Availability of MudPIT data for classification of biological samples

Di Silvestre Dario\textsuperscript{a}, Zoppis Italo\textsuperscript{b}, Brambilla Francesca\textsuperscript{a}, Valeria Bellettato\textsuperscript{a}, Mauri Giancarlo\textsuperscript{b}, Mauri Pierluigi\textsuperscript{a}

\textsuperscript{a} Institute for Biomedical Technologies, ITB-CNR Segrate (Milan), Italy.

\textsuperscript{b} Department of Informatics, Systems and Communication University of Milano-Bicocca, Milan, Italy.

Supplementary Information
• **Experimental details of MudPIT analysis**

Trypsin-digested samples were analyzed by MudPIT approach using a ProteomeX-2 system (Thermo Fisher Scientific, San José, CA, USA) implemented on a LTQ mass spectrometer. Digested peptide mixtures (5 µl) were first loaded onto a strong cation exchange column (Biobasic-SCX column, 0.32 i.d. x 100 mm, 5 µm, Thermo Electron Corporation, Bellefonte, PA, USA) and eluted stepwise with ammonium chloride salt injections of increasing molarity (0, 20, 40, 60, 80, 120, 200, 400, 600, 700 mM). Fractions were captured in turn onto two peptide traps (Zorbax 300 SB-C18, 0.3 i.d. x 5 mm, 5 µm, Agilent Technologies, Palo Alto, CA) for concentration and desalting prior to final separation by reversed-phase C18 column (Biobasic-18, 0.180 i.d. x 100 mm, 5 µm, Thermo Electron Corporation) with an AcN gradient (eluent A, 0.1% formic acid in water; eluent B, 0.1% formic acid in acetonitrile). Gradient profile was set to 5% eluent B for 5 min, 5% to 40% B in 45 min, 40% to 80% B in 10 min, 80% to 95% B in 5 min, 95% B for 10 min; flow-rate 1 µL/min. Finally, the eluted peptides were electrosprayed directly into a linear ion trap LTQ mass spectrometer equipped with a nano-LC electrospray ionization source (Thermo Finnigan, San Jose, CA, USA). Full mass spectra were acquired in positive mode, and over a 400-2000 m/z range, followed by five MS/MS events sequentially generated in a data-dependent manner on the top five most intense ions selected from the full MS spectrum scans using dynamic exclusion for MS/MS analysis.

• **Processing of MS/MS spectra by database search method based on SEQUEST algorithm (SEQUEST defined parameter thresholds)**

Experimental tandem mass spectra (MS/MS) produced by MudPIT runs were matched to in-silico tryptic peptide sequence as previously described (1–2). Briefly, data processing was performed by means of Biowork 3.3.1 software based on SEQUEST algorithm (3). Briefly, matches between spectra were only retained if they had a minimum Xcorr of 2.0 for +1, 2.5 for +2 and 3.5 for +3 charge state, respectively. In addition, maximum value for peptide/protein probability was determined to $10^{-3}$, while the minimum value for the SEQUEST-based SCORE was set to 10. Finally, the percentage of the false positives identification was estimated processing the raw mass spectra by means of decoy database (4), and the procedure revealed a false positives rate (FDR) less than 5%. Specifically, FDR was calculated by dividing the number of peptides identified by using the decoy database (reversed) for the number of peptides obtained by target one. Of course, in both cases were used the same SEQUEST defined parameter thresholds.

$$FDR = \frac{\text{n° peptide decoy database}}{\text{n° peptide target database}} \times 100$$
• Label-free quantification approaches

A direct correlation between the SEQUEST-based SCORE values and the relative abundance of the identified proteins has been previously demonstrated (9-10). Based on this findings, protein profiles of healthy and diseased samples were semi-quantitatively compared using a label-free proteomic approach, based on the application of DAve (Differential Average) and DCI (Differential Confidence Index) indices inserted in MAProMa (11). For the study purpose it was applied a value of DAve ≥ 0.3 (< -0.3), corresponding to a SCORE/SpC ratio ≥ 1.4. As previously demonstrated, it, coupled with a DCI threshold value ≥ 300 (< -300), allows to identify differentially expressed proteins with a good reliability (2, 9, 12).

To assess the accuracy of our criteria of biomarkers selection and to minimize false identification of differentially expressed proteins, spectral count (SpC) values were further evaluated by using the unpaired Student's t-test and the G-test, as previously described (13). In particular, G-test was calculated according to equation (1) (14):

\[
G = 2f_1 \ln \left( \frac{f_1}{\hat{f}_1} \right) + 2f_2 \ln \left( \frac{f_2}{\hat{f}_2} \right)
\] (1)

where \(f_1, f_2\) are the normalized SpC for the protein in sample 1 and sample 2, respectively; while \(\hat{f}_1, \hat{f}_2\) are expected SpC for the protein in sample 1 and sample 2, respectively. All employed statistical analyses were performed with a significance accepted at \(P > 95\%\).

• Rapid Miner

Rapid Miner (RM) is a software environment for rapid prototyping of machine learning and knowledge discovery processes. It is currently employed for classification (5), clustering (6), and also data integrations tasks (7). In RM, knowledge discovery is modeled by a complex nested chain of objects called operators. The user is supported with graphical interfaces where operators can be dropped as nodes onto the working pane and the data-flow is specified by connecting the operator nodes. In other words, RM workflows represent the conceptual sequence of operational steps employed for different data mining experiment.
Reference


