BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 12 2010, pages 1574–1575
doi:10.1093/bioinformatics/btq171

Gene expression

Advance Access publication April 22, 2010

PepC: proteomics software for identifying differentially expressed proteins based on spectral counting

N.L. Heinecke[1], B.S. Pratt[1], T. Vaisar[2] and L. Becker[2,*]

[1]Insilicos LLC and [2]Department of Medicine, University of Washington, Seattle, WA 98109, USA

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Identifying biologically significant changes in protein abundance between two conditions is a key issue when analyzing proteomic data. One widely used approach centers on spectral counting, a label-free method that sums all the tandem mass spectra for a protein observed in an analysis. To assess the significance of the results, we recently combined the *t*-test and *G*-test, with random permutation analysis, and we validated this approach biochemically. To automate the statistical method, we developed PepC, a software program that balances the trade-off between the number of differentially expressed proteins identified and the false discovery rate. This tool can be applied to a wide range of proteomic datasets, making data analysis rapid, reproducible and easily interpretable by proteomics specialists and non-specialists alike.

Availability and implementation: The software is implemented in Java. It has been added to the Trans Proteomic Pipeline project's 'Petunia' web interface, but can also be run as a command line program. The source code is GNU Lesser General Public License and the program is freely available on the web. http://sashimi.svn.sourceforge.net/viewvc/sashimi/trunk/trans_proteomic_pipeline/src/Quantitation/Pepc

Contact: levb@u.washington.edu; brian.pratt@insilicos.com

Received on February 26, 2010; revised on March 26, 2010; accepted on April 14, 2010

1 INTRODUCTION

Two important objectives of proteomics—global assessment of protein expression levels and biomarker discovery—depend critically on measuring relative protein abundance. Mass spectrometry (MS) has emerged as the leading technology for this purpose because it is easily applied to high-throughput analyses and provides broad coverage of the proteome (Abersold and Mann, 2003; Sadygov *et al.*, 2004). A potential limitation is that results are generally semiquantitative (Mason and Liebler, 2003). Isotope labeling of peptides has been developed to circumvent this problem. Labeling methods are expensive and limited in the number of samples that can be coordinately interrogated, making application to samples from large-scale clinical studies difficult.

Spectral counting and peptide ion intensity are alternative, label-free methods for quantifying relative protein abundance (Bondarenko *et al.*, 2002; Liu *et al.*, 2004; Old *et al.*, 2005; Wiener *et al.*, 2004). Spectral counting sums all MS/MS spectra observed for peptides derived from a single protein. Because abundant proteins are more likely to be identified during data-dependent MS/MS scanning, spectral counting has the potential to quantify protein levels. Indeed, studies have shown strong correlations between relative protein abundance, as assessed by peptide ion intensities and spectral counting (Bondarenko *et al.*, 2002; Liu *et al.*, 2004; Old *et al.*, 2005). More recent studies have shown that the accuracy of spectral counting can be improved by normalization methods such as correction for sampling depth, protein molecular weight or number of tryptic peptides expected (reviewed in Zhu *et al.*, 2010).

A variety of statistical approaches have been developed to assess differences in spectral counts (Choi *et al.*, 2008; Fu *et al.*, 2008; Old *et al.*, 2005; Pavelka *et al.*, 2008; Zhang *et al.*, 2006). However, a key issue is correcting for multiple comparisons. Overly stringent significance cutoffs identify few differentially expressed proteins, while looser criteria identify more proteins at the expense of increased false discovery rates (FDRs). Therefore, the development of rigorous statistical strategies that maximize the identification of significant differences in protein expression while minimizing the FDR is of critical importance in proteomics.

We recently developed a novel method for identifying statistically significant differences in protein abundance based on spectral counting, and validated our results biochemically. We then used this approach to identify an atherogenic protein network in macrophages (Becker *et al.*, 2010). In this application note, we describe freely accessible software, PepC, for automating our approach.

2 METHODS

Our method implements two widely used statistical tests for analyzing proteomics data: the *t*-test and *G*-test. Moreover, it *combines the significance cutoffs* from the two tests so that each one's stringency can be lowered without inflating the FDR. The method is successful because the *t*-test and *G*-test scrutinize proteomic data in different ways. Thus, the *G*−test (Sokhal and Rholf, 1994), a likelihood ratio based on a χ^2 distribution with one degree of freedom, determines the confidence level primarily from the *magnitude* of the difference in spectral counts. The *t*−test, which assumes that spectral count measurements are normally distributed, emphasizes the *reproducibility* of spectral counts. An analogous approach widely used to analyze microarray data, supplements *t*-test significance cutoffs with fold change requirements to ensure high confidence when exploring changes in gene expression (Cui and Churchill, 2003).

PepC automates the detection of differentially expressed proteins based on a matrix of different *t*-test and *G*-test confidence intervals. It also empirically calculates the FDR for each combination within the matrix by random permutation analysis (Benjamini and Hochberg, 1995). The confidence intervals for the two tests are then optimized to maximize the detection of differentially abundant proteins at any given user-defined FDR. The program

*To whom correspondence should be addressed.

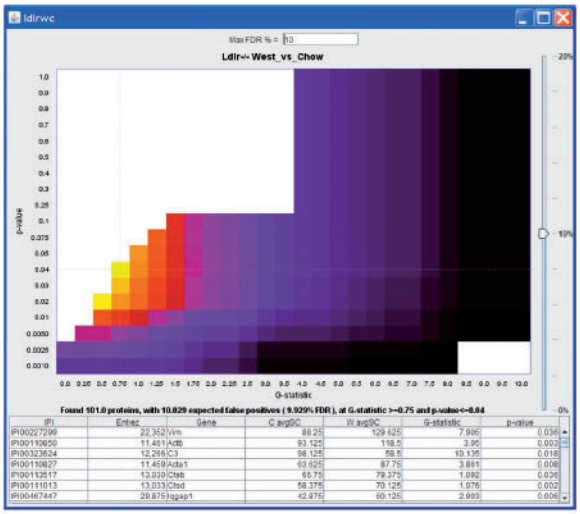

Fig. 1. (**A**) Screen shot from PepC illustrating differentially expressed proteins that were isolated from media of macrophages of low density lipoprotein receptor-deficient (*Ldlr–/–*) mice on a low-fat (chow) or high-fat (Western) diet (Becker *et al.*, 2010). The slider control (right edge of panel) is set to yield a maximum FDR of 10%. The heatmap displays combinations of *t*-test (*y*-axis, *P*-value) and *G*-test (*x*-axis, *G*-statistic) confidence intervals that meet this condition (FDR $\leq$10%). Its color scale (yellow = most, black = least) indicates the number of proteins whose abundances differ significantly between the two groups. When a box is selected (in this case, *G*-statistic $\geq$0.75 and *P*-value $\leq$0.04), the number of proteins identified and the expected false positives are shown in the dialog directly beneath the heatmap. The individual proteins and their respective statistical parameters are displayed in the window at the bottom of the screen.

requires at least three independent values per condition; however, we find that higher replicate numbers make FDR estimation more reliable.

PepC is implemented in Java, enabling use in any operating system. It supports two different types of user interaction. It can act as a straightforward command line application that performs calculations and presents a JFreeChart-based graphical user interface for exploration of the results. Alternatively, a web server module (PepcView) that performs the calculations uses JSON to send its results to a GWT-based web client for interactive exploration by the user. The latter mode is integrated with the Trans Proteomic Pipeline (TPP), a widely used proteomic analysis tool.

3 RESULTS

PepC is a tool for identifying differentially expressed proteins based on spectral count measurements in MS/MS studies. Input data for the program can be (i) ProtXML files that are introduced directly in TPP or (ii) spreadsheets (in csv file format) corresponding to spectral counts extracted from ProtXML files. We recommend the latter as it allows the user to curate ProtXML files to minimize problems associated with detecting multiple identifiers for the same protein (e.g. International Protein Index). Moreover, it enables implementation of spectral counting normalization methods prior to data analysis with PepC.

After the data are uploaded, PepC calculates the *G*-statistic (*G*-test) and *P*-value (*t*-test) for each protein in the biological comparison (group A versus B) and in the randomly permuted data. These measurements are subsequently used to quantify the number of proteins identified (presented as a heatmap) and the FDR at each combination of confidence intervals (Fig. 1). Using the slider control or Max FDR dialog box, the user sets the maximum FDR desired. PepC eliminates significance cutoffs with FDRs that exceed this value and resets the heatmap to display the combination of *G*-statistic and *P*-value that yields the maximal number of differentially expressed proteins (colored yellow). In this manner, the user can monitor the trade-off between the number of proteins identified and the FDR. Proteins that pass the chosen significance criteria are listed in the bottom panel along with their respective average spectral counts and statistical parameters.

Funding: National Institutes of Health (HG004537); Canadian Institutes of Health Research Fellowship Award (L.B.); Pilot and Feasibility Award (T.V.) from the Diabetes and Endocrinology Research Center.

Conflict of Interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Becker,L. *et al.* (2010) A macrophage sterol-responsive network linked to atherogenesis. *Cell Metab.*, **11**, 125–135.

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

Choi,H. *et al.* (2008) Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics*, **7**, 2373–2385.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.

Fu,X. *et al.* (2008) Spectral index for assessment of differential protein expression in shotgun proteomics. *J. Proteome Res.*, **7**, 845–854.

Liu,H. *et al.* (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.*, **76**, 4193–4201.

Mason,D.E. and Liebler,D.C. (2003) Quantitative analysis of modified proteins by LC-MS/MS of peptides labeled with phenyl isocyanate. *J. Proteome Res.*, **2**, 265–272.

Old,W.M. *et al.* (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics*, **4**, 1487–1502.

Pavelka,N. *et al.* (2008) Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol. Cell. Proteomics*, **7**, 631–644.

Sadygov,R.G. *et al.* (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods*, **1**, 195–202.

Sokal,R.R. and Rohlf,F.J. (1994). *Biometry: the Principles and Practice of Statistics in Biological Research*, 3rd edn. Freeman, New York.

Wiener,M.C. *et al.* (2004) Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Anal. Chem.*, **76**, 6085–6096.

Zhang,B. *et al.* (2006) Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.*, **5**, 2909–2918.

Zhu,W. *et al.* (2010) Mass spectrometry-based label-free quantitative proteomics. *J. Biomed. Biotechnol.*, **2010**, 840518.