BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 12 2012, pages 1652–1654
doi:10.1093/bioinformatics/bts247

Sequence analysis

Advance Access publication April 26, 2012

Improving the TFold test for differential shotgun proteomics

Paulo C. Carvalho[1,*], John R. Yates III[2] and Valmir C. Barbosa[3]

[1]Carlos Chagas Institute, Fiocruz, Paraná, Brazil, [2]Department of Chemical Physiology, The Scripps Research Institute, La Jolla, California, USA 2, 92037 and [3]Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Associate Editor: Martin Bishop

ABSTRACT

Summary: We present an updated version of the TFold software for pinpointing differentially expressed proteins in shotgun proteomics experiments. Given an FDR bound, the updated approach uses a theoretical FDR estimator to maximize the number of identifications that satisfy both a fold-change cutoff that varies with the t-test P-value as a power law and a stringency criterion that aims to detect lowly abundant proteins. The new version has yielded significant improvements in sensitivity over the previous one.

Availability: Freely available for academic use at http://pcarvalho.com/patternlab.

Contact: paulo@pcarvalho.com

Received on 27 February, 2012; revised on 23 March, 2012; accepted on 17 April, 2012

1 INTRODUCTION

One of the goals of shotgun proteomics is to perform large-scale comparisons of complex protein mixtures (e.g. biological fluids or whole-cell lysates). It comprises protein digestion followed by peptide chromatographic separation online with tandem mass spectrometry (MS2) for protein identification and quantitation. There are several strategies for protein quantitation; examples are metabolic labeling (Gouw *et al.*, 2010), chemical derivatization (Thompson *et al.*, 2003) and label-free (Bondarenko *et al.*, 2002).

We previously described PatternLab for Proteomics (Carvalho *et al.*, 2010), a computational environment for analyzing shotgun proteomics data. The modules available in PatternLab serve various purposes, ranging from preparing decoy databases, to selecting and sharing reliable protein identifications (Carvalho *et al.*, 2012), to performing gene ontology analyses (Carvalho *et al.*, 2009), and many more. One of its most popular modules is the TFold test for pinpointing differentially expressed proteins quantitated by any of the above strategies. Briefly, in the previous version a PatternLab file containing protein quantitation data from different biological conditions would first be filtered according to specified t-test P-value and fold-change cutoffs. Then the user would interactively experiment with different fold-change cutoffs, each leading to a different list of proteins through the BH FDR estimator (Benjamini and Hochberg, 1995) for the same user-specified q-value, henceforth referred to as α. This process would ultimately lead to a fold-change cutoff that maximized the number of proteins in the list.

Even though this simple strategy has been effective in a number of occasions, imposing a fixed fold-change cutoff could discard proteins with very low P-values but not satisfying the fold-change cutoff. This seems unreasonable because, in the proteomics scenario, the magnitude of a protein's fold change need not correlate with its biological importance. In fact, molecular abundance is only one of the various factors controlling protein activity, not necessarily the most important one (Weiss *et al.*, 2010). Here, we describe how we reformulated the TFold test to better address these limitations and increase sensitivity.

We also tackle another common problem in proteomics, namely that of dealing with lowly abundant proteins (e.g. those having low spectral counts), which usually turn out to be over-represented in complex mixtures. These proteins are more prone to getting around the barriers imposed by common statistical filters (e.g. the t-test) because they tend to artificially acquire low P-values. To exemplify, assume the following sets of spectral counts for a given protein in two biological conditions: {1, 1, 1} and {5, 6, 5}. Notably, in this extreme example, there is a considerable fold change while the P-value is artificially low due to the lack of variation. In our experience, it is much more likely that this protein is a false-positive than another having significantly higher spectral counts but with a higher P-value as well. Our updated TFold approach introduces a method to quickly highlight (and separate) proteins such as the one here exemplified. For these proteins, additional experimentation is recommended before ascertaining their status as differentially expressed.

2 VARIABLE FOLD-CHANGE CUTOFF

The centerpiece of our fold-change reformulation is to allow the fold-change cutoff for a protein to be given as a function of its t-test P-value. In order to address the issues raised above, a larger than 1 fold-change cutoff must decrease as the P-value decreases; likewise, a smaller than 1 fold-change cutoff must increase as the P-value decreases. We postulate a power-law functional form for some exponent $z > 0$, thence the fold-change cutoff is proportional to p^z in the former case and to p^{-z} in the latter, where p is the P-value in question. We resolve the proportionality constant by imposing a fold-change cutoff of 1 when p is the lowest P-value for the proteins in the data set, here denoted by $p_{\min}$. A protein of P-value p is then rejected (i.e. declared not differentially expressed) if its fold change lies between $(p_{\min}/p)^z$ and $(p/p_{\min})^z$.

Thus, in the revised version of the TFold test, one core operation is to filter the proteins according to the P-value-dependent fold-change cutoffs that result from a fixed value of the exponent z. Building on this operation, our software automatically maximizes the list of

*To whom correspondence should be addressed.

identified proteins by varying the value of z given α (decreasing the value of z tightens the fold-change interval for protein rejection). Thus, z works as a fold-change stringency parameter (F-stringency).

3 HANDLING LOWLY ABUNDANT PROTEINS

Another core operation is to filter and highlight proteins that would be considered differentially expressed according to α and the fold-change cutoff function but still have a high chance of being false positives. This occurs mainly in the case of lowly abundant proteins, as their quantitation values are often compromised. Briefly, we begin by assuming that the quantitation value of a randomly chosen quantitated protein, regardless of biological condition, is exponentially distributed with parameter λ which implies that both mean and standard deviation are given by $1/\lambda$. As is well-known, the maximum-likelihood estimates of $1/\lambda$ is the sample mean of all proteins' quantitation values (to which a protein contributes as many times as there are biological conditions in which it appears).

This given, a new parameter, called λ-stringency (L-stringency), is introduced to pinpoint each problematic protein. This is done by specifying a fraction of $1/\lambda$ below which the sample mean of that protein in a certain biological condition is too low for it to qualify as differentially expressed in that condition. We exemplify the use of L-stringency as follows. Suppose that the sets {2, 2, 2} and {7, 8, 7} of spectral counts were acquired for a certain protein in two biological conditions and that $1/\lambda = 2$. Choosing an L-stringency of 0.4 disqualifies the protein as differentially expressed in the second biological condition relative to the first, since $(7+8+7)/3 < 0.4 \times 20$. We recommend this L-stringency of 0.4, but it should be evaluated by the user on a case-by-case basis.

4 RESULTS

The effectiveness of our updated algorithm is exemplified on the data set provided by Fischer *et al.* (2011) comparing A172 cancer cells to those (A172R) resistant to Perillyl Alcohol (POH), a naturally occurring chemotherapeutic agent that induces apoptosis. The authors provide a list of protein identifications quantitated by spectral counting and satisfying a 1% FDR for their identifications according to DTASelect (Cociorva *et al.*, 2006). In the original analysis, the authors converged to a fixed fold-change cutoff of 2.5, resulting in 57 differentially expressed proteins for a P-value cutoff of 0.01 and $\alpha = 0.01$. The new algorithm converged to $z = 0.14$, pinpointing 77 differentially expressed proteins for the same α, plus six proteins satisfying the same criteria but reported in a separate list as having been marked by the L-stringency filter, thus suggesting the need for further experimentation. Figure 1 is a snapshot of Pattern Lab's graphical user interface during analysis.

An example of a protein that was only marked as differentially expressed by the revised method is glyceraldehyde-3-phosphate dehydrogenase (GAPDH; IPI00219018.7). Identifying this protein is important because we hypothesize that it could be related to the initiation of apoptosis when the cell's medium contains POH. The sets of spectral counts for this protein were {117, 71, 141} for the wild-type and {271, 255, 240} for the resistant cell culture, but further discussions on any of the corresponding biological interpretations are beyond the scope of this manuscript.

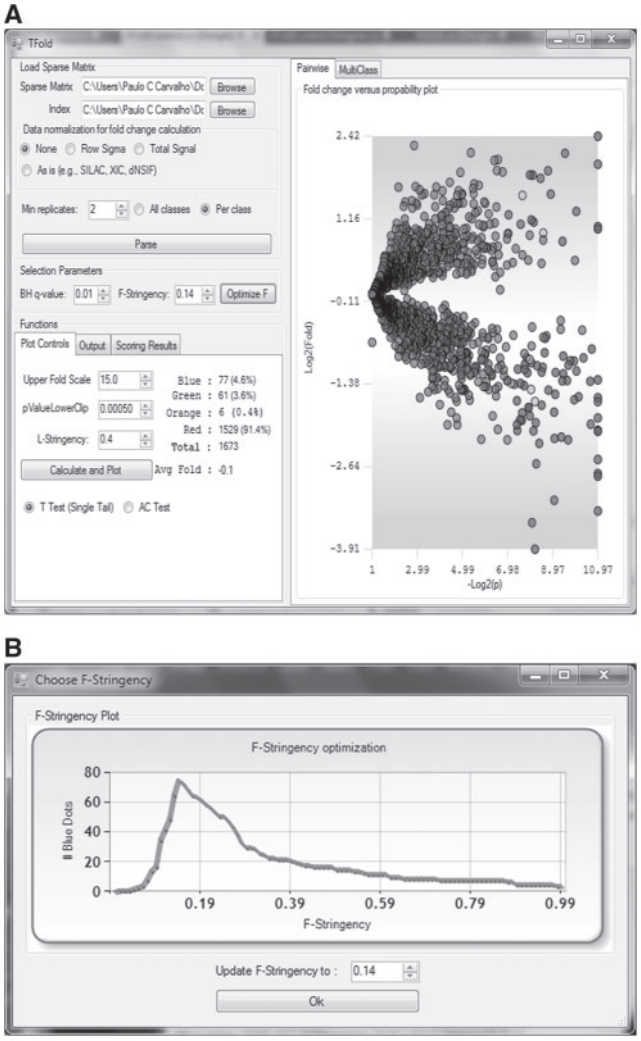

Fig. 1. Panel **A** shows the TFold graphical user interface and the results of our updated analysis relative to Fischer *et al.* (2011). Each protein is mapped as a dot on the plot according to its $-\log_2(P\text{-value})$ (x-axis) and $\log_2(\text{fold change})$ (y-axis). Red dots are proteins that satisfy neither the fold-change cutoff nor the FDR cutoff α. Green dots are those that satisfy the fold-change cutoff but not α. Orange dots are those that satisfy both the fold-change cutoff and α but are lowly abundant proteins and thus require further experimentation to certify their differential expression. Blue dots, finally, are proteins that satisfy all statistical filters. Note that the bounds separating green dots from red do not correspond to a single fold-change cutoff (a 'horizontal' line) as in the previous approach. Instead, separation is achieved through several cutoffs along a 'conical' region (due to the logarithms), one cutoff for each P-value, with vertex at $P\text{-value} = p_{\min}$ and Fold change = 1. However, as a consequence of the BH FDR estimator, we have a single P-value cutoff (a vertical line) separating blue and orange dots from green and red in both approaches. Panel **B** is automatically displayed once the Optimize button is pressed. Given α, it plots the distribution of blue dots as a function of the F-stringency parameter z

5 FINAL CONSIDERATIONS

Computational approaches for pinpointing differentially expressed proteins can, at best, provide a list of putative biomarkers that could ultimately aid in the understanding of pathology or be specific to a biological condition. Further validation is to be carried out in different cohorts.

The application of the F-stringency parameter prior to using the BH FDR estimator eliminates hypotheses (i.e. putative differentially expressed proteins) that are most likely false positives. Applying the L-stringency parameter also prior to using the BH FDR estimator has the potential of further eliminating the likely false positives related to some of the lowly abundant proteins. The final, reduced list to be evaluated by the BH FDR estimator maximizes the TFold test sensitivity, as the BH FDR estimator penalizes the number of hypotheses tested by including it in the denominator of its equation.

To summarize, we have described an algorithm for maximizing the number of differentially expressed protein candidates. It optimizes a fold-change cutoff that varies with the proteins' *P*-values as a power law and provides a means for highlighting (and separating) lowly abundant proteins prior to filtering by the BH FDR estimator. This algorithm resulted in a significant increase in sensitivity when compared to its previous version.

ACKNOWLEDGEMENT

The authors thank Prof. Gilberto Barbosa Domont for critically reviewing the manuscript.

Funding: PCC and VCB were funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) /Fiocruz 30/2006, Programa de Desenolvimento Tecnológico de Insumos para a Saúde (PDTIS), Conselho Nacional de Pesquisa (CNPq,) Fundação de Amparo a Pesquisa do Rio de Janeiro (FAPERJ BBP). JRY was funded by the National Institutes of Health (NIH) [P41 RR011823, P41 GM130533, 5R01 MH067880].

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Bondarenko,P.V. *et al.* (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal. Chem.*, **74**, 4741–4749.

Carvalho,P.C. *et al.* (2009) GO Explorer: a gene-ontology tool to aid in the interpretation of shotgun proteomics data. *Proteome Sci.* **7**, 6.

Carvalho,P.C. *et al*. (2010) Analyzing shotgun proteomic data with PatternLab for proteomics. *Curr. Protoc. Bioinformatics*, Chapter 13, Unit 15.

Carvalho,P.C. *et al.* (2012) Search engine processor: filtering and organizing PSMs. *Proteomics*, **12**, 944–949. doi: 10.1002/pmic.201100529

Cociorva,D. *et al*. (2006) Validation of tandem mass spectrometry database search results using DTASelect. *Curr. Protoc. Bioinformatics*, Chapter 13, 13.4.1–13.4.14.

Fischer,J.S.G. *et al*. (2011) Chemo-resistant protein expression pattern of glioblastoma cells (A172) to perillyl alcohol. *J. Proteome Res.*, **10**, 153–160.

Gouw,J.W. *et al*. (2010) Quantitative proteomics by metabolic labeling of model organisms. *Mol. Cell Proteom.*, **9**, 11–24.

Thompson,A. *et al*. (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.*, **75**, 1895–1904.

Weiss,M. *et al*. (2010) Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics*, **10**, 1297–1306.