BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 11 2009, pages 1397–1403
doi:10.1093/bioinformatics/btp168

Data and text mining

Integrating shotgun proteomics and mRNA expression data to improve protein identification

Smriti R. Ramakrishnan[1,†], Christine Vogel[2,†], John T. Prince[2], Zhihua Li[2], Luiz O. Penalva[3], Margaret Myers[1], Edward M. Marcotte[2,*], Daniel P. Miranker[1,*] and Rong Wang[4]

[1]Department of Computer Sciences, 1 University Station C0500, [2]Department of Chemistry and Biochemistry, Institute for Cellular and Molecular Biology, Center for Systems and Synthetic Biology, 2500 Speedway, The University of Texas at Austin, Austin, TX 78712, [3]Children's Cancer Research Institute, The University of Texas Health Science Center at San Antonio, San Antonio, TX 78229 and [4]Pathogen Functional Genomics Resource Center, J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA

Received on December 24, 2008; revised on February 19, 2009; accepted on March 18, 2009
Advance Access publication March 24, 2009
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Tandem mass spectrometry (MS/MS) offers fast and reliable characterization of complex protein mixtures, but suffers from low sensitivity in protein identification. In a typical shotgun proteomics experiment, it is assumed that all proteins are equally likely to be present. However, there is often other information available, e.g. the probability of a protein's presence is likely to correlate with its mRNA concentration.

Results: We develop a Bayesian score that estimates the posterior probability of a protein's presence in the sample given its identification in an MS/MS experiment and its mRNA concentration measured under similar experimental conditions. Our method, MSpresso, substantially increases the number of proteins identified in an MS/MS experiment at the same error rate, e.g. in yeast, MSpresso increases the number of proteins identified by ~40%. We apply MSpresso to data from different MS/MS instruments, experimental conditions and organisms (*Escherichia coli*, human), and predict 19–63% more proteins across the different datasets. MSpresso demonstrates that incorporating prior knowledge of protein presence into shotgun proteomics experiments can substantially improve protein identification scores.

Availability and Implementation: Software is available upon request from the authors. Mass spectrometry datasets and supplementary information are available from http://www.marcottelab.org/MSpresso/.

Contact: marcotte@icmb.utexas.edu; miranker@cs.utexas.edu

Supplementary Information: Supplementary data website: http://www.marcottelab.org/MSpresso/.

1 INTRODUCTION

The measurement of all mRNA and protein expression levels in organisms is a fundamental biological goal. Though mRNA expression levels are now routinely measured on large scale, methods of high-throughput protein identification like western blotting, 2D gel electrophoresis and green-fluorescent protein (GFP) fusion tagging are very expensive in labor, time and resources. Mass spectrometry (MS) based shotgun proteomics is a simple alternative to these methods. With sensitive tandem mass spectrometry (MS/MS) instruments or extensive biochemical fractionation, several thousand proteins can be identified (Brunner *et al*., 2007; Graumann *et al*., 2007; Peng *et al*., 2003; Washburn *et al*., 2001). However, less costly approaches only identify a few hundred proteins in a complex protein sample.

A shotgun proteomics experiment typically proceeds by MS/MS analysis of peptides from proteolytically digested proteins, followed by *in silico* matching of the MS/MS spectra against a database of theoretical peptide spectra derived from protein sequences (Fig. 1). Proteins are identified from combined evidence for their composite peptides, resulting in a list in which each protein is associated with a confidence score of correct identification. We refer to this score as the 'original', 'primary' or 'raw' protein identification score, e.g. here using ProteinProphet (Nesvizhskii *et al*., 2003). All proteins with scores greater than a chosen threshold are labeled 'present' (Fig. 1).

Protein identification in an MS/MS experiment is hindered by a number of factors: noisy spectra, low-concentration proteins, post-translational modifications and chemical properties that interfere with efficient peptide ionization. For complex samples such as cell lysates, current MS search algorithms typically match a disproportionately small percentage (<20%) of all MS/MS spectra to peptides in a database, and only a small fraction of the expected proteins is identified. In other words, despite their presence in the biological sample, raw MS/MS identification scores of many proteins fall below a given confidence threshold and the proteins are incorrectly labeled as 'not present'.

The vast majority of MS/MS experiments are analyzed without considering any prior information regarding a protein's presence in the sample. MS/MS protein identification scoring schemes, e.g. BioWorks (ThermoFinnigan) or ProteinProphet (Nesvizhskii *et al*., 2003), assume that all proteins are equally likely to be present. In reality, other information may be readily available and can be used to

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

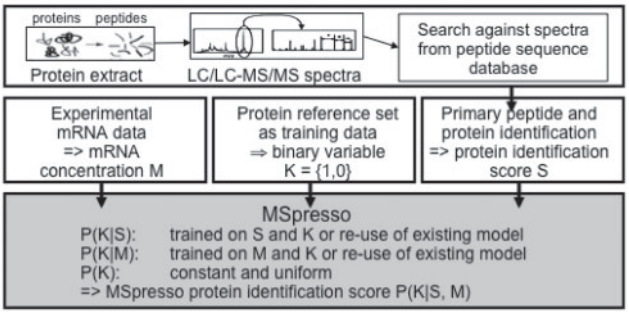

Fig. 1. Boosting protein identifications with prior information on mRNA concentration. A complex protein sample, e.g. cellular extract, is enzymatically digested into peptides and subjected to MS/MS. Raw MS/MS spectra are searched against a database of sequences using primary protein identification software, e.g. Bioworks (ThermoFinnigan), PeptideProphet (Keller *et al*., 2002) and ProteinProphet (Nesvizhskii *et al*., 2003), which produces a list of proteins and scores that signify the probability of correct identification. In a secondary analysis, MSpresso reexamines protein identification scores with respect to their mRNA abundance. MSpresso boosts the protein identification score given sufficient mRNA concentration. Proteins are then labeled 'present' if their MSpresso probability is larger than a newly determined cutoff. The MSpresso score, $P(K=1|S, M)$, estimates the probability of protein presence as the posterior probability of $K=1$ given mRNA abundance M and MS protein identification score S. MSpresso uses three probabilities, $P(K|S)$, $P(K)$ and $P(K|M)$, and their estimation is discussed in the text. K, protein presence; S, MS/MS identification score; M, mRNA concentration.

influence the inferred probability of protein presence when evidence from the MS/MS experiment is weak.

Our method, MSpresso (for MS and exPRESSion data), integrates data from MS/MS experiments with mRNA expression data in a Bayesian framework. MSpresso computes a new protein identification score as the posterior probability of the protein being present in the sample given both its MS/MS and mRNA scores.

We demonstrate the applicability of MSpresso on a yeast sample grown in rich medium analyzed on a low-resolution mass spectrometer (LCQ). We use mRNA concentrations from three independent experiments (Holstege *et al*., 1998; Velculescu *et al*., 1995; Wang *et al*., 2002) and corresponding protein data from four MS experiments (Chi *et al*., 2007; de Godoy *et al*., 2006; Peng *et al*., 2003; Washburn *et al*., 2001). We compare the performance of MSpresso on the yeast sample with the original raw MS/MS identification scores using ROC (Receiver Operator Characteristic) plots, and find an increase of ∼40% in the number of proteins identified at a fixed error rate. We validate 98% of these new identifications by their presence in at least one of the seven independent benchmarking datasets. We also generalize the method and demonstrate its applicability to a data from a high-resolution MS/MS instrument, different biological conditions, as well as to other organisms (*Escherichia coli*, human). To the best of our knowledge, MSpresso is the first integrative approach to analysis of shotgun proteomics data.

2 METHODS

2.1 MSpresso uses a Bayesian probability framework

Primary protein identification is an essential step in the procedure to derive an initial list of proteins and their MS/MS identification scores. The process is outlined in Figure 1. Starting from the MS/MS analysis of a complex protein sample, we used Bioworks (ThermoFinnigan), PeptideProphet (Keller *et al*., 2002) and ProteinProphet (Nesvizhskii *et al*., 2003) to derive an initial list of proteins and their corresponding protein identification scores. We do not require that the raw MS/MS identification score be a probability, though this is the case with ProteinProphet (Nesvizhskii *et al*., 2003). Any other MS analysis software is equally suitable for primary protein identification. MSpresso also does not affect peptide identifications as it only uses identifications at the protein level. MSpresso results do not reflect on the quality of the primary protein identification, they merely produce a new score based on additional information.

The MSpresso protein identification probability combines both direct and inferential evidence of protein presence. Direct evidence is generated by methods that directly measure protein presence e.g. MS/MS analysis. Inferential evidence refers to data that implies protein presence but does not directly measure it, e.g. mRNA abundance.

More formally, K is a Bernoulli variable where $K=1$ is the event that the protein is present in the sample, and $P(K=1)$ is the probability of that event. The MSpresso probability is the posterior probability $P(K|S=s, M=m)$ that a protein is present in the sample given its associated mRNA abundance $M=m$, and its raw MS/MS protein identification score $S=s$. Using Bayes' law,

$$\begin{aligned} P(K|S,M) &\propto P(K)P(S|K)P(M|K) \\ &\propto P(K)\left(\frac{P(K|S)P(S)}{P(K)}\right)\left(\frac{P(K|M)P(M)}{P(K)}\right) \\ &\propto \frac{1}{P(K)}\left(P(K|S)P(S)\right)\left(P(K|M)P(M)\right) \end{aligned} \quad (1)$$

Using a conditional independence assumption between M and S given K, we set $P(M|K, S)=P(M|K)$. In other words, we assume that observed mRNA abundance M is independent of S given that the protein is present (see Supplementary Material for discussion and detailed derivation). $P(K|M)$ and $P(K|S)$ are the posterior probabilities of a protein existing in the sample, given only its mRNA abundance M and primary identification score S, respectively. $P(K)$ is the prior probability of the protein being present. Rewriting and normalizing, we obtain the MSpresso score:

$$P(K|S,M)=\frac{P(K|S)P(K|M)/P(K)}{\sum_{K=0,1} P(K|S)P(K|M)/P(K)} \quad (2)$$

2.2 Evaluation methodology

To compare primary and MSpresso protein identifications, we estimated true positive rate (TPR), false positive rate (FPR) and precision, given the proteins known to be in the sample (positive instances) and known not to be in the sample (negative instances). TPR at score threshold t is the fraction of positive instances with scores $\geq t$. FPR at score threshold t is the fraction of negative instances with scores $\geq t$. False negative rate at score t is the fraction of all positive instances with score $< t$, and is equal to (1 – TPR). Precision at score threshold t is the fraction of all identifications with scores $\geq t$ that are positive instances. Note that FPRs are computed differently from false discovery rates (FDR = 1 – precision) (Nesvizhskii *et al*., 2003), and the number of reported proteins can vary depending on the error model. Supplementary Section 2.5 has a discussion on different error estimates.

We evaluated results using 'protein reference datasets' of large-scale protein identification. Reference sets act as an empirical estimate of the ground truth of proteins truly present in the sample. Proteins present in the reference set were labeled as positive instances. The reference sets were used as training data to estimate the probabilities in Equation (2), and in evaluation to validate the reported proteins and generate ROC plots. Since we typically used the same set for training and evaluation, probability estimates were averaged across 10 runs of 10-fold cross-validation, using a different fold partitioning per run. We constructed protein reference sets by gathering high-confidence protein identifications from published large-scale experiments

run on similar sample conditions. When such data was unavailable, we generated a reference set by pooling high-confidence protein identifications from several technical replicates of our MS/MS experiment e.g. human data (Section 2.3.4). Reference sets for each experiment are given in Section 2.3. In Section 3.4.2, we discuss evaluation without protein reference sets, where we used decoy proteins to represent negative instances.

In each experiment, we generated MSpresso scores for all proteins in the test set of proteins with non-zero primary identification score and mRNA abundance. Since the samples were cytosolic, we expected a bias in sample composition against membrane proteins and therefore excluded all proteins predicted to have one or more membrane helices (Kall *et al*., 2004). In rich-medium yeast, this resulted in a test set of 4165 genes and 3443 proteins in the mRNA and protein reference datasets, respectively. All numbers quoted in Section 3 refer to proteins without membrane helices; analysis of the full proteome gives similar results (Supplementary Fig. S7C).

2.3 Datasets and sample preparation

We applied MSpresso to a variety of organisms, experimental conditions and mass spectrometers: yeast grown in rich and minimal medium, *E. coli* grown in minimal medium and a human cell line. We analyzed each dataset on one or two different mass spectrometers: a ThermoFinnigan Surveyer/DecaXP+ (LCQ) or ThermoFinnigan LTQ-OrbiTrap (ORBI). MS/MS protein identification was conducted using the Bioworks 3.3. (ThermoFinnigan), PeptideProphet (Keller *et al*., 2002) and ProteinProphet (Nesvizhskii *et al*., 2003) pipeline. Details for each dataset are given below, with more details in the Supplementary Material (Section 1).

2.3.1 Yeast (rich medium) Cell lysate from wild-type yeast grown in rich medium was analyzed on both the LCQ and ORBI mass spectrometers. The LCQ data has been published before (Lu *et al*., 2007). For the fractionation data, cellular lysate was separated in a 7–47% sucrose gradient and fractions were monitored by UV for RNA content. We chose the fraction containing 80S ribosomes for further liquid chromatography (LC) MS/MS analysis on the LCQ.

The mRNA data is the average of at least two of three independent absolute expression measurements, all derived from wild-type yeast grown to log-phase in rich medium (Holstege *et al*., 1998; Velculescu *et al*., 1995; Wang *et al*., 2002). The same dataset was also used in Lu *et al*. (2007). The protein reference set was generated from a pool of four independent MS-based protein datasets from yeast grown in rich medium (Chi *et al*., 2007; de Godoy *et al*., 2006; Peng *et al*., 2003; Washburn *et al*., 2001), choosing proteins present in at least two of the four datasets ('YP4gte2'). Proteins that were absent from all four datasets represent the negative instances, i.e. we assume these proteins are not expressed under the given conditions. Unless stated otherwise, all rich-medium yeast results in this article use this reference dataset.

We also compiled a non-MS-based protein reference dataset (YP3, Supplementary Section 1.2, Fig. S7D). For the fractionation data, we assembled a list of ribosomal proteins as published in literature (Planta and Mager, 1998) and proteins annotated as ribosomal, involved in ribosome biogenesis or translation (Nash *et al*., 2007).

2.3.2 Yeast (minimal medium) MS/MS data on wild-type yeast grown in minimal medium (MOPS9) was derived from published work (Lu *et al*., 2007), with cell lysate analyzed on an LCQ mass spectrometer. The mRNA abundance was obtained from one dataset for yeast grown in minimal medium (YMD) (Smirnova *et al*., 2005). Protein reference data comprised of published flow-cytometry analysis of GFP-labeled proteins [2214 proteins (Newman *et al*., 2006), 1792 are non-membrane and have detectable mRNA abundances], combined with two MS-based datasets for a total of 2529 proteins identified at high confidence (2022 non-membrane and with mRNA abundances). See Supplementary Section 1.2 for details.

2.3.3 Escherichia coli (minimal medium) We performed shotgun MS/MS analysis on trypsinized, soluble proteins extract from *E.coli* grown in minimal medium using the LTQ-OrbiTrap mass spectrometer [details in Supplementary Material and Lu *et al*. (2007)]. Three datasets provided information on mRNA concentration (Allen *et al*., 2003; Corbin *et al*., 2003; Covert *et al*., 2004). Reference data comprised of two published 2D-gel electrophoresis datasets (Link *et al*., 1997; Lopez-Campistrous *et al*., 2005) for a total of 370 non-membrane proteins that also had detectable mRNA abundances.

2.3.4 Human We analyzed two human datasets generated by MS/MS analysis on two mass spectrometers (LCQ, LTQ-OrbiTrap). Experimental preparation of human data from the Daoy medulloblastoma cell line is described in the Supplementary Section 1.4. As no matching published large-scale human proteomics dataset was available for use as a reference set, we generated one by combining high-confidence protein identifications ($\leq$5% FDR defined by ProteinProphet) from 10 technical replicates (injections) of MS/MS analysis on the LTQ-OrbiTrap mass spectrometer. We used this as a reference set for the LCQ dataset. For the LTQ-OrbiTrap dataset, we pooled nine replicates into a reference set, and used the 10th replicate as the test set.

2.3.5 Functional analysis of reported proteins Functional analysis of yeast proteins was conducted with saccharomyces genome database (SGD) (Nash *et al*., 2007), FunSpec (Robinson *et al*., 2002) and FuncAssociate (Berriz *et al*., 2003), applying Bonferroni corrections for multiple hypothesis testing. There was no bias towards phosphorylated proteins among MSpresso identifications (Chi *et al*., 2007; Ptacek *et al*., 2005). Functional analysis of *E.coli* proteins was conducted using annotations from GenProtEC (Serres *et al*., 2004).

2.3.6 Data availability Yeast LCQ and *E.coli* data have been published (Lu *et al*., 2007). Other MS/MS data is available at http://www.marcottelab.org/MSpresso/.

3 RESULTS

We first present results on a rich-medium yeast sample, followed by results on *E.coli* and human datasets. We then describe generalizations of the method that apply in the absence of high-quality training data. Finally, we discuss evaluation without reference sets (decoy databases).

3.1 Knowledge of mRNA levels can improve identification of the expressed yeast proteome

We show that incorporating prior evidence of protein presence into the protein identification score can significantly increase the probability of correct protein identification in MS/MS experiments.

To compute the MSpresso score in Equation (2), we must estimate three probabilities (Fig. 1): (i) the prior probability $P(K=1)$ of protein presence in the sample; (ii) the posterior probability $P(K|S)$ of protein presence given only its primary MS/MS identification score S; and (iii) the posterior probability $P(K|M)$ of protein presence given only its mRNA concentration M.

We assume $P(K)$ follows the uniform distribution, and set the probability $P(K=1)=$constant for all proteins. Under this model, $P(K)$ acts as a proportionality constant which does not change the ranking of MSpresso scores, but just their value. For the expressed yeast proteome, we estimate $P(K)=2/3$, as suggested by the size of a reference dataset (Chi *et al*., 2007; de Godoy *et al*., 2006; Peng *et al*., 2003; Washburn *et al*., 2001).

The posterior probability $P(K|S)$ is learned using a logistic regression classifier on the primary identification score [from

ProteinProphet (Nesvizhskii *et al*., 2003)]. The posterior probability $P(K|M)$ is estimated by binning experimentally determined mRNA concentrations (Holstege *et al*., 1998; Velculescu *et al*., 1995; Wang *et al*., 2002). We used the protein reference set YP4gte2 described in Section 2.3.1 as ground truth for training and evaluation. We generated a histogram of mRNA abundances (log-scale) and set $P(K|M)$ to the fraction of proteins present in the protein reference dataset per bin (Fig. 2). We chose bin width to maximize the area under ROC curve (AUC) using cross-validation. AUC was not very sensitive to bin size variation (data not shown).

In general, we expect proteins with high levels of mRNA expression to have a better chance of being present in a proteomics experiment. Indeed, we find that the probability of protein presence in the reference dataset, $P(K=1|M)$, increases with increasing mRNA concentration (Fig. 2). Note that the relationship in Figure 2 refers to the relationship between mRNA abundance M and protein presence K, which is different from the relationship between protein abundance and mRNA abundance that has been studied elsewhere (Futcher *et al*., 1999; Greenbaum *et al*., 2003; Gygi *et al*., 1999; Lu *et al*., 2007). Figure 2 resembles a step function with linear interpolation between steps: below a (log-scale) concentration of ~0.5 mRNA molecules/cell the probability of the protein being present in the reference set is low ($P(K=1|M) \leq 0.10$), while above nine molecules/cell the probability is high [$P(K=1) \geq 0.90$]. The step function is conserved for yeast grown in minimal medium, *E. coli* and human (Supplementary Fig. S2).

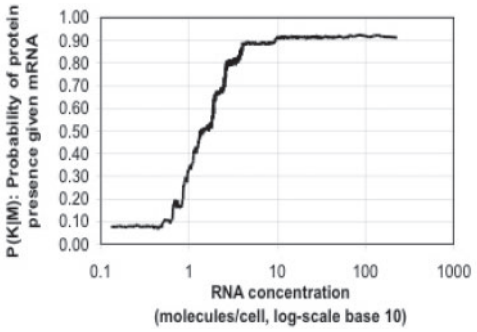

Fig. 2. Experimental data describes the relationship between the probability of protein presence given that the corresponding mRNA is observed at a certain abundance, $P(K|M=m)$. The relationship is modeled by a histogram of the fraction of proteins present in the protein reference set per bin of mRNA concentration, generated from a rank ordered list of mRNA abundances using 225 proteins per mRNA bin. The protein reference dataset contains four MS-based proteomics datasets (Chi *et al*., 2007; de Godoy *et al*., 2006; Peng *et al*., 2003; Washburn *et al*., 2001); the mRNA data is an average of three datasets (Holstege *et al*., 1998; Velculescu *et al*., 1995; Wang *et al*., 2002).

3.2 Results on rich-medium yeast sample

3.2.1 MSpresso identifies up to 63% more proteins than the primary identification Using Equation (2), we calculated the MSpresso protein identification score for each protein in the rich-medium yeast LCQ dataset. A protein that is present in the YP4gte2 reference set (Section 2.3.1) is labeled as true identification (positive instance), and a protein that is absent from it is labeled as a false identification (negative instance). We report the MSpresso score for each protein and a 5% FPR cutoff over all identified proteins.

Table 1 summarizes the results at 5% FPR for yeast and other experiments. MSpresso identifies more proteins at the same error rate than the primary identification.

Figure 3A and B illustrate performance via ROC (TPR versus FPR) and precision–recall plots (TPR versus precision). The MSpresso ROC curve dominates the primary identification curve at a wide range of FPRs. This observation implies that using MSpresso scores is better than just lowering the primary score threshold (choosing a threshold with higher FPR) to obtain more predictions.

Table 1. MSpresso performance in different experiments

Experiment	Test set	Area under the ROC (AUC)			Number of proteins identified at 5% FPR		
		MS/MS	MSpresso	Percentage increase	MS/MS	MSpresso	Percentage increase
Yeast-YPD-LCQ	Cell lysate, rich medium (YPD), LCQ (five injections)	0.75	0.89	19	234	327	40
Yeast-YPD-ORBI	Cell lysate, rich medium (YPD), ORBI (eight injections)	0.80	0.84	5	428[a]	618	63
Yeast-YMD-LCQ	Cell lysate, minimal medium (YMD), LCQ (six injections)	0.73	0.84	15	229	278	21
Yeast-Fraction-LCQ	Cell lysate, fractionated in polysomal gradient, rich medium (YPD), LCQ (three injections)	0.72	0.77	7	21[a]	34	62
Escherichia coli-ORBI	Cell lysate, minimal medium (MOPS9), ORBI (three injections)	0.69	0.80	16	63[a]	87	38
Human-LCQ	Cell lysate from Daoy, LCQ (two injections)	0.71	0.75	6	99	121	22
Human-ORBI	Cell lysate from Daoy, ORBI (one injection)	0.79	0.81	3	105	125	19

In each experiment, we generated MSpresso scores for each protein with observed mRNA abundance and MS/MS identification score. The better the MSpresso-based scoring, the higher the 'Percentage AUC increase' and 'Percentage increase in number of identified proteins'. These experiments use the 'self' MSpresso model: trained and evaluated on experiment-specific reference data. MSpresso results using the 'reuse' model are presented in the Supplementary Material (Table S10).

[a]Data as extrapolated from the ROC curve where there was no data at 5% FPR.

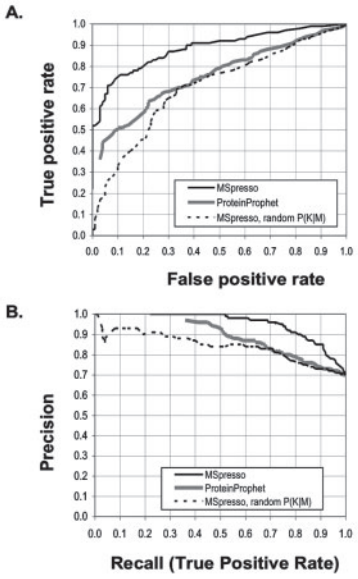

Fig. 3. Performance of MSpresso in yeast grown in rich medium. We evaluate the performance of MSpresso, the original MS/MS identifications (ProteinProphet), and MSpresso using a random $P(K|M)$ model using a ground-truth reference set to determine true and false identifications. (**A**) ROC plot (TPR versus FPR): MSpresso identifies more true positives at a given FPR than the MS/MS identifications, and has a 19% higher AUC. (**B**) Precision–recall plot (TPR versus precision): MSpresso increases precision at fixed recall across different score thresholds.

Similarly MSpresso outperforms the primary identification in Figure 3B, with higher TPR at the same precision.

The AUC is equal to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (Fawcett, 2006). MSpresso, with AUC = 0.89, provided a 19% increase in AUC compared with the primary MS/MS-based protein identification (AUC = 0.75), and a 27% increase compared with MSpresso with random $P(K|M)$ (AUC = 0.70) (Fig. 3A, Table 1).

3.2.2 Functional analysis of MSpresso identifications At a 5% FPR cutoff, MSpresso identifies ∼40% more proteins than ProteinProphet (Table 1)—327 versus 234. Of these identifications, 100 were not identified by the primary analysis, and are new MSpresso identifications. These proteins had sub-threshold ProteinProphet scores and were hence labeled 'not present'. Due to their high-mRNA concentration (>9 molecules/cell), their scores were increased above threshold and they were marked 'present'. The 100 newly identified proteins were not biased in their functions compared with the background: the union of all proteins identified by ProteinProphet and MSpresso (Berriz *et al*., 2003). Both ProteinProphet- and MSpresso-predicted proteins were enriched for molecules of high concentration, e.g. ribosomal proteins, proteins of biosynthesis and metabolism (P-value < 0.001).

We also analyzed the intersection of MSpresso 5% FPR proteins with the two yeast reference datasets described in Section 2.3.1. [YP4gte2 and YP3; see Supplementary Fig. S7 (P-value < 0.001), hypergeometric distribution]. Only two MSpresso-identified proteins were neither present in the reference sets, nor identified by the primary identification: GTO3, glutathione transferase (Nash *et al*., 2007)—a protein not unusual for cells growing (and dividing) in rich medium, and GCN4, a transcription activator of the amino acid starvation response (Lee *et al*., 2007). We did not expect GCN4 to be expressed in rich medium, and it is either a false positive or indicates a weak-starvation response.

MSpresso can also negatively influence protein identification, i.e. low-mRNA concentration shifts MSpresso scores below threshold even if the primary identification labeled such proteins 'present'. There were 15 such proteins in the yeast dataset. They were not enriched for any functional category and had low-mRNA concentration [≤0.88 molecules/cell; median $P(K|M) = 0.26$] in contrast to the median mRNA concentration across all genes [16 molecules/cell; median $P(K|M) = 0.80$]. All but three proteins were as present in the reference sets: two cell cycle proteins (SWE1, SSN3) and a protein of unknown function (MUK1).

3.3 Results on other yeast datasets and other organisms

We tested MSpresso for other biological conditions, organisms and mass spectrometers as detailed in Section 2.3. MSpresso increased the number of identifications at 5% FPR by 19–63% across all datasets (Table 1), while maintaining constant or higher precision than the primary identification (data not shown). MSpresso increased AUC by 3–19% across experiments; a substantial increase since AUC is related to the probability of correct classification. ROC plots are in the Supplementary Figure S8.

3.3.1 Yeast We applied MSpresso to three other yeast datasets: rich-medium yeast reanalyzed on a high-resolution mass spectrometer LTQ-OrbiTrap (Table 1: Yeast-YPD-ORBI), yeast grown in minimal medium (Table 1: Yeast-YMD-LCQ), and a sample from a sucrose gradient experiment (Table 1: Yeast-Fraction-LCQ). MSpresso-predictions for OrbiTrap and YMD experiments were strongly enriched for metabolic and ribosomal functions (P-value < 0.001) (Berriz *et al*., 2003)—proteins of these functions are typically in high concentration in growing and dividing yeast cells. In addition, MSpresso-predicted proteins from yeast grown in minimal medium are enriched for small molecule metabolism (P-value < 0.001), which is expected for growth in minimal medium.

3.3.2 Escherichia coli We applied MSpresso to cytosolic protein extract from *E.coli* grown in minimal medium analyzed on an LTQ-OrbiTrap (Section 2.3.3; Table 1, *Escherichia coli*-ORBI). The MSpresso-predicted proteins were enriched for the same functions as proteins from primary analysis: biosynthesis and translation [P-value < 0.001 using a background of all 3503 *E.coli* proteins with function annotation (Serres *et al*., 2004)]. The reference dataset was very small (∼370 proteins) and hindered immediate verification of the newly identified proteins.

3.3.3 Human We applied MSpresso on two human datasets described in Section 2.3.4 (Table 1: Human-LCQ, Human-ORBI). We found ∼20% more proteins in both datasets than the primary identifications, and these proteins were enriched for functions in metabolism, translation and biosynthesis (P-value < 0.001) (Berriz *et al*., 2003).

3.4 General applicability of the method

So far, we have discussed MSpresso models trained on high-quality protein reference sets available for the respective organism (dubbed the 'self' model). However, the method can be generalized to cases where little or no protein reference data is available. For example, the $P(K|M)$ estimates for *E. coli* (Supplementary Fig. S2) are much smaller than those in yeast, because the protein reference dataset comprises only ∼370 proteins. For this reason, we developed models that reuse $P(K|M)$ and $P(K|S)$ relationships learned from available datasets, albeit in different sample conditions or organisms. Our aim was to investigate the degree to which these relationships can be reused across datasets. The 'reuse' model for $P(K|S)$ involves applying the $P(K|S)$ logistic regression classifier, which was learned on high-quality data (e.g. rich-medium yeast), to other datasets. We now describe 'reuse' models for $P(K|M)$ based on the yeast model (or the best organism-specific model).

3.4.1 Generalizing P(K|M) First, we approximated $P(K|M)$ by a simple step function from Figure 2, estimating $P(K|(\log_{10}M < 0.5)) = 0.10$ and $P(K|(\log_{10}M > 9)) = 0.90$ (results not shown). Next, we derived two 'scaled' models: SCALE-UP scales the $P(K|M)$ values in Figure 2 to a [0,1] interval, and SCALE-DOWN scales $P(K|M)$ to half of the original values (results not shown). The mRNA concentrations from the rich-medium yeast model were also scaled to a [0,1] interval.

A SCALE-UP reuse model derived from Figure 2 (rich-medium yeast data) resulted in 3–14% AUC increase when applied to the other yeast datasets (Supplementary Table S10). We also derived SCALE-UP models from the $P(K|M)$ distributions learned on other organisms (*E.coli*, human) and applied them to the respective organism's datasets. Selected results are presented in the Supplementary Material (Table S10).

In general, we recommend using the self model if a high-quality experiment-specific protein reference set is available. When such data is unavailable, we recommend using an organism-specific SCALE-UP model or using the yeast SCALE-UP model.

3.4.2 Evaluation without a protein reference set So far, we have used protein reference sets as ground truth to define true and false identifications. Though we expect our canonical reference dataset for yeast in rich medium to cover most of the expressed yeast proteins (2/3 of the genome), such large-scale protein datasets are mostly unavailable for other organisms. Thus we need to define false identifications to estimate null score models (Choi and Nesvizhskii, 2008; Choi *et al*., 2008) and error estimates without a reference set.

There have been recent efforts in standardizing null models for peptide identifications; however, no general consensus has yet been reached at the protein level (Elias and Gygi, 2007; Fitzgibbon *et al*., 2008; Kall *et al*., 2008; Kim *et al*., 2008). In general, a set of 'null' or 'decoy' proteins is appended to the sample organism database ('target') before the MS/MS search. Decoy proteins are considered to be negative instances. Proteins from another organism or shuffled/reversed protein sequences have been used as decoys (Kall *et al*., 2008).

To investigate evaluation without using a ground-truth reference set, we applied MSpresso to both real (target) and shuffled (decoy) yeast protein sequences, labeling any identified decoys as false identifications. We first ran the MS/MS analysis on rich-medium yeast, matching experimental spectra against a concatenated database of real and shuffled sequences. This procedure resulted in protein identification scores (S) for both target and decoy proteins, letting us estimate $P(K|S)$ as before. However, $P(K|M)$ cannot be estimated for decoy proteins in the same manner as for the targets, since decoys are artificial proteins and do not have associated mRNA abundances. Hence, we investigated different random $P(K|M)$ distributions for the decoy proteins: e.g. random sampling from the target $P(K|M)$ distribution, random sampling from the $P(K|M)$ values of target proteins that are absent from the reference set (negative instances) and constant at the minimum of the target $P(K|M)$ distribution (Supplementary Section 2.6.2). We measure the increase in AUC and the number of proteins identified using multiple error measures (FPR, FDR, q-value: see Supplementary Section 2.5). Running MSpresso on yeast with five shuffled decoy databases results in up to 5% AUC increase and up to 14% more identified proteins at 5% FPR. We also experimented with different shuffled database sizes ranging from 0.25 to 20 times the size of the real database (Supplementary Section 2.6.1). Detailed results are presented in Supplementary Section 2.6 and Table S6.

3.5 Conclusions

We present a method called MSpresso that improves our ability to identify proteins in large-scale shotgun proteomics experiments. MSpresso learns the relationship between mRNA concentration and the probability of protein presence in a sample, and then applies this relationship to boost sub-threshold protein identifications in data from MS/MS experiments. We assess MSpresso performance with ROC curves over a large range of FPRs, and show a boost of 19% AUC in a yeast sample, as well as 40% increase in protein identifications at 5% FPR, at the same or higher precision. We also generalize the method to other experimental conditions, mass spectrometers and organisms, even in the absence of a high-quality training dataset. By integrating mRNA evidence into the MSpresso score, we improve the confidence in our predicted proteins, which leads to more identifications at similar error rates. MSpresso is not restricted to particular MS-based methods of primary protein identification but could be applied to any large-scale proteomics dataset that contains scores signifying confidence in correct identification.

Our results have interesting biological implications. For example, the relationship between mRNA concentration and protein identification (Fig. 2) is different from what we would expect given a strong correlation between mRNA concentration and protein abundance [$R^2 = 0.73$ (Lu *et al*., 2007)]. In general, yeast proteins are very easily identifiable in shotgun proteomics experiments if their corresponding mRNA is present at 9 molecules/cell (or higher) on average, and at around 1 molecule/cell mRNA, current high-throughput methods largely fail to detect proteins. This empirical relationship between mRNA abundance and protein identification may be refined with increasing experimental sensitivity.

Further, given that we now have several large-scale datasets available, we can attempt to describe the expressed yeast proteome as comprehensively as is currently possible and answer the simple but fundamental question: 'How many proteins are expressed in yeast growing in log-phase under nutrient rich conditions?'. The union of our two MSpresso-predicted datasets (LCQ, ORBI) and the protein reference datasets comprises 3797 cytosolic proteins expressed in yeast growing in rich medium at log-phase; 2364

(62%) of these proteins occur in two or more datasets, and may thus form a core set of reliably identified proteins. Given that there are 4962 non-membrane yeast proteins in total, we can estimate upper boundaries of observed transcription and translation products—and these estimates are impressively high. The majority of all yeast genes, 84% (4165) have observed mRNA, and for 70% (3512) we observe both mRNA and protein. Interestingly, there are 282 genes for which no mRNA but protein is observed: mRNA may exist at only very low levels or is rapidly degraded. Together, these numbers indicate that even in an unperturbed, comparatively simple unicellular eukaryote, a very large number of proteins are expressed and form a complex cellular machinery.

Funding: National Science Foundation (DBI-0640923, IIS-0325116); Welch (F-1515); Packard Foundation; National Institutes of Health (GM06779-01, GM076536-01). International Human Frontier Science Program (to C.V.).

Conflict of Interest: none declared.

REFERENCES

Allen,T.E. *et al*. (2003) Genome-scale analysis of the uses of the Escherichia coli genome: model-driven analysis of heterogeneous data sets. *J. Bacteriol.*, **185**, 6392–6399.

Berriz,G.F. *et al*. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.

Brunner,E. *et al*. (2007) A high-quality catalog of the Drosophila melanogaster proteome. *Nat. Biotechnol.*, **25**, 576–583.

Chi,A. *et al*. (2007) Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. *Proc. Natl Acad. Sci. USA*, **104**, 2193–2198.

Choi,H. and Nesvizhskii,A.I. (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.*, **7**, 47–50.

Choi,H. *et al*. (2008) Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.*, **7**, 286–292.

Corbin,R.W. *et al*. (2003) Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proc. Natl Acad. Sci. USA*, **100**, 9232–9237.

Covert,M.W. *et al*. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, **429**, 92–96.

de Godoy,L.M. *et al*. (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.*, **7**, R50.

Elias,J.E. and Gygi,S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, **4**, 207–214.

Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.

Fitzgibbon,M. *et al*. (2008) Modes of inference for evaluating the confidence of peptide identifications. *J. Proteome Res.*, **7**, 35–39.

Futcher,B. *et al*. (1999) A sampling of the yeast proteome. *Mol. Cell. Biol.*, **19**, 7357–7368.

Graumann,J. *et al*. (2007) SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins. *Mol. Cell Proteomics*.

Greenbaum,D. *et al*. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.*, **4**, 117.

Gygi,S.P. *et al*. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.*, **17**, 994–999.

Holstege,F.C. *et al*. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.

Kall,L. *et al*. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.

Kall,L. *et al*. (2008) Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.*, **7**, 40–44.

Keller,A. *et al*. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **74**, 5383–5392.

Kim,S. *et al*. (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.*, **7**, 3354–3363.

Lee,B. *et al*. (2007) Yeast phenotypic assays on translational control. *Methods Enzymol.*, **429**, 105–137.

Link,A.J. *et al*. (1997) Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12. *Electrophoresis*, **18**, 1259–1313.

Lopez-Campistrous,A. *et al*. (2005) Localization, annotation, and comparison of the Escherichia coli K-12 proteome under two states of growth. *Mol. Cell Proteomics*, **4**, 1205–1209.

Lu,P. *et al*. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.*, **25**, 117–124.

Nash,R. *et al*. (2007) Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res.*, **35**, D468–D471.

Nesvizhskii,A.I. *et al*. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, **75**, 4646–4658.

Newman,J.R. *et al*. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*.

Peng,J. *et al*. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.*, **2**, 43–50.

Planta,R.J. and Mager,W.H. (1998) The list of cytoplasmic ribosomal proteins of Saccharomyces cerevisiae. *Yeast*, **14**, 471–477.

Ptacek,J. *et al*. (2005) Global analysis of protein phosphorylation in yeast. *Nature*, **438**, 679–684.

Robinson,M.D. *et al*. (2002) FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.

Serres,M.H. *et al*. (2004) GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. *Nucleic Acids Res.*, **32**, D300–D302.

Smirnova,J.B. *et al*. (2005) Global gene expression profiling reveals widespread yet distinctive translational responses to different eukaryotic translation initiation factor 2B-targeting stress pathways. *Mol. Cell. Biol.*, **25**, 9340–9349.

Velculescu,V.E. *et al*. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Wang,Y. *et al*. (2002) Precision and functional specificity in mRNA decay. *Proc. Natl Acad. Sci. USA*, **99**, 5860–5865.

Washburn,M.P. *et al*. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.*, **19**, 242–247.