Corresponding author

Pierluigi Mauri

Tel. +39 02 26422728 Fax +39 02 26422770

Email address: pierluigi.mauri@itb.cnr.it

Institute for Biomedical Technologies (ITB-CNR) via F.lli Cervi 93, 20090 Segrate (Milan), Italy.

Availability of MudPIT data for classification of biological samples

Dario Di Silvestre[1] **, Italo Zoppis[2] **, Francesca Brambilla[1], Valeria Bellettato[1], Giancarlo Mauri[2] and Pierluigi Mauri*[1]

[1]Institute for Biomedical Technologies (ITB-CNR), via F.lli Cervi 93, Segrate (Milan), Italy.

[2]Department of Informatics, Systems and Communication, Viale Sarca 336, University of Milano-Bicocca, Milan, Italy.

Email: Dario Di Silvestre - dario.disilvestre@itb.cnr.it; Italo.zoppis - zoppis@disco.unimib.it; Francesca Brambilla - francesca.brambilla@itb.cnr.it; Valeria Bellettato - valeria.bellettato@itb.cnr.it; Giancarlo Mauri - mauri@disco.unimib.it; Pierluigi Mauri - pierluigi.mauri@itb.cnr.it;

*Corresponding author

** Equal contribution

Abstract

Background: Mass spectrometry is an important analytical tool for clinical proteomics. Primarily employed for biomarker discovery, it is increasingly used for developing methods which may help to provide unambiguous

diagnosis of biological samples. In this context, we investigated the classification of phenotypes by applying support vector machine (SVM) on experimental data obtained by MudPIT approach. In particular, we compared the performance capabilities of SVM by using two independent collection of complex samples and different data-types, such as mass spectra (m/z), peptides and proteins.

Results: Globally, protein and peptide data allowed a better discriminant informative content than experimental mass spectra (overall *accuracy* higher than 87% in both collection 1 and 2). These results indicate that sequencing of peptides and proteins reduces the experimental noise affecting the raw mass spectra, and allows the extraction of more informative *features* available for the effective classification of samples. In addition, proteins and peptides *features* selected by SVM matched for 80% with the differentially expressed proteins identified by the MAProMa software.

Conclusions: These findings confirm the availability of the most label-free quantitative methods based on processing of spectral count and SEQUEST-based SCORE values. On the other hand, it stresses the usefulness of MudPIT data for a correct grouping of sample phenotypes, by applying both supervised and unsupervised learning algorithms. This capacity permit the evaluation of actual samples and it is a good starting point to translate proteomic methodology to clinical application.

Keyword: Sample Classification, MudPIT, SVM, Clinical proteomics, Label-free quantification

Background

The identification of proteins changing their quantitative level is a key aspect to investigate biological systems as well as to develop strategies for classifying samples into pre-specified categories, such as healthy and diseased. In fact, one of the main objectives of the clinical proteomics is to use relevant biomarkers for improving disease diagnosis or for monitoring the efficacy of treatments [1].

A procedure for discriminating biological samples involves a preliminary evaluation of experimental data, useful for building classification models [2]. In this context, a wide variety of algorithms has been used for processing raw mass spectra, mainly generated by MALDI [3–10] and SELDI technologies [11–14]. Although results from diagnostic studies based on SELDI have generated both excitement and scepticism, it doesn't allows a direct identification of proteins and it is based on m/z signals, only. On the other hand, MALDI is mainly used for the identification of peptides and its reproducibility is strongly dependent by sample preparation method. Besides, in many studies, selected discriminant mass spectrometry signals have then

been identified by liquid chromatography (LC) coupled to mass spectrometry (MS). Nevertheless, few works have directly taken into consideration LC-MS data for discriminating biological samples [15, 16]. On the contrary, some authors have used them, combined to machine learning algorithms, for improving tandem mass (MS/MS) spectra quality assessment and hence, the protein identification [17–20].

Recently, the improvement of robustness and reproducibility of the MudPIT (Multidimensional Protein Identification Technology) approach, based on two dimensional liquid chromatography coupled to tandem mass spectrometry, has permitted a correct grouping of phenotypes, by using unsupervised algorithms [21–23]. Based on these findings, MudPIT may represent an attractive methodology for improving methods concerning sample classification. It allows to automatically obtain thousands of *features* comprising spectra, peptide sequences and related proteins [24, 25]. In addition, label-free quantification approaches based on spectral count (SpC) or SEQUEST-based SCORE evaluation permit an high-throughput discovering of multiple biomarkers [26–28], which could contain a higher level of discriminatory information.

The present study investigates in-depth the availability of MudPIT data for the classification of biological samples. We focused on classification performances achievable by processing different data-types, such as spectra, peptides and proteins. Specifically, we applied a class of machine learning algorithms, i.e. Support Vector Machine (SVM), to identify most predictive *features* and to score the data-types according to the inference performances of the algorithm [29,30]. Finally, since the identification of *features* allowing a model of classification is a key challenge for high-dimensional data, we evaluated how the applied selection method correlates with an independent label-free quantification approach. Therefore, we measured the overlapping of the *features* selected by SVM with the differentially expressed proteins selected by means of the MAProMa software [31].

Methods

Data collections

For the study purpose, two pre-existing different collections of experimental data were used. They were previously obtained by MudPIT analysis of complex samples, such as adipose and cardiac tissues [23, 32]. Specifically, for collection 1 were considered 30 diseased and 11 healthy controls, while 18 diseased and 18 healthy controls were considered for collection 2 (Supplementary Figure S1). Experimental details of the MudPIT analysis are reported in Supplementary Information.

Data handling of MS results

Raw mass spectra (MS) produced by MudPIT were handled using MZmine software [33]. Peak detection was performed by the chromatogram builder module by using the Centroid algorithm. Each file containing MS spectra was processed individually and converted to pairs of m/z and intensity values by considering all data points above the specified noise level (e^3). Then, m/z data points were connected to form chromatograms. In particular, the minimum time span was set to 1 min, the minimum absolute height to e^3 and the m/z tolerance to 0.5. Finally, peak lists were aligned by Join aligner method applying a ranges of tolerance of 0.5 and 1 min for mass and retention time, respectively.

The experimental tandem mass spectra (MS/MS) were correlated to *in-silico* tryptic peptide sequences, and accordingly to parent proteins, by using a database search method based on the SEQUEST algorithm [34]. The validity of spectrum/peptide matching was assessed using SEQUEST defined parameter thresholds (Supplementary Information). Finally, protein and peptide lists obtained from each sample were handled and aligned using MAProMa software and an in-house R-script, respectively [31, 35].

In order to evaluate the reproducibility of the MudPIT approach, protein lists of technical replicates were aligned and then processed using a linear-regression-based analysis:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where:

$i = 1, ..., n$; with n = number of variables (proteins)

Y_i is the spectral count (SpC) value of the protein i in the first replicate analysis

X_i is the spectral count (SpC)) value of the protein i in the second replicate analysis

β_0 is the intercept of the regression line of the population

β_1 is the slope or gradient of the regression line of the population

u_i is the error term

Proteomic datasets

Each sample belonging to the collection 1 and 2 was represented by five different datasets, including the global protein/peptide profiles and m/z precursor ions from three different chromatographic steps (60, 120, 400 mM) of the applied analytical method (Supplementary Information). Each dataset was formatted in a $s \times f$ matrix, where s represents the number of samples and f the number of *features*. Entries of the protein

data matrix were the spectral count (SpC) values assigned by the SEQUEST algorithm to each identified protein; in the same way, Xcorrelation values and peak area intensity (AUC) were used for the peptide and mass spectra data matrices, respectively (Supplementary Table S1).

Label-free quantification approach

Proteins differentially expressed between the considered phenotype groups were identified by using a label-free quantification approach. In particular, SEQUEST-based SCORE values were processed by means of the DAve and DCI formulas, which are inserted in MAProMa software [31]. In addition, SpC values were evaluated by using the G-test [36] and the unpaired Student's t-test. In this scenario, proteins with DAve ≥ 0.3 (≤ -0.3) and DCI ≥ 300 (≤ -300), or statistical meaningful at least for one test ($P > 95\%$) were considered for the study purpose (Supplementary Information and Figure S2).

Evaluation procedures by SVM

In order to investigate on the classification performance achievable by the different data-types (spectra, peptides and proteins) we designed specific *Rapid Miner* workflows (RM-WF) mainly addressed to implement a class of algorithms widely used in the *machine learning* community, i.e., the Support Vector Machine (SVM) [30].

In our investigation we sequentially applied two main operational processes i.e., feature selection and model construction (and validation), respectively. We briefly summarize in the following issues the RM-WF designed for each phase (a complete description of each operator is reported in Supplementary Figure S3).

1. **Feature selection phase.** Due to the high number of signals, *features selection* may be helpful to improve both the inference quality and the data understanding. For this reason we first applied a standard feature selection procedure [29]. Broadly speaking we weighted each signal by an information theory criterion (i.e., *info–gain ratio* [37]). Then we considered to employ in the forward phase only signals having a weight greater than 0.6; this way, only 10 signals were considered. The RM-WF in this case is simple, providing only the *info–gain* weighting capability as reported in Supplementary Figure S3 (a).

2. **Model construction and validation phase.** To evaluate the classification performance achievable by the different data-types, we employed SVM algorithms as "black boxes" to score each input data-type, according to the inference performances of the algorithm [29]. In order to avoid *over-fitting* we

first sub-sampled a set of different data instances: i.e., for each data set, this phase was applied on (data) instances never used in the above feature selection step. Then, for each instance, we considered only intensity (and counting) values corresponding to the previously suggested 10 signals (i.e., feature selection). This approach has been applied together with an optimization procedure to learn the algorithm parameters. As a matter of fact, different learning model may have many parameters, and often it is not clear which values are best for the learning task at hand; in our case, SVMs involve different *kernel* types and, in turn, each of such functions uses specific values which we need to define in the learning algorithms [30]. In order for the SVMs to perform as better (and homogeneous) as possible for each data-type, we optimized such parameters over the same space of common values. That is, we searched the best parameter values (i.e., providing the highest SVM inference performances) among all the combinations of common ranges for each input data collection. The RM-WF reported in Supplementary Figure S3 (b) specifies the main steps used in this phase.

Finally, standard indices (*i.e.*, sensitivity, specificity, positive (PPV), negative predictive (NPV), accuracy, F-score, balanced accuracy, informedness and Matthews correlation coefficient) were used as performance measures to verify which data-types provide the best SVM classification [2].

Results and Discussion

In this study, we investigated the classification of phenotypes by applying support vector machine (SVM) algorithms on experimental data obtained by MudPIT approach (Figure 1). Identified proteins, peptides and experimental mass spectra (m/z) were processed to evaluate the generalization capability of SVM about the *disease vs. healthy* cases used in this study (Supplementary Figure S1). For this purpose, a RapidMiner workflow was implemented (Supplementary Figure S3). Firstly, a set of data was used as input to SVM learning algorithm. Some learning parameters were optimized over the same common space of values [30]. Finally, data were evaluated according to the inference performance of the algorithm by using standard indices broadly applied to measure the precision and the recall capability [2].

By applying a standard *features* selection procedure, ten *features* having a weight greater than 0.6 were selected from each dataset (see features selection phase in Materials and Methods). Model delivered by the SVM operator was applied on independent validation datasets for estimating the performances concerning the phenotype classification. Tables 1 and 2, reporting the standard indices, show the diagnostic capabilities of SVM by using two independent collection of samples and different data-types. Of note, the results suggest

that SVM allows a better classification capability by using proteins and peptides rather than mass spectra datasets. In fact, better values of accuracy, F-score, informedness and MCC were observed by considering both collection 1 and collection 2. As opposite, samples classification by means of m/z data, resulted to be more difficult. In particular, by using the mass spectra of the collection 1 low values of specificity were observed, while the mass spectra of the collection 2 allowed low overall classification accuracy values.

The different classification performances, obtained by SVM, may be related to the m/z data complexity. In this regard, an overview of the data was performed by means of Principal Component Analysis (PCA) [38]. As opposed to protein and peptide data, PCA showed that mass spectra, especially for the collection 1, didn't allow a clear differentiation in the multidimensional space between disease and healthy groups (Supplementary Figure S4). In this context, the great amount of mass spectra can make it difficult their data-mining. In fact, a single step of liquid chromatography separation allows the collection of a number of *features* (m/z values) about 15 and 3 times bigger than protein and peptide ones, respectively (Figure 2). This great amount of data may be due to the redundant acquisition of spectra, like so to the biological and/or chemical modifications of peptides/proteins (e.g. Post Translational Modifications). Moreover, m/z values may be affected by chemical noise as well as to day-to-day instrument variations. Therefore, preprocessing of the raw data significantly influences the quality of the classification results [39, 40]. Nevertheless, further errors may be introduced during spectra alignment, while overlapping of m/z regions may create ambiguities for peak detection leading to increase the noise and to loss of information and discriminatory ability.

The identification of peptides and proteins by means of the interpretation of tandem mass spectra, can represent a cleaning and a simplifying of m/z data complexity. This aspect probably improved the *features* selection process and consequently the performance of classification by means of SVM model. For each collection about 20% of the selected features resulted common between protein and peptide datasets. Besides, around 80% of proteins and peptides, selected by SVM, matched with the differentially expressed proteins selected by MAProMa software (Figure 3). This correspondence represents a mutual validation of these two different procedures and it means that differentially expressed proteins may be used also for a correct grouping of sample phenotypes. For this reason, the use statistical parameters associated with identified proteins and peptides represents a robust procedure for a rapid extraction of potential biomarkers. In addition, MudPIT approach allows a good analytical reproducibility (Figure 4). In fact, although only 60-80 % of protein are identified in two replicate analyses, most of the variation is due to low abundance proteins which are usually identified with a low number of peptides. However, a statistical model has been proposed for estimating the number of replicates required for saturated sampling of a complex protein mixture [41].

Our findings are in good agreement with the most widely used semi-quantitative methods concerning the identification of biomarkers using LC-MS/MS approach [25, 27]. As for the identification of clinically useful biomarkers, in the last decade, SELDI-TOF analysis has been widely used and many diseases have been mainly studied by serum/plasma protein profiling. Although preliminary results have generated a lot of expectations, later scepticism resulted prevalent [42]. The reasons of this failure is probably due to SELDI profiling based on m/z signals, only, and it doesn't permits a direct identification and quantification of peptides/proteins. In addition, blood samples, although relatively simple to be collected, have a very complex composition with the presence of prominent and unspecific changes, resulting a drawback for the biomarker discovery based on m/z signals. On the contrary, we have evidenced in the present manuscript the improved availability of peptide/protein outcomes to allow biomarker discovery and phenotype discrimination. In comparison to mass spectra, sequenced proteins and peptides are less affected by experimental errors, and their use can be useful to avoid the problems of reproducibility due to different instrumental settings occurring over time. In addition, model of healthy/disease tissues represents a source of biomarkers in higher concentration than to plasma, which may be considered mainly useful in their monitoring using other LC-MS procedures [43].

Conclusion

To realize the potential of MS-based proteomics in the context of clinical utility, for disease diagnosis and prognosis, comparative studies are of great importance. In the present work, MudPIT data, both experimental mass spectra and sequenced peptides/proteins, were processed by SVM for evaluating the corresponding performances of classification. The overall *accuracy* resulted in all investigated cases higher than 77%. In particular, protein/peptide allowed a better discriminant informative content than experimental mass spectra (overall *accuracy* higher than 87% in both collection 1 and 2). This result is probably due to the translation of mass spectra to peptides/proteins, that eliminates the experimental noise and highlight the actual *features* useful for the phenotype classification. Overall, the presented findings indicate that the impressive amount of data produced by MudPIT approach can be processed for identifying multiple biomarkers and for classifying biological samples, by applying both supervised and unsupervised algorithms. These procedures permit the evaluation of actual samples and translate proteomic methodology to clinical application. In this context, MudPIT approach can be a useful tool for improving the extraction of informative *features* and therefore diagnosis procedures. Probably, in the next future new and more efficient algorithms will be applied, and the

discovered biomarkers will be validated by means of fast and high-resolution mass spectrometry and data independent analysis [44, 45]. These aspects will be of primary importance to be combined with clinical data and for investigating mechanisms of pathogenesis. In fact, the improved quality of data has the potential to optimize existing protein quantification methods and address the increasing demand of systems biology studies for correlating molecular expression to biological processes.

List of abbreviations used

- SVM (Support Vector Machine)
- MudPIT (Multidimensional Protein Identification Technology)
- MAProMa (Multidimensional Algorithm Protein Map)
- MALDI (Matrix-Assisted Laser Desorption/Ionization)
- SELDI (Surface-Enhanced Laser Desorption/Ionization)
- LC (Liquid Chromatography)
- MS (Mass Spectrometry)
- MS/MS (Tandem Mass Spectra)
- SpC (Spectral Count)
- DAve (Differential Average)
- DCI (Differential Confidence Index)
- RM-WF (Rapid Miner - WorkFlow)
- PPV (Positive Predicted Value)
- NPV (Negative Predicted Value)
- MCC (Matthews Correlation Coefficient)
- PCA (Principal Component Analysis)

Competing interests

The authors declare that they have no competing interests.

Author's contributions

DDS and IZ carried out the processing of the experimental data, the interpretation of the results and wrote the manuscript; FB performed the MudPIT experiments; VB carried out the administrative management, GM and PLM conceived the project, participated in the design of the study and in the writing of the manuscript.

Author's information

DDS (PhD), Permanent Researcher at the Proteomic and Metabolomic Department of the Institute of Biomedical Technologies - National Research Council (ITB-CNR), located in Segrate (Milan), Italy. His skills include the bioinformatic processing of proteomic data and their functional characterization by using data-derived systems biology approach.

IZ (PhD), Permanent Researcher at the Computer Science Department of the University of Milano-Bicocca, located in Milan (Italy). His skills include the machine learning applications to bioinformatics and computational systems biology.

FB (PhD), Researcher at the Proteomic and Metabolomic Department of the Institute of Biomedical Technologies - National Research Council (ITB-CNR), located in Segrate (Milan), Italy. His skills include the high-throughput proteomic analysis of complex biological samples by means of the MudPIT methodology.

VB, Technician at the Proteomic and Metabolomic Department of the Institute of Biomedical Technologies - National Research Council (ITB-CNR), located in Segrate (Milan), Italy.

GM (PhD), Full Professor and Director at the Computer Science Department of the University of Milano-Bicocca, located in Milan (Italy).

PLM (PhD), Principal Investigator at the Proteomic and Metabolomic Department of the Institute of Biomedical Technologies - National Research Council (ITB-CNR), located in Segrate (Milan), Italy.

Acknowledgements

This study was supported by the Italian Ministry of Economy and Finance to the CNR for the Project 'FaReBio di Qualita', by Italian Ministry of University and Research for the Project FAR Milano-Bicocca 2011-2012 and PON (01 02388), and by Fondazione Cariplo (2010-0653 and 2009-3149). The authors thank M.G. Bitonti for the availability of MAProMa software.

References

1. Palmblad M, Tiss A, Cramer R: **Mass spectrometry in clinical proteomics - from the present to the future.** *Proteomics Clin Appl* 2009, **3**:6–17, [http://dx.doi.org/10.1002/prca.200800090].

2. Ressom HW, Varghese RS, Zhang Z, Xuan J, Clarke R: **Classification algorithms for phenotype prediction in genomics and proteomics.** *Front Biosci* 2008, **13**:691–708.

3. Frenzel J, Gessner C, Sandvoss T, Hammerschmidt S, Schellenberger W, Sack U, Eschrich K, Wirtz H: **Outcome prediction in pneumonia induced ALI/ARDS by clinical features and peptide patterns of BALF determined by mass spectrometry.** *PLoS One* 2011, **6**(10):e25544, [http://dx.doi.org/10.1371/journal.pone.0025544].

4. Sampson DL, Parker TJ, Upton Z, Hurst CP: **A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches.** *PLoS One* 2011, **6**(9):e24973, [http://dx.doi.org/10.1371/journal.pone.0024973].

5. Waloszczyk P, Janus T, Alchimowicz J, Grodzki T, Borowiak K: **Proteomic patterns analysis with multivariate calculations as a promising tool for prompt differentiation of early stage lung tissue with cancer and unchanged tissue material.** *Diagn Pathol* 2011, **6**:22, [http://dx.doi.org/10.1186/1746-1596-6-22].

6. Rajalahti T, Kroksveen AC, Arneberg R, Berven FS, Vedeler CA, Myhr KM, Kvalheim OM: **A multivariate approach to reveal biomarker signatures for disease classification: application to mass spectral profiles of cerebrospinal fluid from patients with multiple sclerosis.** *J Proteome Res* 2010, **9**(7):3608–3620, [http://dx.doi.org/10.1021/pr100142m].

7. Camaggi CM, Zavatto E, Gramantieri L, Camaggi V, Strocchi E, Righini R, Merina L, Chieco P, Bolondi L: **Serum albumin-bound proteomic signature for early detection and staging of hepatocarcinoma: sample variability and data classification.** *Clin Chem Lab Med* 2010, **48**(9):1319–1326, [http://dx.doi.org/10.1515/CCLM.2010.248].

8. Kim HK, Reyzer ML, Choi IJ, Kim CG, Kim HS, Oshima A, Chertov O, Colantonio S, Fisher RJ, Allen JL, Caprioli RM, Green JE: **Gastric cancer-specific protein profile identified using endoscopic biopsy samples via MALDI mass spectrometry.** *J Proteome Res* 2010, **9**(8):4123–4130, [http://dx.doi.org/10.1021/pr100302b].

9. Balog CIA, Alexandrov T, Derks RJ, Hensbergen PJ, van Dam GJ, Tukahebwa EM, Kabatereine NB, Thiele H, Vennervald BJ, Mayboroda OA, Deelder AM: **The feasibility of MS and advanced data processing for monitoring Schistosoma mansoni infection.** *Proteomics Clin Appl* 2010, **4**(5):499–510, [http://dx.doi.org/10.1002/prca.200900158].

10. Chinello C, Gianazza E, Zoppis I, Mainini V, Galbusera C, Picozzi S, Rocco F, Galasso G, Bosari S, Ferrero S, Perego R, Raimondo F, Bianchi C, Pitto M, Signorini S, Brambilla P, Mocarelli P, Galli Kienle M, Magni F: **Serum Biomarkers of Renal Cell Carcinoma Assessed Using a Protein Profiling Approach Based on ClinProt Technique**. *Urology* 2010, **75**(4):842–847.

11. Lin Q, Peng Q, Yao F, Pan XF, Xiong LW, Wang Y, Geng JF, Feng JX, Han BH, Bao GL, Yang Y, Wang X, Jin L, Guo W, Wang JC: **A classification method based on principal components of SELDI spectra to diagnose of lung adenocarcinoma.** *PLoS One* 2012, **7**(3):e34457, [http://dx.doi.org/10.1371/journal.pone.0034457].

12. Fan Y, Wang J, Yang Y, Liu Q, Fan Y, Yu J, Zheng S, Li M, Wang J: **Detection and identification of potential biomarkers of breast cancer.** *J Cancer Res Clin Oncol* 2010, **136**(8):1243–1254, [http://dx.doi.org/10.1007/s00432-010-0775-1].

13. Tang KL, Li TH, Xiong WW, Chen K: **Ovarian cancer classification based on dimensionality reduction for SELDI-TOF data.** *BMC Bioinformatics* 2010, **11**:109, [http://dx.doi.org/10.1186/1471-2105-11-109].

14. Van Gorp T, Cadron I, Daemen A, De Moor B, Waelkens E, Vergote I: **Proteomic biomarkers predicting lymph node involvement in serum of cervical cancer patients. Limitations of SELDI-TOF MS.** *Proteome Sci* 2012, **10**:41, [http://dx.doi.org/10.1186/1477-5956-10-41].

15. Wiesner C, Hannum C, Reckamp K, Figlin R, Dubridge R, Roy SM, Lin S, Becker CH, Jones T, Hiller J, Cheville JC, Wilson K: **Consistency of a two clinical site sample collection: a proteomics study.** *Proteomics Clin Appl* 2010, **4**(8-9):726–738, [http://dx.doi.org/10.1002/prca.200900206].

16. Gambin A, Szczurek E, Dutkowski J, Bakun M, Dadlez M: **Classification of peptide mass fingerprint data by novel no-regret boosting method.** *Comput Biol Med* 2009, **39**(5):460–473, [http://dx.doi.org/10.1016/j.compbiomed.2009.03.006].

17. Perez-Riverol Y, Audain E, Millan A, Ramos Y, Sanchez A, Vizcaino JA, Wang R, Müller M, Machado YJ, Betancourt LH, González LJ, Padrón G, Besada V: **Isoelectric point optimization using peptide descriptors and support vector machines.** *J Proteomics* 2012, **75**(7):2269–2274, [http://dx.doi.org/10.1016/j.jprot.2012.01.029].

18. Ding J, Shi J, Wu FX: **SVM-RFE based feature selection for tandem mass spectrum quality assessment.** *Int J Data Min Bioinform* 2011, **5**:73–88.

19. Webb-Robertson BJM: **Support vector machines for improved peptide identification from tandem mass spectrometry database search.** *Methods Mol Biol* 2009, **492**:453–460, [http://dx.doi.org/10.1007/978-1-59745-493-3_28].

20. Baczek T, Kaliszan R: **Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics.** *Proteomics* 2009, **9**(4):835–847, [http://dx.doi.org/10.1002/pmic.200800544].

21. Gaspari M, Verhoeckx KCM, Verheij ER, van der Greef J: **Integration of two-dimensional LC-MS with multivariate statistics for comparative analysis of proteomic samples.** *Anal Chem* 2006, **78**(7):2286–2296, [http://dx.doi.org/10.1021/ac052000t].

22. Sodek KL, Evangelou AI, Ignatchenko A, Agochiya M, Brown TJ, Ringuette MJ, Jurisica I, Kislinger T: **Identification of pathways associated with invasive behavior by ovarian cancer cells using multidimensional protein identification technology (MudPIT).** *Mol Biosyst* 2008, **4**(7):762–773, [http://dx.doi.org/10.1039/b717542f].

23. Simioniuc A, Campan M, Lionetti V, Marinelli M, Aquaro GD, Cavallini C, Valente S, Di Silvestre D, Cantoni S, Bernini F, Simi C, Pardini S, Mauri P, Neglia D, Ventura C, Pasquinelli G, Recchia FA: **Placental stem cells pre-treated with a hyaluronan mixed ester of butyric and retinoic acid to cure infarcted pig hearts: a multimodal study.** *Cardiovasc Res* 2011, **90**(3):546–556, [http://dx.doi.org/10.1093/cvr/cvr018].

24. Mauri P, Scigelova M: **Multidimensional protein identification technology for clinical proteomic analysis.** *Clin Chem Lab Med* 2009, **47**(6):636–646, [http://dx.doi.org/10.1515/CCLM.2009.165].

25. Yates JR, Ruse CI, Nakorchevsky A: **Proteomics by mass spectrometry: approaches, advances, and applications.** *Annu Rev Biomed Eng* 2009, **11**:49–79, [http://dx.doi.org/10.1146/annurev-bioeng-061008-124934].

26. Mauri P, Scarpa A, Nascimbeni AC, Benazzi L, Parmagnani E, Mafficini A, Della Peruta M, Bassi C, Miyazaki K, Sorio C: **Identification of proteins released by pancreatic cancer cells by multidimensional protein identification technology: a strategy for identification of novel cancer markers.** *FASEB J* 2005, **19**(9):1125–1127, [http://dx.doi.org/10.1096/fj.04-3000fje].

27. Park SK, Venable JD, Xu T, Yates JR 3rd: **A quantitative analysis software tool for mass spectrometry-based proteomics.** *Nat Methods* 2008, **5**(4):319–322, [http://dx.doi.org/10.1038/nmeth.1195].

28. Bergamini G, Di Silvestre D, Mauri P, Cigana C, Bragonzi A, De Palma A, Benazzi L, Döring G, Assael BM, Melotti P, Sorio C: **MudPIT analysis of released proteins in Pseudomonas aeruginosa laboratory and clinical strains in relation to pro-inflammatory effects.** *Integr Biol (Camb)* 2012, **4**(3):270–279, [http://dx.doi.org/10.1039/c2ib00127f].

29. I Guyon and S Gunn and M Nikravesh and L A Zadeh (Ed): *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer 2006.

30. Cristianini N, Schawe-Taylor J: *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press 2000.

31. Mauri P, Dehò G: **A proteomic approach to the analysis of RNA degradosome composition in Escherichia coli.** *Methods Enzymol* 2008, **447**:99–117, [http://dx.doi.org/10.1016/S0076-6879(08)02206-4].

32. Brambilla F, Lavatelli F, Di Silvestre D, Valentini V, Rossi R, Palladini G, Obici L, Verga L, Mauri P, Merlini G: **Reliable typing of systemic amyloidoses through proteomic analysis of subcutaneous adipose tissue.** *Blood* 2012, **119**(8):1844–1847, [http://dx.doi.org/10.1182/blood-2011-07-365510].

33. Pluskal T, Castillo S, Villar-Briones A, Oresic M: **MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.** *BMC Bioinformatics* 2010, **11**:395, [http://dx.doi.org/10.1186/1471-2105-11-395].

34. Ducret A, Van Oostveen I, Eng JK, Yates J 3rd, Aebersold R: **High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry.** *Protein Sci* 1998, **7**(3):706–719, [http://dx.doi.org/10.1002/pro.5560070320].

35. Di Silvestre D, Daminelli S, Brunetti P and Mauri PL: *Bioinformatics tools for mass spectrometry-based proteomics analysis.* Reviews in Pharmaceutical and Biomedical Analysis - BENTHAM SCIENCE PUBLISHERS 2010.

36. Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF: **Detecting differential and correlated protein expression in label-free shotgun proteomics.** *J Proteome Res* 2006, **5**(11):2909–2918, [http://dx.doi.org/10.1021/pr0600273].

37. Mitchell T: *Machine Learning.* McGraw-Hill Education 1997.

38. Jackson JE: *A Users' Guide to Principal Components.* Wiley: New York 1991.

39. Arneberg R, Rajalahti T, Flikka K, Berven FS, Kroksveen AC, Berle M, Myhr KM, Vedeler CA, Ulvik RJ, Kvalheim OM: **Pretreatment of mass spectral profiles: application to proteomic data.** *Anal Chem* 2007, **79**(18):7014–7026, [http://dx.doi.org/10.1021/ac070946s].

40. Zoppis I, Gianazza E, Borsani M, Chinello C, Mainini V, Galbusera C, Ferrarese C, Galimberti G, Sorbi S, Borroni B, Magni F, Antoniotti M, Mauri G: **Mutual Information Optimization for Mass Spectra Data Alignment**. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 2012, **9**(3):934–939.

41. Liu H, Sadygov RG, Yates JR 3rd: **A model for random sampling and estimation of relative protein abundance in shotgun proteomics.** *Anal Chem* 2004, **76**(14):4193–4201, [http://dx.doi.org/10.1021/ac0498563].

42. Albrethsen J, Bøgebo R, Møller CH, Olsen JA, Raskov HH, Gammeltoft S: **Candidate biomarker verification: Critical examination of a serum protein pattern for human colorectal cancer.** *Proteomics Clin Appl* 2012, **6**(3-4):182–189, [http://dx.doi.org/10.1002/prca.201100095].

43. Gallien S, Duriez E, Crone C, Kellmann M, Moehring T, Domon B: **Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer.** *Mol Cell Proteomics* 2012, **11**(12):1709–1723, [http://dx.doi.org/10.1074/mcp.O112.019802].

44. Bern M, Finney G, Hoopmann MR, Merrihew G, Toth MJ, MacCoss MJ: **Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry.** *Anal Chem* 2010, **82**(3):833–841, [http://dx.doi.org/10.1021/ac901801b].

45. Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, Aebersold R: **Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.** *Mol Cell Proteomics* 2012, **11**(6):O111.016717, [http://dx.doi.org/10.1074/mcp.O111.016717].

Figures

Figure 1 - MudPIT workflow.

Enzymatic digested peptides are first separated by Strong Cation Exchange (SCX), using steps of increasing salt concentration, followed by Reverse Phase (RP) chromatography, using an acetonitrile gradient. Eluted peptides are then directly analyzed by tandem mass spectrometry producing MS and MS/MS spectra. By specific algorithm, such as SEQUEST, and applying appropriate criteria of data filtering (see Supplemental Information), the comparison of experimental MS and MS/MS spectra with those *in-silico* predicted from a protein sequence database allows the characterization of the peptide sequences and the corresponding proteins, without limits of isoelectric point (pI), molecular weight (Mw) or hydrophobicity. Using MudPIT, five different datasets per sample were collected for the study purposes. Specifically, in addition to complete protein and peptide profiles, m/z data, corresponding to 60 mM, 120 mM, 400 mM of salt concentration steps, were mined collecting three different datasets of spectra.

Figure 2 - *Features* selected for the study purposes.

Number of *features* (m/z ions, peptides and proteins) collected analyzing, by MudPIT, all samples belonging to collection 1 and collection 2. DB1, DB2 and DB3 correspond to m/z data mined from 60 mM, 120 mM, 400 mM of salt concentration steps, respectively.

Figure 3 - DAve values for proteins and peptides selected by SVM.

DAve evaluates changes in protein expression and is defined as: $((X-Y)/(X+Y))/0.5$, while DCI, which describes the confidence of differential expression, is defined as: $(X+Y)*(X-Y))/2$, where X and Y represent the SEQUEST-based SCORE values (or spectral count) of a given protein in two compared samples. Conventionally, signs (+/-) of DAve (and DCI) indicate if proteins are up-regulated in the first or in the second sample, respectively.

Figure 4 - MudPIT repeatability

Linear regression analysis obtained by considering SpC values of proteins identified in two technical replicates of MudPIT analysis. R2 and Slope values resulted near to 1. Red rectangle highlights the proteins identified with a low number of peptides and which represent the portion of data less reproducible.

Tables

Table 1 - Performance of classification obtained by using SVM - Collection1

Specificity (Spec.), sensitivity (Sens.), positive predictive value (PPV), negative predictive value (NPV), accuracy (Acc.), F-score, balanced accuracy (Bal. Acc.), informedness and Matthews correlation coefficient (MCC) of collection 1. Evaluation capabilities have been obtained using observations not considered in the signal selection phase.

	Spec.	Sens.	PPV	NPV	Acc.	F-score	Bal. Acc.	Informedness	MCC
Proteins	75%	91%	75%	91%	87%	0.46	83%	67%	0.66
Peptides	75%	100%	100%	92%	94%	0.48	88%	75%	0.72
m/z-DB1	50%	96%	80%	85%	84%	0.45	72%	46%	0.43
m/z-DB2	75%	96%	86%	92%	90%	0.47	85%	71%	0.69
m/z-DB3	37%	100%	100%	83%	84%	0.45	68%	37%	0.34

Table 2 - Performance of classification obtained by using SVM - Collection2

Specificity (Spec.), sensitivity (Sens.), positive predictive value (PPV), negative predictive value (NPV), accuracy (Acc.), F-score, balanced accuracy (Bal. Acc.), informedness and Matthews correlation coefficient

(MCC) of collection 2. Evaluation capabilities have been obtained using observations not considered in the signal selection phase.

	Spec.	Sens.	PPV	NPV	Acc.	F-score	Bal. Acc.	Informedness	MCC
Proteins	93%	93%	93%	93%	93%	0.46	92%	85%	0.85
Peptides	100%	100%	100%	100%	100%	0.50	100%	100%	1
m/z-DB1	62%	92%	89%	71%	77%	0.40	77%	54%	0.47
m/z-DB2	77%	77%	77%	77%	77%	0.38	77%	54%	0.54
m/z-DB3	85%	85%	85%	85%	85%	0.42	84%	70%	0.69

Additional Files

1) Supplementary Information (PDF file format).

2) Supplementary Figure S1 (PNG file format) — Sample collections and related experimental data selected and used for the study purpose.

For each sample five different datasets were used. In addition to the global protein and peptide profiles, m/z precursor ions, specifically detected from the chromatographic steps corresponding to 60, 120 and 400 mM of ammonium chloride concentration, were considered. They cover the central part of the salt gradient elution range (0-700mM) and assure the identification of most of the peptides.

3) Supplementary Table S1 (PNG file format) — Matrix of high-dimensional proteomic data obtained analyzing sample by means of the MudPIT approach.

Rows represent features (e.g., m/z values, peptides or proteins), while columns indicate samples. In each cell it is reported a value corresponding to the parameter associated with feature. In particular, peak area intensity (AUC) was used for m/z mass points, Xcorrelation (Xcorr) values for peptides and spectral count (SpC) values for proteins.

4) Supplementary Figure S2 (PNG file format) — Venn diagram.

Venn diagram of differentially expressed proteins identified in collection 1 (A) and collection 2 (B). Evaluation of quantitative level was performed by applying DAve and DCI formulas, G-test and Student's t-test. In brackets is reported the number of proteins matching with the *features* selected by SVM.

5) Supplementary Figure S3 (PNG file format) — Rapid Miner workflow.

Rapid Miner WF for the Feature selection (a) and model construction/validation (b) phases. Blocks correspond to simple processes in the whole design: each operator receives an input and delivers an output to the forward operator. The function of each block is shortly reported as follow:

- **Input Operator** reads data from files.
- **Info Gain Weighting Operator** (Fig. a). Each signal is weighted by an information theory criterion (i.e., *info–gain ratio*). The forward phase (Fig. b) employees only signals having weight greater then 0.6;
- **Cross Validation Operator** encapsulates a cross validation (k–fold) process [37]: the input data set S is split up into subsets $\{S_1, S_2, ..., S_k\}$. The inner operators are applied k times using at each iteration i the set S_i as the test set and $S \backslash S_i$ as the training set.
- **Parameter Optimization Operator** In order for the SVMs to perform as better (and homogeneous) as possible for each datatype, we optimized the learning parameters over the same space of common values. That is, starting from common ranges (for every datatypes the same ranges of values are used) this operator finds the optimal combination (i.e., providing the highest SVM inference performance) of parameter values by using a *cross validation process*. Here, we briefly report the applied common ranges for the selected combinations (some documentation on Rapid Miner can be downloaded at http://rapid-i.com)
  - *SVM.kernel.type* $\in$ {ANOVA,DOT,POLYNOMIAL,RADIAL},
  - *SVM.kernel.degree* $\in \{2, \ldots, 6\}$,
  - *SVM.C*, *SVM.*$\epsilon \in \{1, 1.5, \ldots, 8\}$.
- **Training SVM Operator** implements a Support Vector Machine algorithm to deliver an inference model.
- **Model Applier Operator** applies the model delivered by the SVM operator.
- **Performance Operator** collects the performance evaluation of the classification task and outputs performance measures.

6) Supplementary Figure S4 (PNG file format) — Principal Component Analysis of peptide, protein and m/z, data of collection 1 and 2.

Overview of protein, peptide and mass spectra data matrices performed by Principal Component Analysis (PCA) (15). PCA was applied by RapidMiner software. High-dimensionality of each data matrix was preliminarily reduced by eliminating features identified with an identification frequency (IF) below a certain threshold. In detail, for protein and peptide datasets were retained features with IF >1, while concerning mass spectra datasets were retained features with IF >4. Finally, the principal components that account for most of the variation (PC1-PC2-PC3) in the original multivariate data were plotted in the multidimensional space.

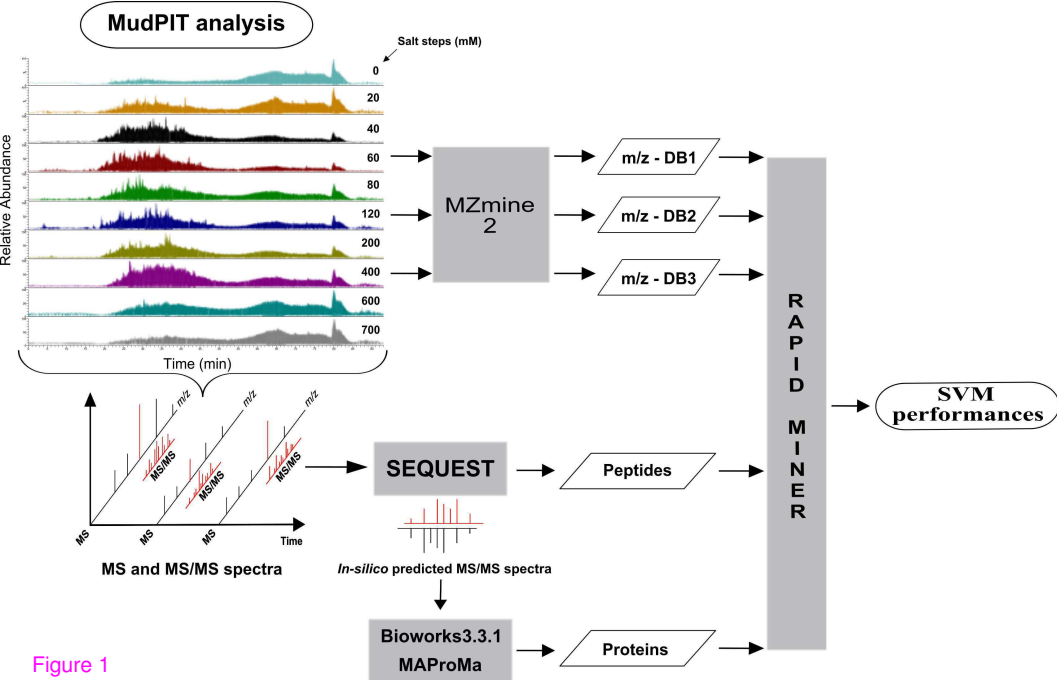
MudPIT analysis
Salt steps (mM)
0
20
40
60
80
120
200
400
600
700
Relative Abundance
Time (min)
MZmine 2
m/z - DB1
m/z - DB2
m/z - DB3
RAPID MINER
SVM performances
m/z
MS/MS
MS
Time
MS and MS/MS spectra
SEQUEST
Peptides
In-silico predicted MS/MS spectra
Bioworks3.3.1
MAProMa
Proteins

Figure 1

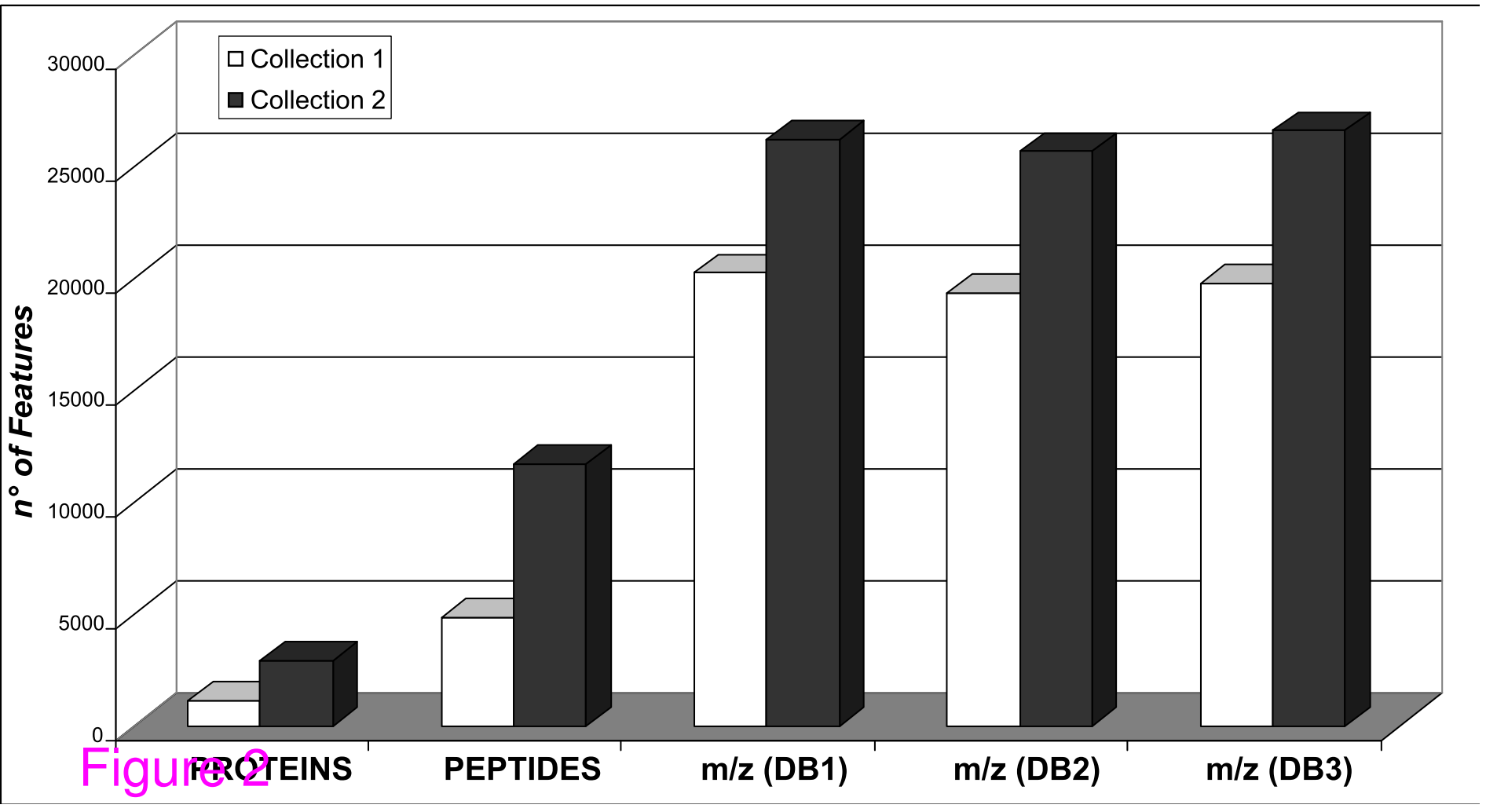
Collection 1
Collection 2
30000
25000
20000
15000
10000
5000
0
n° of Features
PROTEINS
PEPTIDES
m/z (DB1)
m/z (DB2)
m/z (DB3)

Figure 2

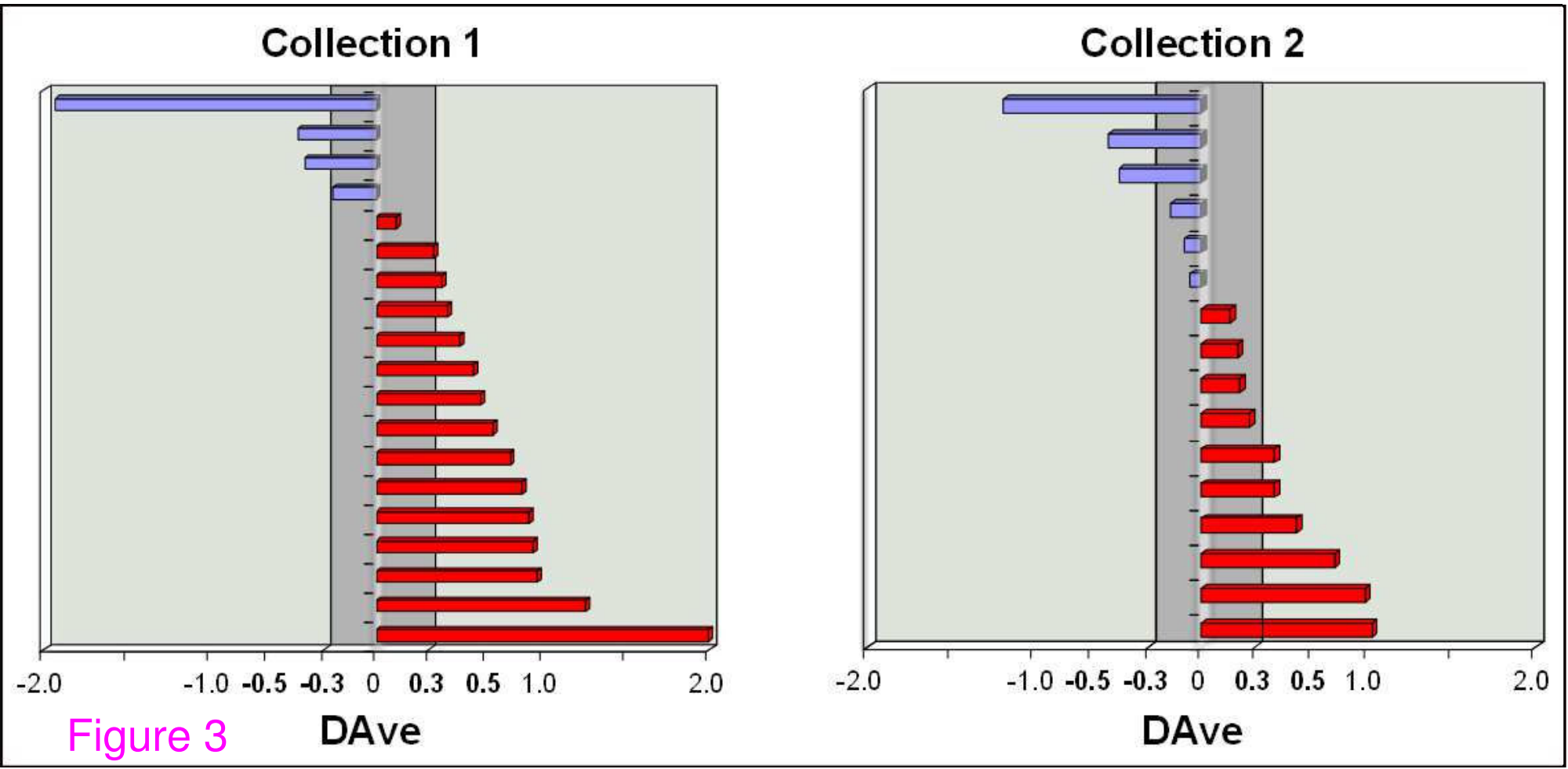
Collection 1
-2.0
-1.0
-0.5
-0.3
0
0.3
0.5
1.0
2.0
DAve
Collection 2
-2.0
-1.0
-0.5
-0.3
0
0.3
0.5
1.0
2.0
DAve

Figure 3

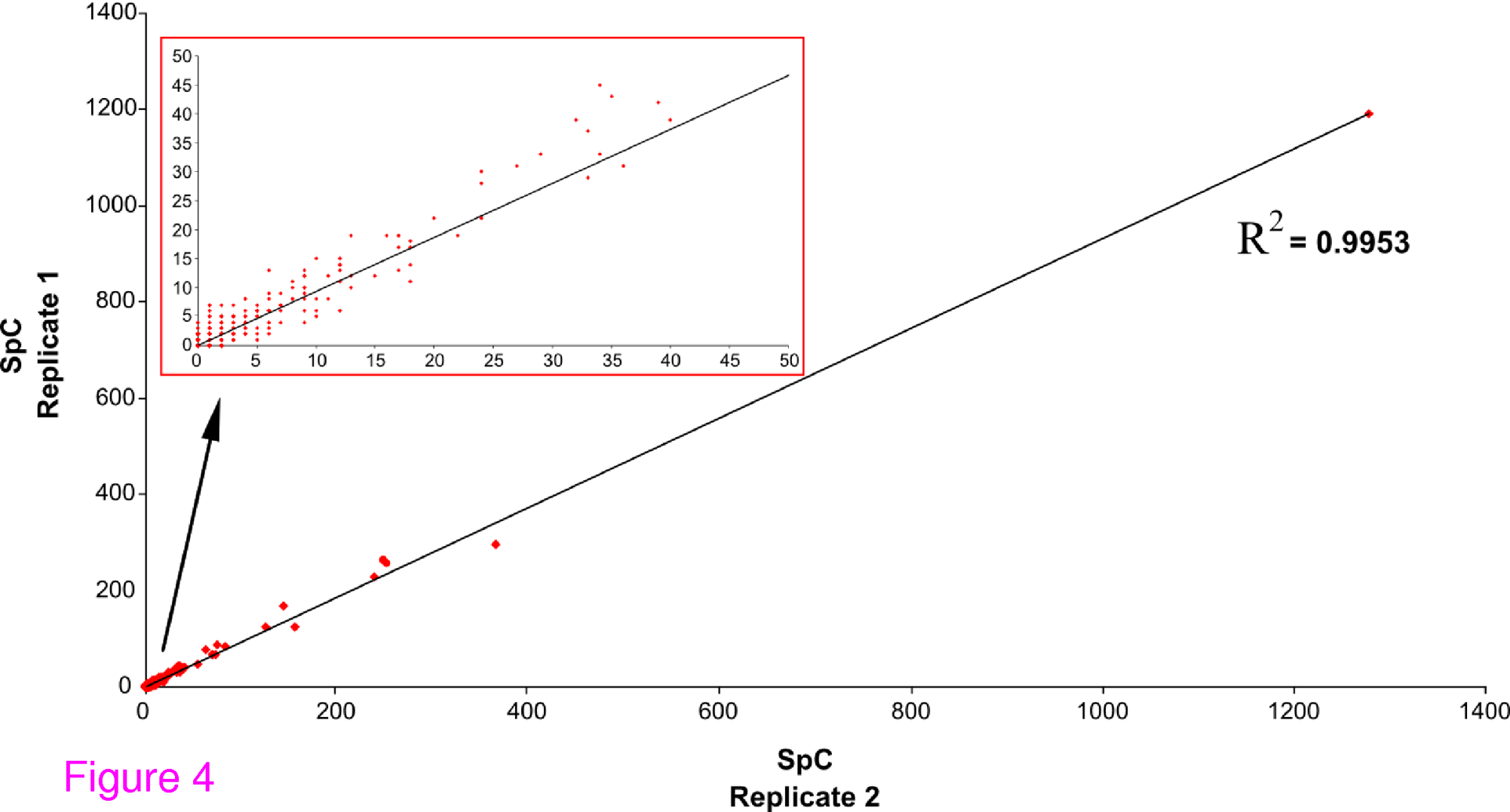
SpC
Replicate 1
SpC
Replicate 2
R^2 = 0.9953
0
200
400
600
800
1000
1200
1400
0
5
10
15
20
25
30
35
40
45
50

Figure 4

Additional files provided with this submission:

Additional file 1: Supplementary Information.pdf, 30K
http://www.jclinbioinformatics.com/imedia/1466050649871159/supp1.pdf
Additional file 2: Figure S1.png, 139K
http://www.jclinbioinformatics.com/imedia/1467280308871159/supp2.png
Additional file 3: Figure S2.png, 27K
http://www.jclinbioinformatics.com/imedia/1178584650871158/supp3.png
Additional file 4: Figure S3.png, 166K
http://www.jclinbioinformatics.com/imedia/2055743995871158/supp4.png
Additional file 5: Figure S4.png, 1597K
http://www.jclinbioinformatics.com/imedia/1306425980871158/supp5.png
Additional file 6: Table S1.png, 24K
http://www.jclinbioinformatics.com/imedia/7346580278711644/supp6.png