Analyzing marginal cases in differential shotgun proteomics

Paulo C. Carvalho1,2,*, Juliana S. G. Fischer1,3, Jonas Perales2, John R. Yates4, Valmir C. Barbosa5 and Elias Bareinboim6

1Center for Technological Development in Health, 2Department of Physiology and Pharmacodynamics, Laboratory of Toxilog, 3Department of Biochemistry and Molecular Biology, Laboratory for Functional Genomics and Bioinformatics, Oswaldo Cruz Institute, Rio de Janeiro, Brazil, 4Department of Chemical Physiology, The Scripps Research Institute, La Jolla, CA 92037, USA, 5Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil and 6Computer Science Department, University of California, Los Angeles, CA 90024, USA

Associate Editor: John Quackenbush

ABSTRACT

Summary: We present an approach to statistically pinpoint differentially expressed proteins that have quantitation values near the quantitation threshold and are not identified in all replicates (marginal cases). Our method uses a Bayesian strategy to combine parametric statistics with an empirical distribution built from the reproducibility quality of the technical replicates.

Availability: The software is freely available for academic use at http://pcarvalho.com/patternlab.

Contact: paulo@pcarvalho.com

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on August 24, 2010; revised on October 18, 2010; accepted on November 5, 2010

1 INTRODUCTION

Shotgun proteomics describes a large-scale approach to analyzing complex peptide mixtures (i.e. mixtures originating from biological fluids, cell lysates, etc.). Briefly, the strategy is to perform protein digestion followed by peptide chromatographic separation online with tandem mass spectrometry (MS2) for protein identification and moreover that it may occur in any of the replicates from state B1. We assume throughout that the appearance of any given protein in a replicate from B2 is subject to the same underlying laws governing its appearance in replicates from B1, and moreover that it may occur in any of the replicates from B2 independently with the same probability. This implies that the number of replicates from B2 containing that protein is distributed binomially. Henceforth, we use the smoother, approximate formula of the Poisson distribution instead. Accordingly, the probability that the protein appears in at least one of the replicates from B2 with mean λ is denoted by Pois(u, λ). In our estimates, we always choose the value of λ in reference to what is observed or hypothesized with respect to state B1.

From a Bayesian perspective, we begin by estimating the prior probability P(H) that protein P does not appear in any replicate from state B2. If r is the number of replicates from state B1 in which P is detected, then we set P(H) = Pois(r, λ). Similarly, computing the desired probability, P(H|D), requires that first we obtain P(D|H) and P(D|not H), that is, the probabilities that P is detected in at least one replicate from B1 conditioned, respectively, on the fact that it does not or does appear in replicates from B2. In order to estimate either probability, we first partition the B1-replicate proteins into four groups of approximately the same size, each corresponding to one of the categories low, medium, high, or very high, according to the average signal of each protein (e.g. spectral count, peak area, etc.) over the replicates in which it appears. Let G denote the group to which protein P belongs. Our estimates of P(D|H) and P(D|not H) are relative to G, therefore specific to a certain range of average reproducibility, many marginal cases that are likely to be an artifact of chance may be included in the results and shadow important aspects. Moreover, many false negative cases may be lost.

2 PROBLEM FORMULATION AND MODELING

Consider two biological states B1 and B2 and two experimental datasets, one containing replicates from state B1, the other as many replicates from state B2. We address the question of estimating the probability that a protein appearing in at least one replicate from state B1 is differentially expressed with respect to state B2, i.e. that it is found in none of the replicates from B2.

If P is the protein in question, then our aim is to estimate the probability P(H|D), where H stands for ‘P is not detected in any replicate from B2’ and D for ‘P appears in at least one of the replicates from B1’. We assume throughout that the appearance of any given protein in a replicate from B2 is subject to the same underlying laws governing its appearance in replicates from B1, and moreover that it may occur in any of the replicates from B2 independently with the same probability. This implies that the number of replicates from B2 containing that protein is distributed binomially. Henceforth, we use the smoother, approximate formula of the Poisson distribution instead. Accordingly, the probability that the protein appears in at least one of the replicates from B2 with mean λ is denoted by Pois(u, λ). In our estimates, we always choose the value of λ in reference to what is observed or hypothesized with respect to state B1.

From a Bayesian perspective, we begin by estimating the prior probability P(H) that protein P does not appear in any replicate from state B2. If r is the number of replicates from state B1 in which P is detected, then we set P(H) = Pois(r, λ). Similarly, computing the desired probability, P(H|D), requires that first we obtain P(D|H) and P(D|not H), that is, the probabilities that P is detected in at least one replicate from B1 conditioned, respectively, on the fact that it does not or does appear in replicates from B2. In order to estimate either probability, we first partition the B1-replicate proteins into four groups of approximately the same size, each corresponding to one of the categories low, medium, high, or very high, according to the average signal of each protein (e.g. spectral count, peak area, etc.) over the replicates in which it appears. Let G denote the group to which protein P belongs. Our estimates of P(D|H) and P(D|not H) are relative to G, therefore specific to a certain range of average reproducibility, many marginal cases that are likely to be an artifact of chance may be included in the results and shadow important aspects. Moreover, many false negative cases may be lost.

*To whom correspondence should be addressed.
signal. In what follows, we use \( f_t \) to denote the fraction of group-\( G \) proteins that occur in \( t \) replicates from state \( B_1 \).

We estimate \( P(D/H) \) as the sum of probabilities of pairs of independent events. If \( n \) is the total number of replicates from either state, we consider one pair for each possible number \( t \) of replicates from state \( B_1 \), \( t = 1, 2, \ldots, n \). The two independent events for each pair are that a randomly chosen protein from group \( G \) appears in \( t \) replicates from state \( B_1 \), and that it appears in none of the replicates from state \( B_2 \). Thus,

\[
P(D/H) = \sum_{t=1}^{n} f_t \text{Poi}(0, t).
\]

The case of \( P(D/\text{not } H) \) is similar, but now the invalidity of \( H \) implies that we must sum up the probabilities that the randomly chosen protein from group \( G \) appears in \( u \) replicates from state \( B_2 \), \( u = 1, 2, \ldots, n \). We then obtain

\[
P(D/\text{not } H) = \sum_{t=1}^{n} f_t \sum_{u=1}^{n} \text{Poi}(u, t).
\]

The desired probability, finally, follows from the Bayesian inversion formula,

\[
P(D/H) = \frac{P(D/H|P(H))}{P(D/H|P(H)) + P(D/\text{not } H)(1 - P(H))}
\]

and is henceforth used as a \( p \)-value for all proteins in \( G \) that appear in \( r \) replicates from state \( B_1 \).

3 DATA ACQUISITION

For evaluation of the above methodology, we used two shotgun proteomic datasets acquired by Fischer et al. (2010). Briefly, the authors employed Multi-dimensional Protein Identification Technology (MudPIT; Washburn et al., 2001) to compare the A172 cell line in two biological states, here identified with the \( B_1 \) and \( B_2 \) states of Section 2. Each state was analyzed in triplicates (i.e. \( n = 3 \)). Relative quantitation was performed by spectral counting. A protein required a minimum of two peptides (thus, two unique spectral counts) to be considered.

4 RESULTS

Each of Supplementary Figures 1A, B and C shows a Venn diagram (VD) of identified proteins from \( B_1 \) and \( B_2 \) appearing in at least one, at least two, and all three replicates, respectively. Supplementary Figures 2A and B show VDs comparing uniquely identified proteins among the technical replicates from \( B_1 \) and \( B_2 \), respectively. Both Supplementary Figures 1 and 2 corroborate the great variability claimed by Liu et al. (2004).

The model described in Section 2 has been implemented as part of the PatternLab for proteomics suite (Carvalho et al., 2008).

Results on the biological states to which Section 3 refers are shown in Supplementary Tables I and II, respectively, to verify differential expression in state \( B_1 \) relative to state \( B_2 \) and conversely (i.e. reversing the roles of the two states in the discussion of Section 2). Clearly, proteins that are more reproducible (appear in more replicates) yield lower \( p \)-values. The resulting algorithm was also incorporated into PatternLab’s area-proportional VD module (Carvalho et al., 2010). The user can now choose between generating VDs by filtering proteins that appear in at least a certain number of replicates, or by using the new approach through a user-specified \( p \)-value. The new option can be used to eliminate proteins that cannot be claimed to be statistically differentially expressed. Supplementary Figure 3 shows a VD that considers a \( p \)-value cutoff of 0.05 for the two biological states of Section 3, instead of the replicate-cutoff criterion used in Supplementary Figure 1.

5 FINAL CONSIDERATIONS

An alternative, simple strategy to pinpoint marginal proteins representative of a biological state is to consider only proteins that appear in a minimum number of replicates. Such an approach, however, is arbitrary and lacks proper foundation. The approach we have described, on the other hand, is well-founded and therefore amounts to a more refined method. It is useful especially in generating VDs, such as the one in Supplementary Figure 3, for the study of proteins that are representative of a given biological state. We note, in relation to VDs such as this, that uniquely identified proteins in the VD are not to be claimed as being unique to a state; instead, they are most likely differentially expressed.

Funding: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES-Fiocruz 30/2006); Programa de Desenvolvimento Tecnológico em Insumos para Saúde (PDTIS-Fiocruz); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); BBP grants from Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro; National Institutes of Health (NIH 5R01MH07880 and P41 RR011823).

Conflict of Interest: none declared.

REFERENCES


