3] (the total coverage of a protein sequence by the peptides identified in a given LC/LC−MS/MS analysis) and the *peptide count*[6] (the total number of peptides identified from a protein in a given LC/LC−MS/MS analysis). Thus, the use of sampling statistics seems to be a simple and practical approach to measuring relative protein abundance. The question remains as to which sampling statistics perform the "best" (i.e., are the most robust and sensitive) for protein quantification.

Because sample handling in proteome measurements is highly complex, proteome quantification requires rigorous statistical approaches. If the measurements are assumed to be normally distributed, the *t*-test is often used for testing the equality of the means of two populations. However, this assumption cannot be guaranteed for shotgun proteomics data. The *t*-test also requires multiple replicates (three or more) for the estimation of variance associated with the measurements, but proteomic studies frequently have only one or two replicates. The Local-Pooled-Error (LPE) test,[7] a variant of the *t*-test, identifies differentially expressed proteins with at least two replicates. If no experimental replicates are available, the *G* test[8] and its analogues (Fisher's exact test, χ^2 test, and AC test) are used. They require no variance estimation. If multiple replicates exist, however, these statistical tests are not able to incorporate the inherent variability across these replicates.

Most of the published studies on differential protein expression have focused on a *pairwise* comparison of experimental conditions. Many biological experiments, however, are designed to study *multiple* conditions. Although pairwise comparisons among multiple conditions could still be used, they may introduce type I errors (false positives, i.e., proteins identified as differentially expressed, though they are not) due to the increased number of comparisons.[9] When attempting to reduce type I errors, no method is currently available for the *simultaneous*, rather than *pairwise*, comparison of multi-condition experiments in label-free shotgun proteomics. To fill this gap, one might consider applying a common statistical method for testing the equality of the means among three or more populations simultaneously with the one-way analysis of variance (ANOVA). Similar to the *t*-test, one-way ANOVA assumes that the data points are normally distributed and requires multiple replicates for each population. The issue remains whether it is possible to detect differential protein expression in multi-condition proteomics studies while eliminating the multiple replicates requirement.

In addition to *differential* protein expression, *correlated* protein expression is a complementary measurement that could be applied in the analysis of multi-condition protein expression data. For instance, cluster analysis of quantified protein expression data from SILAC (Stable Isotope Labeling with Amino acids in Cell culture)[10] has revealed clusters of functionally related proteins. However, for label-free shotgun proteomics data, the value of such an approach remains to be established.

In this paper, we perform comparative analysis of various approaches to measuring differential and correlated protein expression using eukaryotic *Saccharomyces cerevisiae*[4] and prokaryotic proteobacteria *Rhodopseudomonas palustris*[11] LC−MS/MS label-free shotgun proteomic data. We first study the technical reproducibility of the sampling statistics (peptide count, spectral count, and sequence coverage) using the *R. palustris* data set. For the most reproducible sampling statistics, the *S. cerevisiae* data set with spiked proteins of various concentrations is then used to evaluate how well various statistical tests detect differential expression in *pairwise* experiments. To enable the quantification of differential expression under *multiple* conditions simultaneously, we extend the methodology through the generalized *G*-test. We then demonstrate its applicability using an *R. palustris* data set that contains replicate LC−MS/MS measurements of proteomes from six different metabolic states. Finally, we perform cluster analysis of such quantified *R. palustris* data to detect the groups of coordinately expressed proteins and statistically evaluate the functional significance of the identified protein clusters.

Materials and Methods

Data Sets. Two data sets were used in this study. The first came from *S. cerevisiae* cell lysate samples spiked with three different concentrations (2.5, 1.25, and 0.25%) of a protein marker mixture,[4] composed of bovine carbonic anhydrase (CAH2), bovine serum albumin (ALBU), soybean trypsin inhibitor (ITRA), chicken lysozyme (LYC), chicken ovalbumin (OVAL), and rabbit phosphorylase b (PHS2). Three parallel multidimensional protein identification technology (MudPIT) experiments were performed for each concentration of protein markers. To rebuild the complete data set from the nine experiments, data from the spike-in proteins were mapped to those from the yeast proteins as follows:

(a) for the 2.5% concentration: experiment 1 → 0503-Y10_SC, experiment 2 → 0508-Y10_SC, experiment 3 → 0517-Y10_SC;

(b) for the 1.25% concentration: experiment 1 → 0505-Y5_SC, experiment 2 → 0510-Y5_SC, experiment 3 → 0520-Y5_SC;

(c) for the 0.25% concentration: experiment 1 → 0504-Y1_SC, experiment 2 → 0509-Y1_SC, experiment 3 → 0518-Y1_SC.

Note that the mapping information between the spectral counts for the spiked proteins and those for the yeast proteins was kindly provided by Dr. Hongbin Liu (Agilent Technologies, Inc., Wilmington, DE 19808).

The second data set came from our proteomics study[11] of *R. palustris* CGA010 wild-type strain grown under six different conditions labeled as:

(i) *Dark* for chemoheterotrophic aerobic growth in the dark with succinate;

(ii) *Light* for photoheterotrophic anaerobic growth in the light with succinate;

(iii) *Benzoate* for photoheterotrophic anaerobic growth in the light with benzoate;

(iv) *Auto* for photoautotrophic anaerobic growth in the light with sodium bicarbonate and H_2;

(v) *N2* for photoheterotrophic nitrogen fixing growth in the light; and

(vi) *Stationary* for photoheterotrophic stationary phase growth in the light with succinate.

Except for the *Stationary* condition, in which the cells were grown into stationary phase, the cells were harvested at the mid-log phase for all of the other five conditions. The proteins were extracted, fractionated, and then digested with trypsin. Digested samples were analyzed in duplicate by an automated 1D-LC−MS/MS technique employing multiple mass range scanning. The MS/MS spectra were identified by SEQUEST.[12] The output files from SEQUEST were organized by growth state and run number (all fractions from a single proteome analysis were combined) and then filtered by DTASelect.[13] All data is available at the *R. palustris* proteome study website (http://compbio.ornl.gov/rpal_proteome).

Table 1. Arrangement of the Spectral Count Data in a Two-Way Table

	condition 1	condition 2	total
spectral count for a target protein *X*	x_1	x_2	x
spectral count for any other protein	y_1	y_2	y
total spectral count across all proteins	n_1	n_2	n

Statistical Tests for Detecting Differential Expression in Pairwise Experiments. For the *S. cerevisiae* data set, five statistical tests on pairwise comparisons of different spiked concentrations were used. Those that did not require replicates included the *G*-test,[8] Fisher's exact test,[14] and AC test,[15] and those that needed replicates comprised the *t*-test[16] and LPE test.[17] To ease readability, the usage of these tests in our analysis is briefly described below. Detailed descriptions and formulas can be found in the Supporting Information.

(a) Statistical Tests that Do Not Require Replicates. Given a target protein (*X*), the spectral counts from a pairwise experiment are arranged in a two-way table (Table 1). For a target protein *X*, we denote the proportion of its spectral count as $r_1 = x_1/n_1$ and $r_2 = x_2/n_2$ under Condition 1 and Condition 2, respectively. The total spectral count for a specific condition defines the sampling number for this experiment. Then *X* is differentially expressed under the two conditions if the difference between these two proportions is statistically significant. If a statistical significance test does not require replicates, then the spectral count data is pooled to "mimic" replicated experiments. Specifically, the *G*-test[8] tests the null hypothesis that the observed frequencies result from random sampling from a distribution with the given expected frequencies. The distribution of the *G* statistic approximately follows a χ^2 distribution with one degree of freedom. This approximation can lead to higher false positive rates, which is adjusted by the William's correction (w).[8]

Fisher's exact test[14] assumes that the row totals and the column totals are fixed in the two-way table. Hence, any entry in the table completely determines the others. For the expected spectral counts in Fisher's exact test, a hypergeometric distribution is assumed.

The AC test[15] calculates the conditional probability of finding x_2 spectral counts in Condition 2 given x_1 spectral counts that have been observed in Condition 1.

(b) Statistical Tests that Require Replicates. If a statistical significance test does require replicates, then the spectral count data for replicates are kept separately. The data are globally normalized before using the test. Normalization is performed by dividing the protein spectral count in a particular experiment by the average spectral count across all the proteins in that experiment. This is done so that the global average count is the same across all experiments. For the *t*-test, the statistical difference between $Mean_1$ and $Mean_2$ is assessed, where $Mean_1$ and $Mean_2$ are average spectral counts across all replicates under Condition 1 and Condition 2, respectively. The LPE test[17] pools proteins with similar counts by percentile intervals and fits a smooth local regression curve to the variance estimates on the percentiles. The difference between the medians rather than averages is tested. The variances are estimated from the pooled values. The LPE test implemented in bioconductor (http://www.bioconductor.org) is used for this study.

Performance Metric for the Comparison of Statistical Tests. There are different ways to evaluate the performance of a statistical test (e.g., *t*-test, LPE test, or *G* test) to detect differentially expressed proteins with one or more replicates for various protein abundance levels. We focus on the false positive rate, namely the frequency of mistaking random fluctuations in sampling statistics for significant differences in the test result.[15]

Let us consider a data set with spiked protein markers as a benchmark set. Then a good statistical test should be able to highlight the true changes in a spiked protein marker while minimizing false identification of the nonspiked proteins. Suppose that *N* proteins are identified in a pairwise comparison, and *n* proteins are ranked equal to or more significant than a spiked protein marker by a statistical test of interest. Then, for the spiked protein, the false positive rate of this test can be formally defined as an n/N ratio.

Detecting Differential Expression in Multi-condition Experiments. In a multi-condition experiment with *m* conditions, a 2 × *m* table is generated for a protein of interest (*X*). The *G*-test for the 2 × 2 table is generalized for the 2 × *m* table as follows:

$$G = 2 \times \left(\sum_{i=1}^{m} x_i \ln x_i + \sum_{i=1}^{m} y_i \ln y_i - \sum_{i=1}^{m} n_i \ln n_i - x \ln x - y \ln y + n \ln n\right)$$

William's correction (w) can be similarly used for adjustment:

$$w = 1 + \frac{\left(\sum_{i=1}^{m} \frac{n}{n_i} - 1\right)\left(\frac{n}{x} + \frac{n}{y} - 1\right)}{6n(m-1)}$$

The adjusted *G* statistic is defined as: $G_{adj} = G/w$.

The corresponding *p*-value is calculated based on the assumption of a χ^2 distribution with $m - 1$ degrees of freedom. Because more than 1000 genes are tested simultaneously in our study, the Benjamini and Hochberg correction for multiple-testing adjustment[18] is applied. A protein with a *p*-value of *p* is then called differentially expressed if $p \le (j/t)q$, where *t* is the total number of proteins tested, *j* is the rank of this protein in a list of the *t* proteins sorted in increasing order by their *p*-value, and *q* is the desired false discovery rate (FDR, e.g., 0.01).

Sources of Functional Annotation and Operon Prediction for *R. Palustris*. For *R. palustris*, the following information was downloaded from the corresponding web-sites: (a) The predicted proteome (http://compbio.ornl.gov/rpal_proteome/databases/); (b) KEGG pathways (http://www.genome.ad.jp/kegg/pathway.html); (c) TIGR cellular role and sub-role (http://www.tigr.org/tigr-scripts/CMR2/gene_table.spl?db=ntrp02); and (d) Predicted operons (http://www.microbesonline.org).

Identification and Visualization of Correlated Protein Expression. The *hopach* package[19] in bioconductor (http://bioconductor.org) was used to identify clusters of coordinately expressed proteins. *Hopach* combines the strengths of both partitioning and agglomerative hierarchical clustering methods. It uses the Median Split Silhouette (MSS) criterion to prune the tree automatically and produces partitions of homogeneous clusters. The dissimilarity matrix was generated by the *distance matrix ( )* function using the parameter "cor" for the calculation. The default parameters of the *hopach* function were used for the clustering and automatic determination of the number of clusters. The clustering result was visualized in the Maple Tree program (http://mapletree.sourceforge.net/).

Functional Interpretation of Clustering Results. Each protein cluster was compared with the KEGG pathways and the

TIGR sub-role categories to identify functional categories with significantly enriched protein numbers using the hypergeometric test as described in Zhang et al.[20]

Results

Reproducibility of the Sampling Statistics between Technical Replicates. We used an *R. palustris* data set to investigate the reproducibility of sampling statistics (peptide count, spectral count, and sequence coverage) between technical replicates in label free LC−MS/MS. The data set included two technical replicates for each of the six growth conditions.[11] If a protein was found in only one replicate, we assigned the number zero to the sampling statistics for the other replicate.

Linear regression based on different sampling statistics was performed for each of the six duplicated experiment pairs in *R. palustris* data. The *R*-squared values from linear regression for different sampling statistics were averaged across the six experiment pairs. The average *R*-squared value for the spectral count, the peptide count, and the sequence coverage were 0.96, 0.94, and 0.73, respectively. Scatter plots and *R*-squared values for individual experiment pairs are available in the Supporting Information File S2. Thus, both the spectral count and peptide count had a high technical reproducibility. The spectral count was used in what follows.

Evaluation of Statistical Tests for a Pairwise Comparison of Experimental Conditions. For the *S. cerevisiae* data set with spiked proteins, the abundance of individual proteins can be considered consistent across the experiments with different spiked concentrations (i.e., 0.25, 1.25, and 2.5%). This is due to the relatively small proportion of spiked proteins (at a maximum level of 2.5%). In fact, the protein spectral counts were reproducible across all experiments of various spiked concentrations.[4] Consequently, the differences among the spectral count frequencies of the nonspiked proteins in experiments of various spiked concentrations were due to random fluctuations.

Using the *S. cerevisiae* data as a benchmark data set,[4] we compared the false positive rate for the Fisher's exact test, *G*-test, AC test, *t*-test, and LPE test. All spiked protein concentration pairs (0.25 vs 2.5%, 0.25 vs 1.25%, and 1.25 vs 2.5%) using data from one to three available replicates in each test were considered. Namely, comparison of the 0.25% experiments with the 2.5% experiments with only one replicate considered gave nine results for each test; each of the three replicates in one experiment was compared to each of the three replicates in the other. Similarly, we had nine (one) results for each test with only two (three) replicates considered. Data from replicated experiments were pooled for the Fisher's test, *G*-test, and AC test. The *t*-test and LPE test could only be used with two or three replicates.

Figure 1 shows the results of these comparisons using data from one to three replicates. Large differences (10-fold or 5-fold) in spiked protein concentrations were identifiable with a low false positive rate across all the tests. Specifically, the false positive rate was less than 0.4% (0.7%) for a 10-fold (5-fold) change detection even using data from a single replicate in the Fisher's exact test, *G* test, or AC test (an average of 15 452 total spectral counts per condition).

A 2-fold spiked protein abundance change was much more difficult to identify with a single replicate. A false positive rate greater than 10% was observed for the Fisher's test, *G*-test, and AC test. However, increasing the sampling number reduced the false positive rate. The false positive rate was reduced to less than 5% and 3% using pooled data from two replicates (an average of 30 904 total spectral counts per condition) and three replicates (an average of 46 356 total spectral counts per condition), respectively.

For 10- or 5-fold change detection with two or three replicates, the LPE test performed similarly to the aforementioned three tests. However, for 2-fold change identification, it performed the worst. When only two replicates were considered, the *t*-test performed the worst, but when three replicates were considered, the *t*-test outperformed all of the other tests for 2-fold change detection. However, it occasionally generated a small false positive rate (less than 1%) for a 10- or 5-fold change, while the other tests gave a false positive rate of zero.

In summary, the *t*-test is a better choice for differential proteomics studies with three or more replicates. However, when a study has less than three replicates, the Fisher's exact test, *G*-test, and AC test perform reasonably well. Moreover, the performance of these tests improves with increased sampling numbers.

Detection of Differential Expression Using Simultaneous Comparison of Multiple Conditions. For the multi-conditional comparison, we applied the generalized *G*-test to the proteome of *R. palustris* grown under six conditions: *Dark*, *Light*, *Benzoate*, *Auto*, *N2*, and *Stationary*. Among the 1622 identified proteins in the LC−MS/MS experiment, the generalized *G*-test detected 1119 proteins that were differentially expressed under the six conditions with a desired FDR (*q* value) of 0.01. The *p*-values and the corresponding significance calls based on the *q* values for all the proteins are available in the Supporting Information. For these 1119 differentially expressed proteins, the average spectral count for a protein under one growth condition is 54.1. Many of the nonsignificant proteins, however, are expressed at a low level. The average spectral count for the remaining 503 nonsignificant proteins is 7.1. For some of these proteins expressed at a low level, it might be difficult to make a significance call when using the generalized *G*-test.

For the pairwise comparison of the same *R. palustris* proteome performed in VerBerkmoes et al.,[11] the *Light* growth condition was used as a reference. Proteomes under other growth conditions (*Dark*, *Benzoate*, *Auto*, *N2*, and *Stationary*) were compared to the reference. Differentially expressed proteins were identified based on the following manually set rules: (1) a replicated difference of at least four peptides and/or 30% sequence coverage; and (2) a 2-fold difference in the spectral count level. All of the 236 differentially expressed proteins (http://compbio.ornl.gov/rpal_proteome/supplemental/excel/Table_S4.xls) found using these rules were also identified by the generalized *G*-test. Thus, compared to manual elucidation, the generalized *G*-test provides a practical method for sensitive detection of differential expression under multiple conditions.

Consistency between Protein Coexpression and Co-location in Predicted Operons. Proteins co-located in the same operon are more likely to be coexpressed than those from different operons. We tested the protein coexpression derived from the *R. palustris* quantification results against the 802 computationally predicted operons from the MicrobesOnline web site (http://www.microbesonline.org). The unordered protein pairs among the 1119 differentially expressed proteins were assigned to one of two groups: (1) both proteins in the same operon and (2) each protein in a different operon. The Pearson's correlation coefficient was used to measure protein

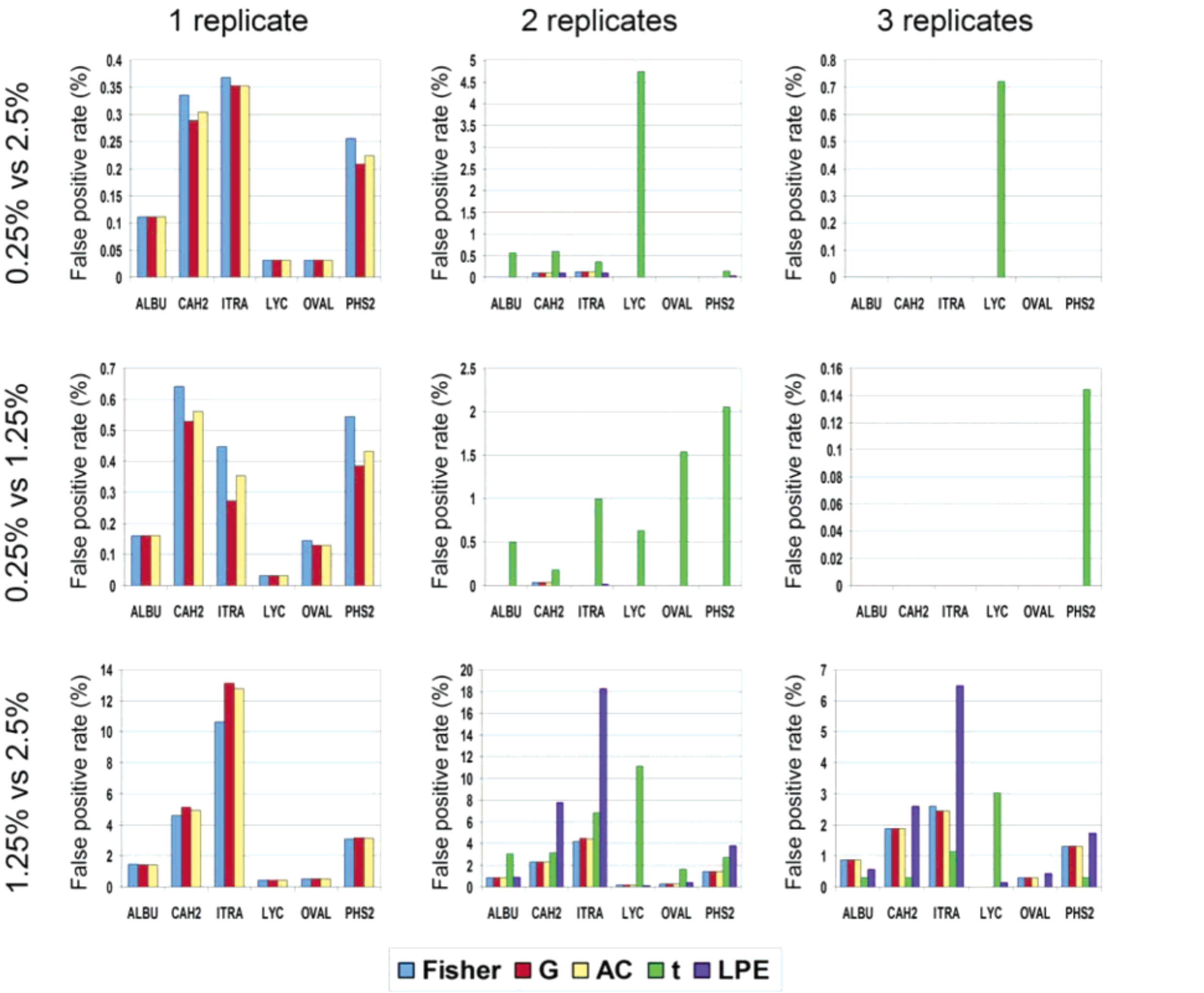

Figure 1. False positive rate generated by different statistical tests for a pairwise comparison of different spiked concentrations. Rows represent different levels of concentration changes for the protein markers. Columns represent the number of replicates used for the statistical tests. ALBU (bovine serum albumin), CAH2 (bovine carbonic anhydrase), ITRA (soybean trypsin inhibitor), LYC (chicken lysozyme), OVAL (chicken ovalbumin), and PHS2 rabbit phosphorylase b (PHS2) are the six protein markers. Fisher: Fisher's exact test; G: *G*-test; AC: AC test; t: *t*-test; LPE: local pooled error test.

coexpression levels for each pair. Protein pairs in each group were placed in bins at intervals of 0.01 based on their Pearson's correlation coefficients. For each bin, the proportion of protein pairs was calculated. Operon pairs were strongly coexpressed relative to the nonoperon pairs with a *p*-value of 2.52*E*-78 in the Kolmogorov-Smirnov test[21] (Figure 2). Thus, multi-condition protein expression data from label-free LC−MS/MS can reveal biologically meaningful coexpression.

Discovery of Biologically Significant Coexpressed Protein Groups with Cluster Analysis. We applied *hopach* clustering[19] to the 1119 differentially expressed proteins to identify correlated protein expression. To reduce noise and improve the quality of the clustering results, proteins with nonsignificant differential expression were excluded following a common strategy in cluster analysis.[22] The *hopach* algorithm clearly identified six clusters, which are visualized in Figure 3. The cluster membership of the 1119 proteins is available in the Supporting Information. As expected, these six clusters contained proteins that are predominantly expressed in one of the six growth conditions. To interpret the biological significance of these clusters, the proteins in each cluster were compared with those in the KEGG pathways and the TIGR role categories. Table 2 lists all functional categories for each cluster with a significantly enriched number of proteins ($p < 0.001$ in hypergeometric test[20]).

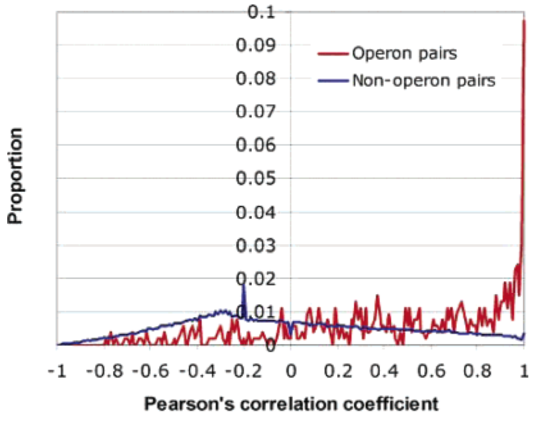

Figure 2. Protein coexpression consistent with protein colocation in predicted operons. Distribution of protein coexpression level (Pearson's correlation coefficient) for operon pairs (red) and nonoperon pairs (blue). Operon pairs are strongly coexpressed relative to the nonoperon pairs.

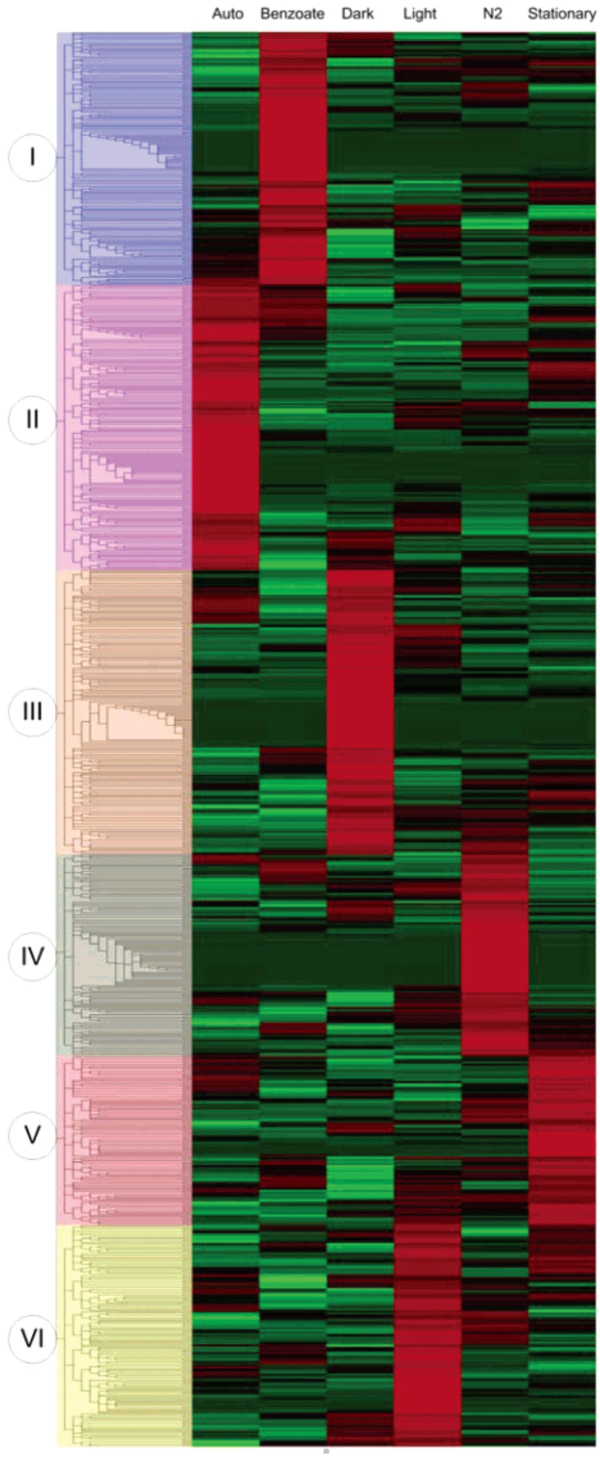

Figure 3. The six clusters of coexpressed proteins revealed by cluster analysis. The 1119 differentially expressed proteins in the *R. palustris* data set were clustered into six clusters (I–VI) by the hopach program. Individual proteins are represented by a single row, and each growth condition is represented by a single column. Each cell represents the expression level of a protein under one growth condition, relative to the mean expression level of the protein across all six conditions. Red represents overexpression, and green represents underexpression.

Some of the important functional categories identified in this study have already been reported in our previous manual elucidation.[11] For example, cluster I, dominated by proteins highly expressed in the *Benzoate* growth condition, is highly enriched in proteins from the category “Benzoate degradation”. Cluster IV, dominated by proteins highly expressed in the *N2* growth condition, is highly enriched in proteins from the category “Nitrogen fixation”. Actually, all of the nine nitrogen fixation proteins predicted by TIGR are found in this cluster. Although these correlations may appear obvious, they illustrate the ability of our approach to reveal the expected proteins, like a positive control in a computational experiment.

Our results not only confirmed the previous manual elucidation but also discovered new information. Proteins highly expressed in the *Auto* growth condition dominate cluster II. This cluster is significantly enriched in proteins from the category “chemotaxis and motility,” including RPA0139, RPA1627, RPA1640, RPA1676, RPA1884, RPA2479, RPA3185, RPA3546, RPA4302, RPA4638, RPA4639, and RPA4691. The *Auto* growth condition was harsh for the cells relative to the other conditions studied. The cells had to obtain energy from light and carbon from carbon dioxide. Expression of proteins in the category of “chemotaxis and motility” may help the cell detect and move toward the light and carbon dioxide sources, which provide a favorable environment for growth.[23]

Cluster III and cluster VI are dominated by proteins highly expressed in the *Dark* and *Light* growth conditions, respectively. We observed in the experiments that cells grew the best under these two conditions. Accordingly, the category “Ribosomal proteins: synthesis and modification” is enriched in the *Dark* condition, suggesting active protein synthesis. The categories “TCA cycle”, “Purine metabolism”, and “Alanine and asparate metabolism” are enriched in the *Light* condition, indicating active energy, nucleic acids, and protein metabolism in the cells grown under this condition.

Proteins highly expressed in the *Stationary* growth condition dominate cluster V. This cluster is significantly enriched in proteins from the category “protein folding and stabilization,” including RPA0306, RPA0331, RPA0333, RPA0334, RPA0511, RPA1198, RPA2164, RPA2889, and RPA4815. It has been reported in yeast that a large percent of genes involved in the maintenance of the stationary phase encode proteins involved in protein stabilization.[24] Our results from the *R. palustris* shotgun proteomics data seem to be consistent with this report.

Clustering results, such as those reported in this study, could help improve functional annotation of *hypothetical* proteins or proteins with *unknown function* using, for example, the “guilt-by-association” principle.[25] Specifically, in the predicted operon RPA3930-RPA3932, RPA3930 is a hypothetical protein, whereas RPA3931 and RPA3932 are both annotated in the functional category “chemotaxis and motility” based on TIGR annotation. Clustering of these three proteins in the same cluster (cluster III) in the *R. palustris* shotgun proteomics data set provides experimental evidence for the potential involvement of RPA3930 in “chemotaxis and motility.” We downloaded the protein sequence for RPA3930 from the Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene), and BLASTed against the UniProt Knowledgebase (http://www.expasy.org/tools/blast/). The RPA3930 protein has two HAP3 (putative flagellar hook-associated protein 3) domains, and showed sequence similarity to other flagellar hook-associated proteins. This supports the involvement of RPA3930 in “chemotaxis and motility”. In the operon RPA4611-RPA4620, RPA4613, and PRA4614 are two proteins with unknown function in TIGR, whereas the other proteins are annotated in the categories “nitrogen fixation” or “nitrogen metabolism”. Clustering of the two proteins in the same cluster (cluster IV) as the other operon

Table 2. Functional Categories with Significantly Enriched Protein Numbers for Each Cluster

ID[a]	size	predominant conditions	enriched KEGG categories[b]	enriched TIGR role categories[b]
I	200	benzoate	benzoate degradation via CoA ligation (21/47, $p = 2.06E{-}8$) carbon fixation (9/23, $p = 9.55E{-}4$)	
II	226	auto	bacterial chemotaxis (8/14, $p = 1.78E{-}4$)	chemotaxis and motility (12/29, $p = 2.34E{-}4$)
III	225	dark	ribosome (30/53, $p = 1.07E{-}13$)	ribosomal proteins: synthesis and modification (30/59, $p = 4.72E{-}12$)
IV	160	N2	nitrogen metabolism (7/15, $p = 2.6E{-}4$)	nitrogen fixation (9/9, $p = 7.19E{-}10$) nitrogen metabolism (6/12, $p = 4.7E{-}4$)
V	134	stationary		protein folding and stabilization (9/22, $p = 2.71E{-}5$)
VI	174	light	citrate cycle (TCA cycle) (8/20, $p = 6.07E{-}4$) purine metabolism (12/38, 3.54E−4) alanine and aspartate metabolism (6/12, 7.46E−4)	

[a] Cluster ID. [b] Each category is followed by the number of proteins in the cluster, the number of proteins identified in all growth conditions, and the *p* value.

members indicates the potential involvement of these two proteins in these biological processes. Through BLAST analysis, the RPA4614 protein showed high sequence similarity to nifK in *Synechococcus sp.*, which supports its role in "nitrogen fixation". Proteins with high sequence similarity to the RPA4613 protein are all hypothetical proteins. Although this sequence analysis did not confirm our inference, it suggests that functional annotation through "guilt-by-association" could be complementary to the sequence-based approach.

Discussion

Sampling Statistic as a Quantitative Measurement in Label-free LC−MS/MS Shotgun Proteomics. Peptide count, spectral count, and sequence coverage have all been used to quantify protein abundance. It was not clear which one was the most appropriate for protein quantification. Our results showed that both spectral count and peptide count are highly reproducible in technical replicates. In addition, in a previous study, spectral count showed a linear correlation with relative protein abundance; much higher than the correlation shown by peptide count.[4] A linear correlation between sequence coverage and the relative protein abundance was not observed.[4] On the basis of these results, we can conclude that spectral count is the most reproducible and the most strongly correlated with protein abundance among the three sampling statistics.

Detecting Differential Expression in Label-free LC−MS/MS Shotgun Proteomics. We have compared five statistical tests for detecting differential expression in a pairwise comparison of experimental conditions. Both the *t*-test and LPE test require replicates. We observed that the *t*-test works well with at least three replicates available. Although the Fisher's exact test, *G*-test, and AC test do not require replicates, the accuracy of these tests improves with the increased sampling numbers that result when pooling replicates. These statistical tests assume random sampling. Random sampling is obviously violated by the dynamic exclusion filtering commonly used in shotgun proteomics to maximize the number of identified proteins.

Despite this violation, these tests give reasonable false positive rates even with limited sampling numbers from a single replicate. This can likely be attributed to how these statistical tests are decided based on the spectral count for the protein under study and the spectral count for all the other proteins. Dynamic exclusion filtering has little effect on the second count, as the effect on individual proteins tends to average out. It has little effect on the counts for low abundance proteins, as the elution times for these proteins are very short. Obviously, dynamic exclusion filtering will decrease the spectral counts for abundant proteins. However, at high expression levels, as discussed below, these statistical tests require a very low fold-change for a difference to be called significant (Figure 4A). This might ensure the robustness of the test against dynamic exclusion filtering and explain the reasonable false positive rates.

For the Fisher's exact test, *G*-test, or AC test, a higher fold-change is required for a difference to be called significant at lower expression levels (Figure 4A). To examine the intensity-dependent variability of the spectral count data, we generated MA plots for the replicated experiment pairs from the six growth conditions in the *R. palustris* data set (Figure 4B). In an MA plot, the log intensity ratio (*M*) is plotted as a function of the average intensity (*A*). It is manifest from these plots that the variability of the spectral count is much greater at lower expression levels. This backs up the use of the Fisher's exact test, *G*-test, and AC test on the data. As we can see in Figure 4A, the Fisher's exact test is the most conservative, while the *G*-test is the least conservative. However, the performance of these three tests is very similar, which is consistent with the false positive rate results. Due to its computational simplicity, the *G*-test might be a better choice than the other two. Another advantage of the *G*-test is that it can be generalized and applied to multi-condition comparisons.

A recent publication claimed that 35% of proteins in a replicated MudPIT analysis are novel when compared to the first analysis.[26] Therefore, using the presence or absence of a protein in a particular run for biomarker discovery or differential display could lead to high false positive rates, as it may simply reflect analytical incompleteness. Using statistical analysis based on the quantitative information from the spectral count could reduce the false positive rates. For example, suppose that a protein is not detected in one condition out of two conditions with 10 000 spectral counts each. Then, for this protein to be called differentially expressed, the presence/absence approach will require one spectral count for the protein in the other condition. In contrast, Fisher's exact test will require seven spectral counts.

As both data sets include only technical replicates, applying the findings of this study to experiments with biological replicates should be made with caution. Biological variations are usually much higher than technical variations, and are likely

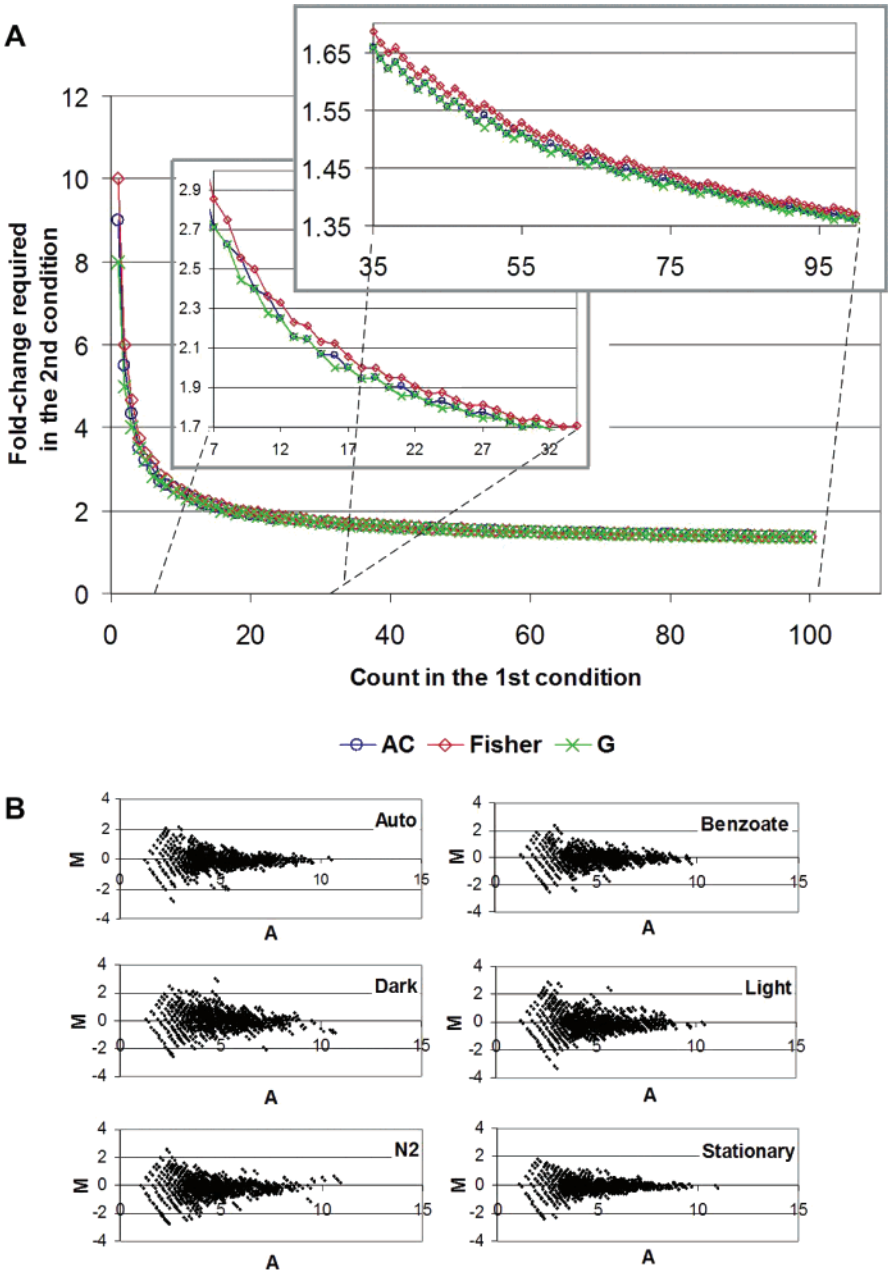

Figure 4. Differential expression detection and spectral count variability at various protein expression levels. (A) Fold changes required in the second condition for a difference to be called significant ($p < 0.01$) in the AC test (AC), Fisher's exact test (Fisher), and G-test (G), assuming 10 000 total spectral counts in the first condition. (b) MA plots showing log intensity ratio (M) as a function of average intensity for the replicated experiment pairs from the six growth conditions.

to change the reliability of the tests, such as the *t*-test, which assumes a normal data distribution.

Identifying Correlated Expression in Label-free Shotgun Proteomics. Methods for pairwise correlation and cluster analysis have been well developed in transcriptomic studies.[22] Nevertheless, attempts to study protein coexpression in label-free shotgun proteomics data are very limited. In this study, we have demonstrated a relationship between operon predictions and protein coexpression in the label-free shotgun proteomics data from *R. palustris*. Cluster analysis has revealed protein groups with functional significance. These results are promising and advocate the application of label-free shotgun proteomics to biological studies.

Through manual checking of the clustering results, we have also noticed some limitations of the clustering algorithm. For example, a protein can be assigned to only one cluster, which contradicts the fact that one protein can be involved in more than one biological process. As an example, one would expect the chaperonin GroEL2 (RPA2164) and its co-chaperonin GroES2 (RPA2165) to appear in the same cluster, as the genes are co-located in the same operon on the genome, and the proteins form the well-known multi-subunit GroELS protein complex,[27] which is important in assisting protein folding, and are responsive under various stress conditions.[28] But the clustering algorithm assigns RPA2164 and RPA2165 to different clusters. RPA2164 is grouped with proteins highly expressed in the *Stationary* growth condition (cluster V), whereas RPA2165 is grouped with proteins highly expressed in the *N2* growth condition (cluster IV). Both proteins showed similar expression patterns in the original data, with relatively high expression levels under both the *Stationary* and the *N2* conditions. RPA2164 had a slightly higher expression level under the *Stationary* condition, whereas RPA2165 had a slightly higher expression level under the *N2* condition. Previously, the induction of GroE homologues by the nutrient starvation associated with the onset of stationary phase has been observed.[29,30] The involvement of GroEL in *nif* gene regulation and nitrogenase assembly has also been reported.[31] These biological studies suggest the association of the GroELS protein complex with multiple biological processes. In such a case, a fuzzy clustering algorithm that allows multi-class memberships for a protein could have performed better in revealing the induction of the protein complex in multiple growth conditions. Several such clustering algorithms, used in microarray data analysis, could be applied, including fuzzy k-means clustering,[32] model-based clustering,[33] non-negative matrix factorization,[34] and clique-based clustering.[35] Alternatively, fine-tuning the parameters of the *hopach* program could have potentially produced different clusters with proteins that were responsive under multiple growth conditions.

Conclusion

Label-free LC−MS/MS shotgun proteomics was developed for protein profiling and has been proposed recently for quantitative studies. We have demonstrated the value of label-free shotgun proteomics in detecting differential and correlated protein expression using two published data sets. Among the sampling statistics that have been used for the quantitative evaluation of protein abundance, we were able to show that spectral and peptide counts are highly reproducible between technical replicates, and that sequence coverage is relatively nonreproducible. We evaluated different statistical tests for pairwise comparison of experimental conditions and showed that the Fisher's exact test, *G*-test, and AC test can be used when the number of replications is limited (one or two), whereas the *t*-test might be useful with three or more replicates available. The Fisher's exact test, *G*-test, and AC test performed similarly in our evaluation, thereby giving the *G*-test some advantage due to its computational simplicity. Moreover, we generalized the *G*-test to detect differential expression under multiple conditions considered simultaneously. Using the spectral count as a measure of protein abundance and the Pearson's correlation coefficient as a measure of protein coexpression level, we demonstrated that operon pairs are strongly coexpressed relative to nonoperon pairs. Cluster analysis combined with protein function and metabolic pathway annotation revealed groups of biologically significant coexpressed proteins from the *R. palustris* proteomics data set.

Acknowledgment. We are grateful to the anonymous reviewers for their insightful comments that helped us improve the paper. We thank Christopher Symons for insightful suggestions on how to improve the clarity of the paper. The computational and experimental work was supported by the DOE Genomics: GTL grants DOE LAB-04-32 and DOE DE-FG02-01ER63241, respectively, from the U.S. Department of Energy (Office of Biological and Environmental Research, Office of Science). The work of N.F.S. was partially funded by a Laboratory Directed Research and Development Program of Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by the University of Tennessee-Battelle, L.L.C. for the Department of Energy under contract DOE-AC05-00OR22725.

Supporting Information Available: Supporting Information File S1 describes in detail the statistical tests used for pairwise experiments. Supporting Information File S2 shows scatter plots and linear regressions for different sampling statistics in the replicated experiments from six growth conditions. Supporting Information File S3 lists differential expression analysis results for all 1622 proteins in the *R. palustris* data set. Supporting Information File S4 lists cluster membership for the 1119 differentially expressed proteins in the *R. palustris* data set. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198−207.
(2) Lin, D.; Tabb, D. L.; Yates, J. R., 3rd, Large-scale protein identification using mass spectrometry. *Biochim. Biophys. Acta* **2003**, *1646* (1−2), 1−10.
(3) Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J. A proteomic view of the Plasmodium falciparum life cycle. *Nature* **2002**, *419* (6906), 520−6.
(4) Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **2004**, *76* (14), 4193−201.
(5) Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G. Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics. *Mol. Cell Proteomics* **2005**, *4* (10), 1487−502.
(6) Gao, J.; Opiteck, G. J.; Friedrichs, M. S.; Dongre, A. R.; Hefta, S. A. Changes in the protein expression of yeast as a function of carbon source. *J. Proteome Res.* **2003**, *2* (6), 643−9.
(7) Colinge, J.; Chiappe, D.; Lagache, S.; Moniatte, M.; Bougueleret, L. Differential Proteomics via probabilistic peptide identification scores. *Anal. Chem.* **2005**, *77* (2), 596−606.
(8) Sokal, R. R.; Rohlf, F. J. *Biometry*, 3rd ed.; W. H. Freeman and Company: New York, 1995.
(9) Dudoit, S.; Shaffer, J. P.; Boldrick, J. C. Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **2003**, *18* (1), 71−103.
(10) Romijn, E. P.; Christis, C.; Wieffer, M.; Gouw, J. W.; Fullaondo, A.; van der Sluijs, P.; Braakman, I.; Heck, A. J. Expression Clustering Reveals Detailed Co-expression Patterns of Functionally Related Proteins during B Cell Differentiation: A Proteomic Study Using a Combination of One-Dimensional Gel Electrophoresis, LC-MS/MS, and Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC). *Mol. Cell Proteomics* **2005**, *4* (9), 1297−310.
(11) VerBerkmoes, N. C.; Shah, M. B.; Lankford, P. K.; Pelletier, D. A.; Strader, M. B.; Tabb, D. L.; McDonald, W. H.; Barton, J. W.; Hurst, G. B.; Hauser, L.; Davison, B. H.; Beatty, J. T.; Harwood, C. S.; Tabita, F. R.; Hettich, R. L.; Larimer, F. W. Determination and comparison of the baseline proteomes of the versatile microbe Rhodopseudomonas palustris under its major metabolic states. *J. Proteome Res.* **2006**, *5* (2), 287−98.
(12) Eng, J. K.; McCormack, A. L.; Yates, J. R., 3rd. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.
(13) Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002**, *1* (1), 21−6.
(14) Fisher, R. A. *Statistical Methods for Research Workers*; Oliver & Boyd: Edinburgh, 1925.
(15) Audic, S.; Claverie, J. M. The significance of digital gene expression profiles. *Genome Res.* **1997**, *7* (10), 986−95.
(16) Student, On the error of counting with a haemacytometer. *Biometrika* **1907**, *5*, 351−360.
(17) Jain, N.; Thatte, J.; Braciale, T.; Ley, K.; O'Connell, M.; Lee, J. K. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **2003**, *19* (15), 1945−51.

(18) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **1995**, *57* (1).

(19) van der Laan, M.; Pollard, K. A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *J. Statist. Plan Infer.* **2003**, *117*, 275−303.

(20) Zhang, B.; Kirov, S.; Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W741−8.

(21) Sheskin, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures*, 3rd ed.; Chapman & Hall/CRC: Boca Raton, FL, 2003.

(22) Shannon, W.; Culverhouse, R.; Duncan, J. Analyzing microarray data using cluster analysis. *Pharmacogenomics* **2003**, *4* (1), 41−52.

(23) Neidhardt, F. C.; Ingraham, J. L.; Schaechter, M. *Physiology of the Bacterial Cell: A Molecular Approach*; Sinauer Associates, Inc.: Sunderland, MA, 1990.

(24) Werner-Washburne, M.; Braun, E.; Johnston, G. C.; Singer, R. A. Stationary phase in the yeast *Saccharomyces cerevisiae*. *Microbiol. Rev.* **1993**, *57* (2), 383−401.

(25) Wolfe, C. J.; Kohane, I. S.; Butte, A. J. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics* **2005**, *6*, 227.

(26) Wilkins, M. R.; Appel, R. D.; Van Eyk, J. E.; Chung, M. C.; Gorg, A.; Hecker, M.; Huber, L. A.; Langen, H.; Link, A. J.; Paik, Y. K.; Patterson, S. D.; Pennington, S. R.; Rabilloud, T.; Simpson, R. J.; Weiss, W.; Dunn, M. J. Guidelines for the next 10 years of proteomics. *Proteomics* **2006**, *6* (1), 4−8.

(27) Xu, Z.; Horwich, A. L.; Sigler, P. B. The crystal structure of the asymmetric GroEL-GroES−(ADP)7 chaperonin complex. *Nature* **1997**, *388* (6644), 741−50.

(28) Segal, R.; Ron, E. Z. Regulation and organization of the groE and dnaK operons in Eubacteria. *FEMS Microbiol. Lett.* **1996**, *138* (1), 1−10.

(29) Parsons, L. M.; Limberger, R. J.; Shayegani, M. Alterations in levels of DnaK and GroEL result in diminished survival and adherence of stressed Haemophilus ducreyi. *Infect. Immun.* **1997**, *65* (6), 2413−9.

(30) Siegele, D. A.; Kolter, R. Life after log. *J. Bacteriol.* **1992**, *174* (2), 345−8.

(31) Govezensky, D.; Greener, T.; Segal, G.; Zamir, A. Involvement of GroEL in nif gene regulation and nitrogenase assembly. *J. Bacteriol.* **1991**, *173* (20), 6339−46.

(32) Gasch, A. P.; Eisen, M. B. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome. Biol.* **2002**, *3* (11), RESEARCH0059.

(33) Yeung, K. Y.; Fraley, C.; Murua, A.; Raftery, A. E.; Ruzzo, W. L. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **2001**, *17* (10), 977−87.

(34) Brunet, J. P.; Tamayo, P.; Golub, T. R.; Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (12), 4164−9.

(35) Voy, B. H.; Scharff, J. A.; Perkins, A. D.; Saxon, A. M.; Borate, B.; Chesler, E. J.; Branstetter, L. K.; Langston, M. A. Extracting gene networks for low dose radiation using graph theoretical algorithms. *PLoS Compt. Biol.* **2006**, *2* (7), e89.

PR0600273