S2P: A Software Tool to Quickly Carry Out Reproducible Biomedical Research Projects Involving 2D-Gel and MALDI-TOF MS Protein Data

Hugo López-Fernández[1,2,3*$], José E. Araújo[3,4*], Susana Jorge[3,4], Daniel Glez-Peña[1,2], Miguel Reboiro-Jato[1,2], Hugo M. Santos[3,4], Florentino Fdez-Riverola[1,2], José L. Capelo[3,4]

[1]ESEI - Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, Universidad de Vigo, 32004 Ourense, Spain; [2]CINBIO - Centro de Investigaciones Biomédicas, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310, Vigo, Spain; [3]UCIBIO-REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Portugal; [4]ProteoMass Scientific Society, Madan Parque, Rua dos Inventores, 2825-182, Caparica, Portugal

* These authors contributed equally to this work

$ Corresponding author

Abstract

Background and Objective: 2D-gel electrophoresis is widely used in combination with MALDI-TOF mass spectrometry in order to analyze the proteome of biological samples. For instance, it can be used to discover proteins that are differentially expressed between two groups (e.g. two disease conditions, case vs. control, etc.) thus obtaining a set of potential biomarkers. This procedure requires a great deal of data processing in order to prepare data for analysis or to merge and integrate data from different sources. This kind of work is usually done manually (e.g. copying and pasting data into spreadsheet files), which is highly time consuming and distracts the researcher from other important, core tasks. Moreover, engaging in a repetitive process in a non-automated, handling-based manner is prone to error, thus threatening reliability and reproducibility. The objective of this paper is to present S2P, an open source software to overcome these drawbacks. **Methods**: S2P is implemented in Java on top of the AIBench framework, and relies on well-established open source libraries to accomplish different tasks. **Results:** S2P is an AIBench based desktop multiplatform application, specifically aimed to process 2D-gel and MALDI-mass spectrometry protein identification-based data in a computer-aided, reproducible manner. Different case studies are presented in order to show the

usefulness of S2P. **Conclusions:** S2P is open source and free to all users at http://www.sing-group.org/s2p. Through its user-friendly GUI interface, S2P dramatically reduces the time that researchers need to invest in order to prepare data for analysis.

Keywords: protein identification, data processing, 2D-gel, MALDI-TOF-MS, LC-MS/MS, emPAI.

1. Introduction

2D-gel electrophoresis and mass spectrometry using matrix assisted laser desorption ionization coupled to time of flight analyzers (MALDI-TOF-MS), are widely used in conjunction in order to perform proteome analysis [1,2]. In brief, while the comparison of 2D-gels allows obtaining a set of differentially expressed spots, MALDI-TOF-MS allows identifying the proteins separated in such spots.

The scientific community is particularly interested in the challenging task of finding proteins that can be used to differentiate different conditions of health with the aim to aid in the diagnosis, prognosis and development of new targeted therapies [3–5] . In order to find such proteins, known as biomarkers, a typical experimental workflow combining 2D-gel and MALDI-TOF-MS can involve the following steps: (*i*) separating the proteins present in a complex proteome; (*ii*) comparing the 2D-gels across samples to obtain the spots that were found expressed differentially; (*iii*) excising such spots and treating them for protein identification; (*iv*) linking the protein identifications to the 2D-gel spots; and (*v*) performing different types of data analysis to discover the potential biomarkers and extract meaningful biological knowledge. Such workflow generates a large amount of data, which need to be processed before they can be properly analyzed. A considerable part of the aforementioned data processing is usually carried out manually by laboratory researchers (e.g. using text editors and spreadsheet software). However, such a repetitive and non-automated process presents important drawbacks: it is time consuming, it is error-prone, and it tends to lack reliability and reproducibility.

To overcome the aforementioned drawbacks we have developed the S2P software application (http://www.sing-group.org/s2p/), a free software that aims to help researchers overcome these tedious but necessary data processing steps.

S2P was created with a twofold purpose: to improve reproducibility and to save time. Currently, lack of reproducibility is a growing concern in science [6]. The S2P software aims to improve reproducibility by avoiding human errors due to manual

data processing. For instance, this issue has been particularly important in recent genomics bioinformatics, where gene name errors have been shown to be widespread in the scientific literature due to the use of Excel [7,8]. Through its user-friendly GUI interface, S2P dramatically reduces the time that researchers need to invest in order to prepare data for analysis. To the best of our knowledge there is currently no other application offering similar functionalities.

The usefulness of S2P is illustrated by three case studies. The first is a case study experiment that aims to establish a biomarker-based method to allow better diagnosis and monitoring of patients with bladder cancer. The second aims to develop a longitudinal study to unravel the evolution of proteome of the peritoneal dialysate with time, so that biomarkers and molecular profiles for diagnosis and prognosis can be obtained. Finally, the third case study, which demonstrates how S2P has been extended to support new types of data, shows how it can be used to determine the relative abundance of serum protein using Mascot emPAI quantification data.

2. Materials and methods

2.1 Case study datasets

2.1.1 Case study 1 dataset

The first case study uses a dataset composed of 14 patients plus 1 healthy group of 6 individuals. Plasma samples were collected from 7 anonymous patients diagnosed with bladder cancer, 7 anonymous patients diagnosed with lower urinary tract symptoms (LUTS) and 6 healthy individuals, following standard procedures. All patients and healthy volunteers were informed about the project and their consent was obtained in written form. The local ethics committee approved the study. This experiment was developed as a proof of concept to find potential biomarkers that allow differentiating bladder cancer from LUTS.

Once in the laboratory, the samples were centrifuged, the supernatant was then withdrawn, aliquoted and stored at -80 °C until analysis. Most abundant proteins (MAPs) in plasma can mask or interfere with the detection of proteins belonging to the low-abundance protein fraction [9]. To avoid this problem, protein equalization from plasma samples was performed with dithiothreitol, DTT, according to the protocol described by Warder et al. [10] with minor modifications as described by Fernández et al. [11] and Araújo et al. [12–14]. This process was performed with five

replicates for each patient. Then, the total protein content was determined using a Bradford protein assay [15].

Two dimensional gel electrophoresis separation was carried out in duplicate for each patient and for the healthy pool. The 2D-gels obtained for each patient and the pool of healthy volunteers were then compared using the Progenesis SameSpots software v4.0 (NonLinear Dynamics) to ascertain the differentially expressed proteins. All spots of interest were excised and subjected to in-gel protein(s) digestion and then to protein fingerprint identification by mass spectrometry using MALDI-TOF-MS [16]. Finally, S2P was used to process the spots data (i.e. differentially expressed spots) obtained with the SameSpots software as well as to analyze them along with the protein identifications obtained from Mascot. This procedure is explained in the Results and discussion section in more detail.

2.1.2 Case study 2 dataset

The second case study uses a dataset composed of ten patients receiving a peritoneal dialysis treatment. Peritoneal dialysis effluent (PDE) samples were collected from patients following a peritoneal equilibrium test in different time points. Table 1 summarizes the clinical information for each of the ten patients. Patients were enrolled in a longitudinal study at different stages of dialysis and were followed during the same period. Therefore, the number of samples taken every 6 months for each patient is not equal: there are four samples from patient P01, three samples from patients P02, P03 and P04 and two samples from each of the rest.

Table 1. Second case study patients.

Patient	Lab. Reference	Age	Gender	Time in PD (months/years)	Classification	Nº sampling (months)
P01	IE	71	F	25.4/2	Unknown	4 (1st, 7th, 12th, 24th)
P02	MIR	71	F	20.47/1.71	Diabetic Nephropathy	3 (1st, 7th, 19th)
P03	JM	54	M	19.47/1.62	Hypertensive Nephrosclerosis	3 (1st, 7th, 19th)
P04	VA	63	F	21/1.75	ANCA Vasculitis	3 (1st,7th, 19th)
P05	LP	61	M	18.83/1.57	Unknown	2 (1st, 7th)
P06	SL	39	F	8.3/0.69	Chronic Glomerulonephritis	2 (1st, 7th)
P07	ML	25	M	7.30/0.61	Obstructive Uropathy	2 (1st, 7th)
P08	MLC	70	F	7.6/0.64	ANCA Vasculitis	2 (1st,7th)
P09	MC	51	F	8.53/0.71	Unknown	2 (1st, 7th)
P10	JP	74	M	6.97/0.55	Hypertensive Nephrosclerosis	2 (1st, 7th)

All volunteers were informed about the project and their consent was obtained in written form. The local ethics committee approved the study. This experiment was

developed as a preliminary study with the aim of identifying and following changes in the peritoneal membrane at the molecular level by proteomics using longitudinal studies to unravel morphological and biochemical changes in the long-term PDE.

Once in the laboratory, the samples were centrifuged at 9,000g for 20 min at 4ºC, cell debris was discarded, and peritoneal dialysis effluent supernatant was aliquoted in 15 mL tubes and stored at -80ºC until use. After peritoneal dialysate concentration, the total protein content was determined using a Bradford protein assay [15].

Two dimensional gel electrophoresis separation was carried out in duplicate for each sample. All of the 2D-gels corresponding to the samples obtained for each patient were then compared using the Progenesis SameSpots software v4.0 (NonLinear Dynamics) to ascertain the differentially expressed proteins. All spots of interest were excised and subjected to in-gel protein(s) digestion and then to protein fingerprint identification by mass spectrometry using MALDI-TOF-MS [16]. Finally, S2P was used to process the spots data corresponding to each patient obtained with the SameSpots software, and then to analyze them along with the protein identifications obtained from Mascot. This procedure is explained in the Results and discussion section in more detail.

2.1.3 Case study 3 dataset

In the third case study, the blood serum of a patient with osteoarthritis, were depleted with DTT as described previously [11]. The serum proteins were then alkylated with IAA (2µL of IAA 600 mM prepared in Ammonium Bicarbonate 12.5 mM, incubated at room temperature for 45 min in the dark). Protein digestion was carried out using (*i*) the classic overnight method [17] and (*ii*) the ultrasonic accelerated method using the microplate horn assembly device [18]. Briefly, for protein digestion, four aliquots of 2 µg of serum proteins each were mixed with trypsin to a ratio 1:20 (Wt./Wt.). Two samples were digested overnight at 37 ºC while the other two samples were digested using the microplate horn assembly device with the following operating conditions: 25% of ultrasonic amplitude and 4 min of ultrasonic time in a pulsed mode (30 sec on, 15 sec off). After digestion, 1 µL of formic acid 50% (Vol./Vol.) were added to stop the enzymatic activity, and the digested samples were evaporated to dryness.

The LC-MS/MS analysis was carried out using an EASY-nLC II on-line coupled to an IMPACT HD (Bruker Daltonics) with a CaptiveSpray ion source (Bruker Daltonics). All samples were diluted to 40 ng/µL with 0.1% (v/v) aqueous formic acid (2 µg of digested protein + 50 µL of aqueous formic acid) before loading onto an EASY-nLC II equipped with an EASY-Column, 2cm, ID100µm, 5µm, C18-A1 (Thermo Fisher

Scientific) and an EASY-Column, 10cm, ID75µm, 3µm, C18-A2 (Thermo Fisher Scientific). Chromatographic separation was carried out using a linear gradient of 0–35% buffer B (90% Acetonitrile and 0.1% Formic Acid) at a flow rate of 300 nl/min over 120 min followed by a gradient of 35-90% buffer B in 15 minutes and an isocratic flow of 90% buffer B for 5 minutes. Total run time was 140 minutes. For each sample, two replicate injections were performed (200 ng loading per injection).

Raw data were processed in DataAnalysis 4.2 and subsequently exported to Protein-Scape 4.0 for automated protein identification. For protein identification, CID-MS2 spectra were first searched against the human subset of the Swiss-Prot database, using the Mascot search engine with the following parameters: (*i*) two missed cleavage; (*ii*) fixed modifications: carbamidomethylation (C); (*iii*) variable modifications: oxidation of methionine, Acetyl (Protein N-term), Glu->pyro-Glu (N-term E), Gln->pyro-Glu (N-term Q), (*vi*) peptide mass tolerance up to 20 ppm, (v) fragment mass tolerance 0.05 Da (vi) Adjust FDR 1%. The significance threshold was set to a minimum of 95% ($p \leq 0.05$).

Finally, Mascot reports containing emPAI quantification values were obtained using the “Export” option in Mascot Search Results, which creates a CSV file of each LC-MS/MS run (i.e. 8 files in this case). S2P was used to process these reports in order to determine the relative abundance of serum protein. This procedure is explained in the Results and discussion section in more detail.

2.2 Implementation

S2P v1.2.0 is implemented in Java and was constructed using the AIBench framework [19], which has been demonstrated to be suitable for rapid development of scientific applications [20–23]. The Graphical User Interface (GUI) was created in Java Swing using freely available extensions such as SwingX or GC4S. S2P also makes use of several well-established open-source libraries such as JFreeChart, charts4j, iText and the Apache Commons Mathematics library. For enhanced table visualization, the JSparklines [24] library was used.

The source code of the project is freely available at https://github.com/sing-group/S2P under a GNU GPL 3.0 License (http://www.gnu.org/copyleft/gpl.html). It is divided into three modules: (*i*) core, which contains the default implementation API; (*ii*) gui, which contains several reusable GUI components; and (*iii*) aibench, which contains a GUI application based on the AIBench framework.

3. Results

With the goal of showing the main features of *S2P* as well as its usefulness to analyze real data, this section shows how it has been used to process and analyze the three case study datasets presented.

3.1 Case study 1: bladder cancer biomarker discovery

Figure 1 illustrates the main S2P workflow with the five main steps where it was used to process this case study data: (1) to merge the SameSpots reports into a single table where all samples can be compared; (2) to design the MALDI plate; (3) to load and filter the Mascot identifications; (4) to link the Mascot identifications with their corresponding spots using the MALDI plate; and (5) to examine and analyze spots data along with Mascot identifications. All data needed to reproduce the steps explained below are available at http://www.sing-group.org/s2p/tutorial-cs-1.html, along with a detailed quick-start tutorial that guides users using S2P for the first time.

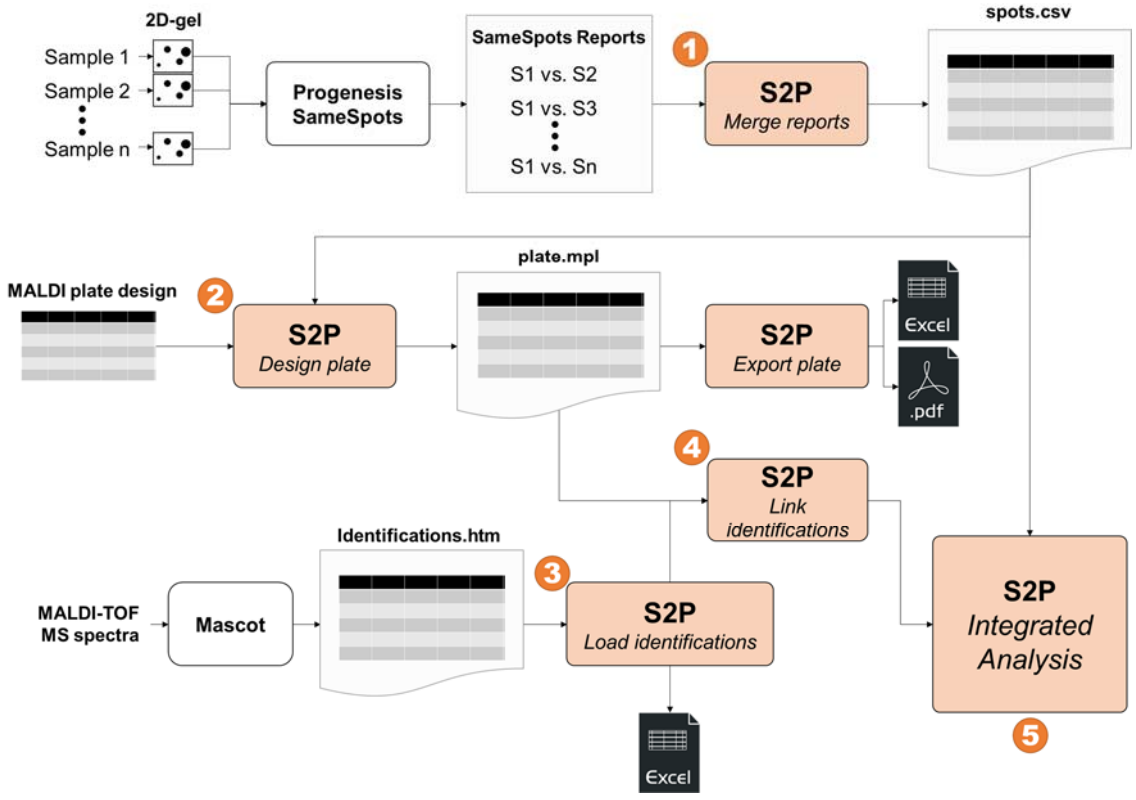


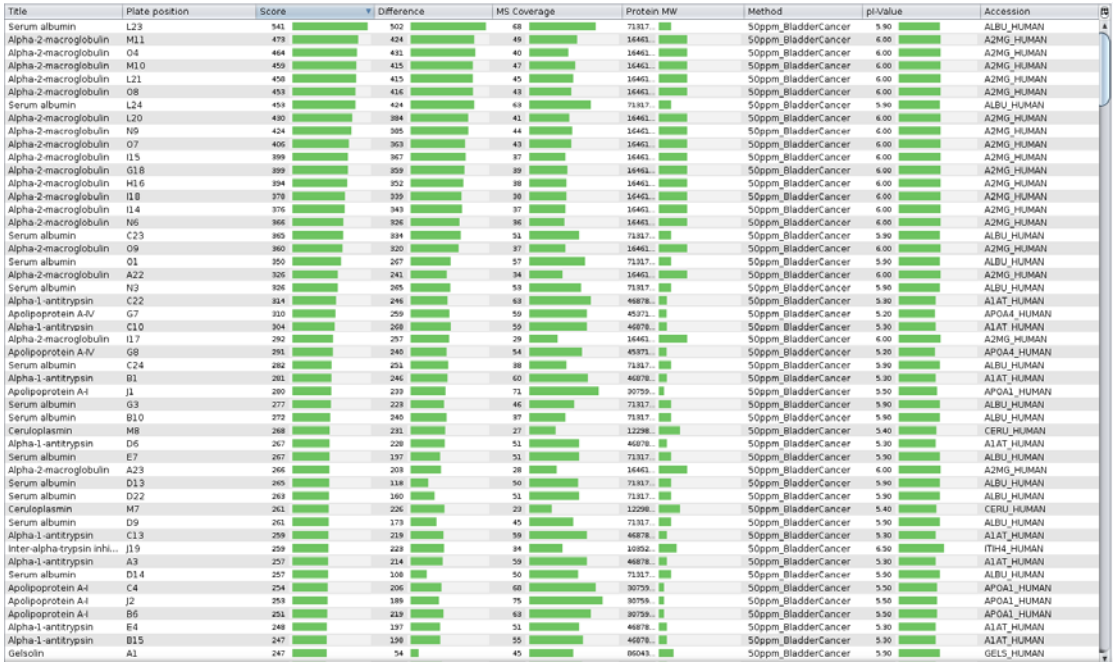
Figure 1. Schematic S2P flow diagram.

The case study dataset was composed of 7 anonymous patients diagnosed with bladder cancer, 7 anonymous patients diagnosed with lower urinary tract symptoms (LUTS) and 6 healthy individuals that were pooled. The Progenesis SameSpots software was used to align and compare the 2D-gels corresponding to each

individual against the health pool's 2D-gels to obtain the differentially expressed spots. These results were exported using the "Export report" option of SameSpots, which creates one HTML file per comparison (i.e. 14 files in this case). S2P was then used to parse and merge these reports into a single table with samples in columns and spots in rows (Step 1 of Figure 1). This table was exported into a comma-separated values (CSV) file that can be easily reopened with S2P as well as external applications such as Excel, LibreOffice or R.

These differentially expressed spots were first treated and then analyzed through MALDI-TOF MS in order to identify their protein content. To do so, a dedicated sample treatment is performed [16] and the pool of peptides obtained is spotted twice into a MALDI plate, which is then introduced into the MALDI apparatus for analysis. Researchers usually fill a sheet with the position of the spots in the plate so that they can trace back where each spot was placed. This is important as it allows researchers to know which spot is associated to each MALDI spectrum and, therefore, to know which Mascot identifications are associated to each spot. However, keeping a single handwritten copy of this key information is risky as it can be lost, damaged or misfiled and, most likely, there will be no way to recover this information. For these two reasons, S2P incorporates a MALDI plate editor that allows storing digital copies of experiment plates and printing them into PDF files (Step 2 of Figure 1). S2P also allows filling the plate automatically by using a set of previously loaded spots (Step 1 of Figure 1), thus permitting the user to define parameters such as matrix dimensions (i.e. number of rows and columns) or the number of replicates of each spot. In our case study, S2P was used to create the MALDI plate and to obtain a printed copy of it that was used to guide the experimental work.

Once the MALDI-TOF MS analysis was done, the MALDI-based spectra of the digested proteins were submitted to Mascot in order to identify the proteins. They were then exported into an HTML file that was loaded into S2P (Figure 2) in order to remove duplicated entries and exclude identifications with a Mascot score under 57 (Step 3 of Figure 1). This processed list of Mascot identifications was exported into a CSV file so that it can be directly loaded into S2P later or used in other applications (e.g. spreadsheet software). Next, these Mascot identifications were integrated with the spots data using the MALDI plate (Step 4 of Figure 1) to know which identifications are associated with each spot.



Title	Plate position	Score	Difference	MS Coverage	Protein MW	Method	pI-Value	Accession
Serum albumin	L23	541	502	68	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Alpha-2-macroglobulin	M11	473	424	49	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	O4	464	431	40	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	M10	459	415	47	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	L21	458	415	45	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	O8	453	416	43	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Serum albumin	L24	453	424	69	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Alpha-2-macroglobulin	L20	430	384	41	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	N9	424	385	44	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	O7	406	363	43	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	I15	399	367	37	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	G18	399	359	39	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	H16	394	352	39	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	I18	378	339	36	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	I14	376	343	37	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Alpha-2-macroglobulin	N6	366	326	36	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Serum albumin	C23	365	334	51	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Alpha-2-macroglobulin	O9	360	320	37	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Serum albumin	O1	350	267	57	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Alpha-2-macroglobulin	A22	326	241	34	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Serum albumin	N3	326	265	59	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Alpha-1-antitrypsin	C22	314	246	63	46878...	50ppm_BladderCancer	5.30	A1AT_HUMAN
Apolipoprotein A-IV	G7	310	259	59	45371...	50ppm_BladderCancer	5.20	APOA4_HUMAN
Alpha-1-antitrypsin	C10	304	268	59	46878...	50ppm_BladderCancer	5.30	A1AT_HUMAN
Alpha-2-macroglobulin	I17	292	257	29	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Apolipoprotein A-IV	G8	291	240	54	45371...	50ppm_BladderCancer	5.20	APOA4_HUMAN
Serum albumin	C24	282	251	38	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Alpha-1-antitrypsin	B1	281	246	60	46878...	50ppm_BladderCancer	5.30	A1AT_HUMAN
Apolipoprotein A-I	J1	280	239	71	30759...	50ppm_BladderCancer	5.50	APOA1_HUMAN
Serum albumin	G3	277	228	46	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Serum albumin	B10	272	240	37	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Ceruloplasmin	M8	268	231	27	12298...	50ppm_BladderCancer	5.40	CERU_HUMAN
Alpha-1-antitrypsin	D6	267	228	51	46878...	50ppm_BladderCancer	5.30	A1AT_HUMAN
Serum albumin	E7	267	197	51	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Alpha-2-macroglobulin	A23	266	203	28	16461...	50ppm_BladderCancer	6.00	A2MG_HUMAN
Serum albumin	D13	265	118	50	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Serum albumin	D22	263	160	51	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Ceruloplasmin	M7	261	226	23	12298...	50ppm_BladderCancer	5.40	CERU_HUMAN
Serum albumin	D9	261	173	45	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Alpha-1-antitrypsin	C13	259	219	59	46878...	50ppm_BladderCancer	5.30	A1AT_HUMAN
Inter-alpha-trypsin inhi...	J19	259	228	84	10952...	50ppm_BladderCancer	6.50	ITIH4_HUMAN
Alpha-1-antitrypsin	A3	257	214	59	46878...	50ppm_BladderCancer	5.30	A1AT_HUMAN
Serum albumin	D14	257	108	50	71317...	50ppm_BladderCancer	5.90	ALBU_HUMAN
Apolipoprotein A-I	C4	254	206	68	30759...	50ppm_BladderCancer	5.50	APOA1_HUMAN
Apolipoprotein A-I	J2	253	189	75	30759...	50ppm_BladderCancer	5.50	APOA1_HUMAN
Apolipoprotein A-I	B6	251	219	63	30759...	50ppm_BladderCancer	5.50	APOA1_HUMAN
Alpha-1-antitrypsin	E4	248	197	51	46878...	50ppm_BladderCancer	5.30	A1AT_HUMAN
Alpha-1-antitrypsin	B15	247	196	55	46878...	50ppm_BladderCancer	5.30	A1AT_HUMAN
Gelsolin	A1	247	54	45	86043...	50ppm_BladderCancer	5.90	GELS_HUMAN

Figure 2. Mascot identifications table shown in S2P, enhanced with the use of the JSparklines library.

Finally (Step 5 of Figure 1), S2P allows an integrated analysis of the spots data and the Mascot identifications (Figure 3). In the context of our case study, this option was first used to try to identify potential biomarkers of the two conditions of interest. When the healthy pool was compared with the bladder cancer patients, four differentially expressed spots present in at least 5 of 7 bladder cancer patients (Figure 4A) were found. The corresponding proteins were: (*i*) Serum albumin (Spot Number [SN]=137), (*ii*) Gelsolin (SN=137), (*iii*) Fibrinogen gamma chain (SN=337), (*iv*) Ig alpha-1 chain C region (SN=360), (*v*) Ig alpha-2 chain C region (SN=360) and (*vi*) Haptoglobin (SN=266). When the healthy pool was compared with the LUTS patients, we found five differentially expressed spots that were present in at least 4 of 7 LUTS patients (Figure 4B). The associated proteins were the following: (*i*) CD5 antigen-like (SN=244), (*ii*) Heparin cofactor 2 (SN=175 and SN=190), (*iii*) Hemopexin (SN=175), (*iv*) Serum albumin (SN=192 and SN=190) and (*v*) Inter-alpha-trypsin inhibitor heavy chain H4 (SN=88).

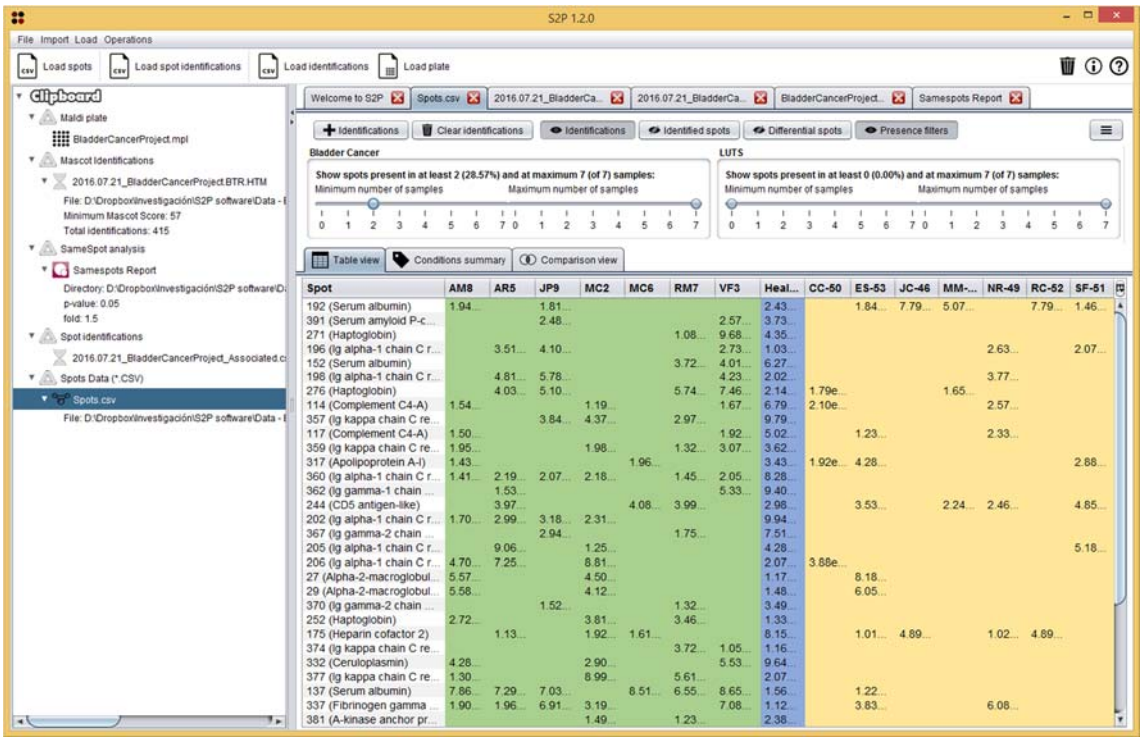


Figure 3. Screenshot of the S2P integrated analysis window.

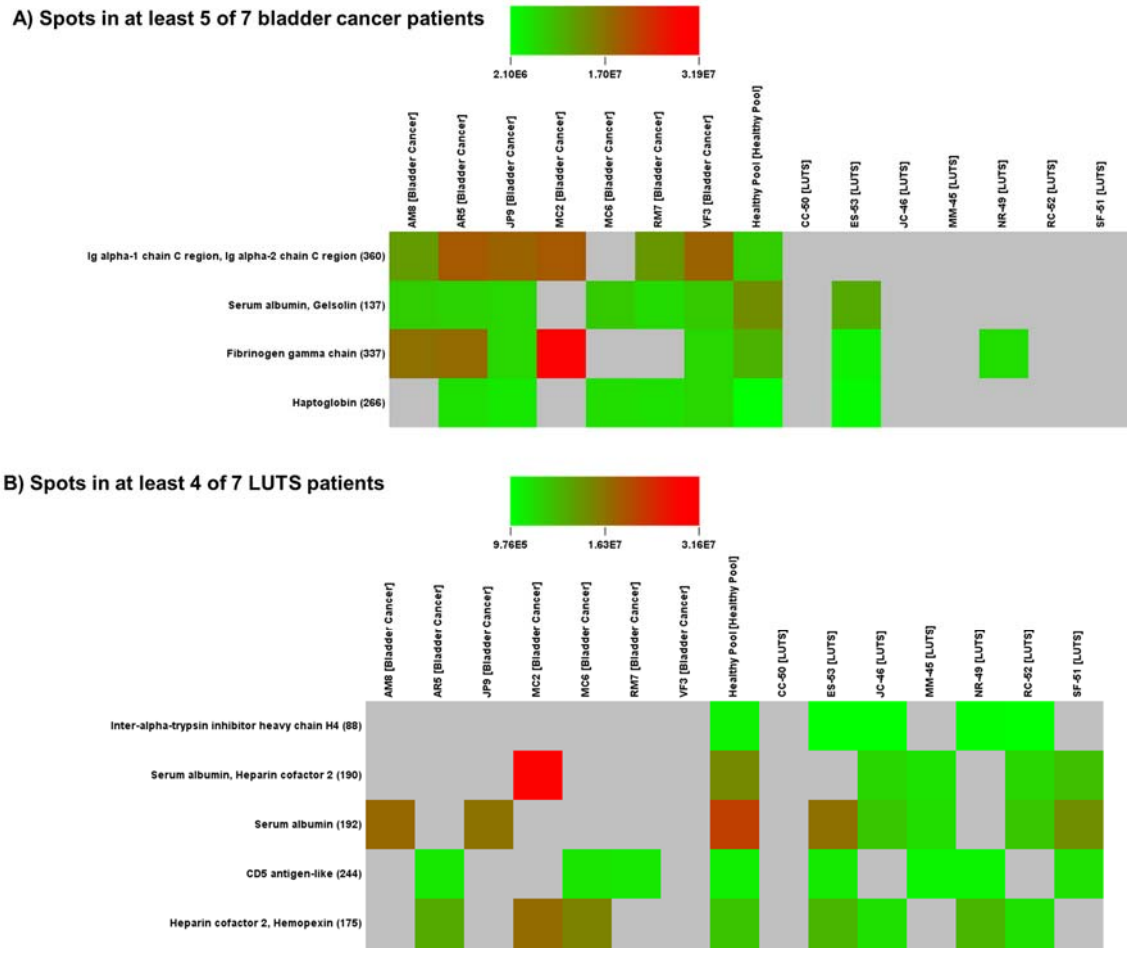


Figure 4. Heat maps showing the differentially expressed spots.

As seen in Figure 4, a small set of candidate biomarkers that can be associated to each disease was identified. Consequently, a complementary approach was experimented: exporting all spots data from SameSpots rather than exporting only those spots that were differentially expressed when each individual and the healthy pool were compared. To do so, we used S2P to process these new dataset (analogously to step 1) and then to find spots whose average value was statistically different between bladder cancer and LUTS patients. Following this strategy, 40 differentially expressed spots (i.e. having t-test p-values corrected using Benjamini-Hochberg, or q-values, less than 0.05) between bladder cancer and LUTS were found, 27 of which have associated protein identifications (corresponding to 14 unique proteins). This also allowed us to compare the distribution of the expression values of each condition using box plots. For instance, Figure 5 shows the box plots of the two spots identified in Figure 4 that are differentially expressed between bladder cancer and LUTS patients. This information must be carefully analyzed, but the usefulness of S2P to quickly and accurately process and analyze data is thus proven.

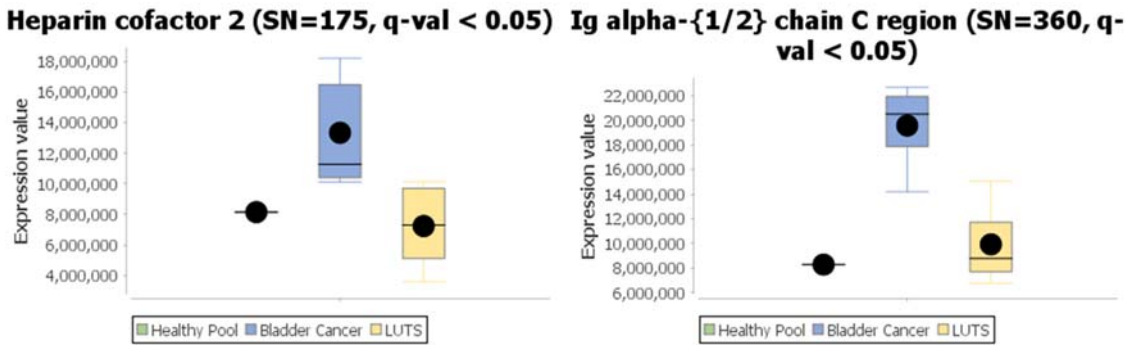


Figure 5. Box plots of the differentially expressed spots.

This case study is important because it led to the development of the initial version of S2P (v1.0.0). Other studies following the same workflow were subsequently carried out, allowing us to fix bugs and collect ideas for new features in later S2P versions. Nevertheless, a slightly different scenario appeared since, in short, it was necessary to process each sample individually instead of comparing all of them. Case study 2 shows how S2P was adapted to address this new requirement.

3.2 Case study 2: longitudinal monitoring of peritoneal dialysis effluent

The case study dataset was composed of 10 patients receiving a peritoneal dialysis treatment. As stated in the Materials and methods section, the number of samples taken every 6 months for each patient is not equal: there are four samples from

patient P01, three samples from patients P02, P03 and P04 and two samples from each of the rest. The Progenesis SameSpots software was used to compare the 2D-gels corresponding to each patient and obtain the differentially expressed spots between the different samples (i.e. time points). These results were exported using the “Export report” option of SameSpots, which creates one HTML file per comparison (i.e. 10 files in this case). S2P was then used to parse and merge these reports into a single table with samples in columns and spots in rows (Step 1 of Figure 1). All S2P data regarding this experiment is available at http://www.sing-group.org/s2p/tutorial-cs-2.html, along with a detailed quick-start tutorial that guides users working S2P in handling this dataset.

As this second case study represents a scenario different from that of the first case study, this first step must be done in a slightly different way. The main differences are that:

- In the first case study, each report contains information about two samples and the aim was to compare all samples in the dataset. In contrast, in this second case study, each report contains information about the same patient in different time points (or conditions). This is because it is a longitudinal study where each patient must be analyzed separately. Therefore, each report must be processed independently from the other reports.
- Unlike in the first case study, where all 2D-gel images were aligned against the master gel, in this experiment, only 2D-gel images from the same patient were aligned. This means that spots identification numbers in different samples (or reports) from different patients do not correspond to the same spot. For this reason, spots were identified in the MALDI matrix using the sample laboratory reference as suffix. To be able to match the spot ids written in the MALDI matrix (with suffixes) with those in the reports (without suffixes), when each patient report is processed with S2P, their suffix must be indicated.

Due to these differences, the first step of the workflow depicted by Figure 1 was adapted in S2P version 1.2.0 in order to allow single report processing and the possibility of changing the spot ids. S2P was then used to parse each patient report and create a table with its samples (i.e. time points) in columns and spots in rows (Step 1 of Figure 1). Each patient’s table was exported into a comma-separated values (CSV) file, which can be easily reopened with S2P as well as external applications such as Excel, LibreOffice or R.

As in the previous case study, once the MALDI-TOF MS analysis was completed, the MALDI-based spectra of the digested proteins were submitted to Mascot in order to identify the proteins. They were then exported into an HTML file that was loaded into S2P in order to remove duplicated entries and exclude identifications with a Mascot score under 56 (Step 3 of Figure 1). This processed list of Mascot identifications was exported into a CSV file so that it can be directly loaded into S2P at a later time, or used in other applications (e.g. spreadsheet software). These Mascot identifications were then integrated with the spots data using the MALDI plate (Step 4 of Figure 1) to know which identifications are associated with each spot.

Finally (Step 5 of Figure 1), S2P was used to integrate each patient's spots data with the Mascot identifications (Figure 2). It is important to note that while in the first case study the whole dataset was analyzed at once, in the second case study each patient is processed independently in S2P. Specifically, S2P was used to: (*i*) obtain the counts of analyzed spots, identified spots, and different identified proteins (summarized in Table 2); (*ii*) export the spots data along with the corresponding protein identifications into a CSV file for further analysis with spreadsheet processing software; and to (*iii*) export the necessary data for the functional protein categorization and integrative analysis with the STRAP and Cytoscape software.

Table 2. Counts of analyzed spots, identified spots and different proteins identified.

Nº sampling (months)	Patient	Analyzed spots	Identified spots	Different identified proteins
4 (1st, 7th, 12th, 24th)	P01 (IE)	31	19	19
3 (1st, 7th, 19th)	P02 (MIR)	24	11	12
	P03 (JM)	36	20	19
	P04 (VA)	40	21	21
2 (1st, 7th)	P05 (LP)	16	8	6
	P06 (SL)	13	12	16
	P07 (ML)	14	13	14
	P08 (MLC)	22	16	14
	P09 (MC)	10	7	5
	P10 (JP)	14	13	10

The CSV files exported with S2P contains the expression levels of each spot (along with the associated protein identification, if available) in the different time points. Each patient's data was analyzed to find proteins with increasing, decreasing or inconsistent (neither increasing nor decreasing) temporal trends, and study each group separately. This discussion is beyond the scope of this paper, which aims to

show the utility of S2P. In fact, at the time of writing this paper, results are not fully analyzed yet and the discussion is not completed.

3.3 Case study 3: protein quantification based on Exponentially Modified Protein Abundance Index (emPAI)

The Exponentially Modified Protein Abundance Index (emPAI) offers approximate, label-free, relative quantitation of the proteins in a mixture based on protein coverage by the peptide matches in a database search result [25]. In this case study, we explain how to use S2P in order to determine the relative abundance of serum protein using Mascot emPAI quantification data.

The Mascot search engine was used for protein identification and to determine the emPAI values for each identified protein. These results were obtained as described in the Materials and Methods section. Mascot reports containing emPAI values were exported using the “Export” option in Mascot Search Results, which creates a CSV file of each LC-MS/MS run (i.e. 8 files in this case). S2P was then used to parse and merge these reports in order to extract the emPAI values and provide different ways of analyzing the same data: (*i*) a table with one row by each single protein quantification in the whole dataset; (*ii*) a table where replicates can be compared by protein identification; and (*iii*) a table where experimental conditions can be compared by protein identification using the quantification values. In this case study, we used this latter table to perform three comparisons: (*i*) Overnight sample 1 vs. Overnight sample 2; (*ii*) Ultrasonic sample 1 vs. Ultrasonic sample 2; and (*iii*) Overnight vs. Ultrasonic samples. Then, the three tables corresponding to the three comparisons were exported into comma-separated values (CSV) files, which can be easily opened with external applications such as Excel, LibreOffice or R. In our case, these CSV tables were processed with Excel in order to calculate the ratios and generate Figure 6. The Excel spreadsheet containing these tables is provided as Supplementary Material. All S2P data regarding this experiment is available at http://www.sing-group.org/s2p/tutorial-cs-3.html, along with a detailed quick-start tutorial that guides users using S2P to process this type of datasets.

Our experimental setting combined with S2P allows us to (*i*) determine the accuracy of the quantification because we can correlate the replicate measurements of the same sample (first two comparisons); and (*ii*) detect bias towards specific proteins due to the digestion method (Ultrasonic digestion vs. Overnight Digestion).

As can be seen in Figure 6, the correlation analysis of the two sample replicates shows good correlation, R2 = 0.93 (A53ON_1 two LC-MS/MS replicates and A53ON_2 two replicates, 157 quantified proteins) for the Overnight digestion, and R2 = 0.99 (A53US_1 two LC-MS/MS replicates and A53US_2 two replicates, 172 quantified proteins) for the Ultrasonic Digestion. Because these samples are replicates, the expected ratio of A53ON_1: A53ON_2 and A53US_1: A53US_2 is 1. It was observed that both methods exhibit the expected ratio in more than 80% of the quantified proteins.

The same experimental data allows us to detect bias towards specific proteins due to the digestion method (Ultrasonic digestion vs. Overnight Digestion). The analysis revealed that 31 proteins (q-value < 0.05, Ratio US:ON > 1) were preferentially digested by the Ultrasonic method (UM) leading to an apparent higher concentration of these proteins in the samples digested with the UM in comparison with the samples digested overnight. On the other hand, only two proteins (q-value < 0.05, Ratio US:ON > 1) were preferentially digested using the Overnight method.

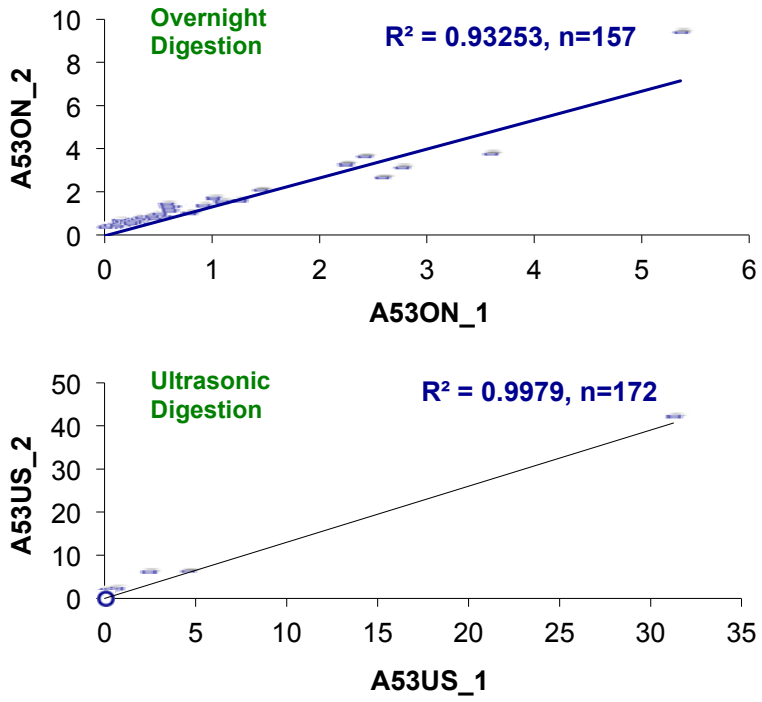


Figure 6. Correlation analysis of emPAI quantification of two serum samples digested overnight (Upper graph) and ultrasonic digestion (Lower graph).

4. Discussion

We provided an overview of the S2P v1.2 software and presented three real-world use cases to illustrate is capabilities. The two main advantages of S2P, when compared to alternative approaches to achieve the same data analysis workflow are

time saving and reproducibility. To our knowledge, the unique alternative approach to such workflow is that laboratory researchers do the processing by themselves using text editors and spreadsheet software.

Regarding the time saving, it is important to highlight that manually performing the steps required for the case study 1 dataset took more than two weeks. With the help of S2P, this data processing time was dramatically reduced to a few minutes. Similar scenarios occur in the other two case studies. On the other hand, by establishing a set of consecutive operations (Figure 1) with well-defined input and output files, S2P also allows researchers to gain control on the operations applied and, therefore, achieve reproducible analysis. We had experienced that manual processing of the case study 1 dataset resulted in unnoticeable errors that were only raised thanks to S2P. For these reasons, S2P offers a reproducible and reliable way of handling experimental data.

In a literature search, it can be found that most 2D gel electrophoresis analyses are meant to identify the set of differential expressed proteins present in biological samples when comparing different conditions. This comparison is done through the use of different commercial software packages. Then, after excision of the interest proteins, identification can be achieved with the help of mass spectrometry [26]. Most publicly available tools focus on the 2D-gel images analysis in order to detect spots [27,28]. Commercial software packages such as Progenesis SameSpots or DeCyder offers a vast amount of functionalities to deal with 2D data, including advanced statistical comparisons. However, they may not serve for such an integrative analysis in the way that S2P does. For instance, Progenesis SameSpots lacks of an option to automatically associate spots with protein identifications, an essential step of the described approach. This can only be done by providing the list of associations, which must be manually created. S2P does this automatically. For this reasons, S2P be viewed as a complement to traditional, well-established 2D analysis software, offering additional data analysis features (e.g. creation of heatmaps and Venn diagrams) in a single software that allow researchers to save a lot of their valuable time.

A possible criticism to S2P concerns data formats. As explained, the workflow applied to the first two case studies takes spots data generated from Progenesis SameSpots and protein identifications from Mascot HTML reports. Nevertheless, S2P also allows loading the same data from CSV files. For this reason, it has the

potential to analyze data from other sources, which only need to provide an option to export data into S2P CSV-compatible formats.

Finally, it is important to note that S2P has also been designed as a general platform open to different, closely related analyses, as the third case study reflects. As in the other two case studies, manual processing of Mascot reports in order to extract emPAI quantification values and create useful, easy to analyze tables, is not feasible. The third case study demonstrates how S2P has been extended to deal with Mascot CSV quantification reports in order to allow a comprehensive evaluation of the emPAI label-free quantitation results. By using this S2P feature, a researcher could compare two experimental groups (i.e. case vs. control) in order to find proteins with statistically different average quantification values, thus discovering potential biomarkers for further experiments.

5. Conclusions

S2P (http://www.sing-group.org/s2p/) is freely distributed under license GPLv3, providing a friendly graphical user interface designed to allow researchers to save time in data processing tasks related to 2D-gel electrophoresis and MALDI mass spectrometry protein identification-based data. The usefulness of S2P has been demonstrated by its application to real experiments, where it notably speeds up data processing as well as improves experiment reproducibility and reliability. S2P is open to further extensions and we are currently developing support for more types of datasets.

6. Mode of availability

The S2P software is licensed under a GNU GPL 3.0 License (http://www.gnu.org/copyleft/gpl.html). The S2P software along with full documentation and training tutorials are free and publicly available at http://www.sing-group.org/s2p.

Authors' contributions

HL-F, JEA, HMS, and JLC conceived, coordinated and designed S2P. HL-F, DG-P, and MR-J developed S2P. HL-F, JEA, SJ, and HMS tested the application. HL-F created the S2P website. JEA, SJ, and HMS performed the laboratory experiments. HL-F, JEA, HMS, and FF-R wrote the manuscript. FF-R, DG-P, MR-J, and JLC drafted the manuscript critically. All authors read and approved the final version of the manuscript.

Supplementary material

Supplementary Material.xlsx: Excel spreadsheet for case study 3 containing protein identifications and quantifications for the experimental samples along with its comparison.

Acknowledgements

We are thankful to Doctors Pedro M. Baltazar, Luís Campos Pinheiro, and Fernando Calais da Silva from the Urology Service at *Centro Hospital de Lisboa Central* (Lisboa, Portugal) for providing case study 1 samples. We are also thankful to Doctors Fernando Teixeira e Costa and Aura Ramos from the Nephrology Department at *Hospital Garcia de Orta* (Almada, Portugal) for providing case study 2 samples. SING group thanks CITI (*Centro de Investigación, Transferencia e Innovación*) from University of Vigo for hosting its IT infrastructure.

Funding

This work has been partially funded by: (*i*) the *"Platform of integration of intelligent techniques for analysis of biomedical information"* project (TIN2013-47153-C3-3-R) from Spanish Ministry of Economy and Competitiveness; (*ii*) the *"Discovery of biomarkers for bladder carcinoma diagnosis"* project from Nova Medical School (Nova Health Project Pilot 01); (*iii*) the *"Biomarkers for rheumatic inflammatory diseases diagnosis"* project from Nova Medical School (Nova Health Project Pilot 02); (*iv*) the UCIBIO (*Unidade de Ciências Biomoleculares Aplicadas*), which is financed by national funds from FCT/MEC (UID/Multi/04378/2013) and co-financed by the ERDF under the PT2020 Partnership Agreement (POCI-01-0145-FEDER-007728); (*v*) the *Associate Laboratory for Green Chemistry LAQV*, which is financed by national funds from FCT/MEC (UID/QUI/50006/2013) and co-financed by the ERDF under the PT2020 Partnership Agreement (POCI-01-0145-FEDER-007265); and (*vi*) Consellería de Cultura, Educación e Ordenación Universitaria (Xunta de Galicia) and FEDER (European Union). The Proteomass Scientific Society is acknowledged for the funding provided to the *Isabel Moura* biological mass spectrometry laboratory. H. López-Fernández is supported by a post-doctoral fellowship from Xunta de Galicia. H. M. Santos is funded by the FCT 2015 Investigator Program (IF/00007/2015). J. E. Araújo and S. Jorge are supported by the Portuguese Foundation for Science and Technology under doctoral grant numbers SFRH/BD/109201/2015 and SFRH/BD/120537/2016, respectively.

Conflict of interest statement

The authors have no conflicts of interest to declare.

References

[1] I. Susnea, B. Bernevic, M. Wicke, L. Ma, S. Liu, K. Schellander, M. Przybylski, Application of MALDI-TOF-Mass Spectrometry to Proteome Analysis Using Stain-Free Gel Electrophoresis, in: Z. Cai, S. Liu (Eds.), Appl. MALDI-TOF Spectrosc., Springer Berlin Heidelberg, Berlin, Heidelberg, 2012: pp. 37–54. http://link.springer.com/10.1007/128_2012_321 (accessed November 9, 2016).

[2] J.L.C. Martinez, C.L. Espiño, H.M. Santos, Mass Spectrometry-based Proteomics: What is it expecting ahead?, J. Proteomics. 145 (2016) 1–2. doi:10.1016/j.jprot.2016.07.017.

[3] S.R. Nagalla, J.A. Canick, T. Jacob, K.A. Schneider, A.P. Reddy, A. Thomas, S. Dasari, X. Lu, J.A. Lapidus, G.M. Lambert-Messerlian, M.G. Gravett, C.T. Roberts, D. Luthy, F.D. Malone, M.E. D'Alton, Proteomic Analysis of Maternal Serum in Down Syndrome: Identification of Novel Protein Biomarkers, J. Proteome Res. 6 (2007) 1245–1257. doi:10.1021/pr060539h.

[4] V. Thongboonkerd, K.R. Mcleish, J.M. Arthur, J.B. Klein, Proteomic analysis of normal human urinary proteins isolated by acetone precipitation or ultracentrifugation, Kidney Int. 62 (2002) 1461–1469. doi:10.1111/j.1523-1755.2002.kid565.x.

[5] C.-T. Hsueh, D. Liu, H. Wang, Novel biomarkers for diagnosis, prognosis, targeted therapy and clinical trials, Biomark. Res. 1 (2013) 1. doi:10.1186/2050-7771-1-1.

[6] M. Baker, Reproducibility: Seek out stronger science, Nature. 537 (2016) 703–704. doi:10.1038/nj7622-703a.

[7] B.R. Zeeberg, J. Riss, D.W. Kane, K.J. Bussey, E. Uchio, W.M. Linehan, J.C. Barrett, J.N. Weinstein, Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics, BMC Bioinformatics. 5 (2004) 80. doi:10.1186/1471-2105-5-80.

[8] M. Ziemann, Y. Eren, A. El-Osta, Gene name errors are widespread in the scientific literature, Genome Biol. 17 (2016). doi:10.1186/s13059-016-1044-7.

[9] N.L. Anderson, The Human Plasma Proteome: History, Character, and Diagnostic Prospects, Mol. Cell. Proteomics. 1 (2002) 845–867. doi:10.1074/mcp.R200007-MCP200.

[10] S.E. Warder, L.A. Tucker, T.J. Strelitzer, E.M. McKeegan, J.L. Meuth, P.M. Jung, A. Saraf, B. Singh, J. Lai-Zhang, G. Gagne, J.C. Rogers, Reducing agent-mediated precipitation of high-abundance plasma proteins, Anal. Biochem. 387 (2009) 184–193. doi:10.1016/j.ab.2009.01.013.

[11] C. Fernández, H.M. Santos, C. Ruíz-Romero, F.J. Blanco, J.-L. Capelo-Martínez, A comparison of depletion versus equalization for reducing high-abundance proteins in human serum, ELECTROPHORESIS. 32 (2011) 2966–2974. doi:10.1002/elps.201100183.

[12] J.E. Araújo, T. Santos, S. Jorge, T.M. Pereira, M. Reboiro-Jato, R. Pavón, R. Magriço, F. Teixeira-Costa, A. Ramos, H.M. Santos, Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry-based profiling as a step forward in the characterization of peritoneal dialysis effluent, Anal Methods. 7 (2015) 7467–7473. doi:10.1039/C5AY00620A.

[13] J.E. Araújo, S. Jorge, F. Teixeira e Costa, A. Ramos, C. Lodeiro, H.M. Santos, J.L. Capelo, A cost-effective method to get insight into the peritoneal dialysate effluent proteome, J. Proteomics. 145 (2016) 207–213. doi:10.1016/j.jprot.2016.05.010.

[14] J.E. Araújo, S. Jorge, R. Magriço, T. e Costa, A. Ramos, M. Reboiro-Jato, F. Fdez-Riverola, C. Lodeiro, J.L. Capelo, H.M. Santos, Classifying patients in peritoneal dialysis by mass spectrometry-based profiling, Talanta. 152 (2016) 364–370. doi:10.1016/j.talanta.2016.02.026.

[15] M.M. Bradford, A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding, Anal. Biochem. 72 (1976) 248–254. doi:10.1016/0003-2697(76)90527-3.
[16] E. Oliveira, J.E. Araújo, S. Gómez-Meire, C. Lodeiro, C. Perez-Melon, E. Iglesias-Lamas, A. Otero-Glez, J.L. Capelo, H.M. Santos, Proteomics analysis of the peritoneal dialysate effluent reveals the presence of calcium-regulation proteins and acute inflammatory response, Clin. Proteomics. 11 (2014) 17. doi:10.1186/1559-0275-11-17.
[17] H.M. Santos, R. Rial-Otero, L. Fernandes, G. Vale, M.G. Rivas, I. Moura, J.L. Capelo, Improving Sample Treatment for In-Solution Protein Identification by Peptide Mass Fingerprint Using Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry, J. Proteome Res. 6 (2007) 3393–3399. doi:10.1021/pr0702518.
[18] S. Jorge, J.E. Araújo, F.M. Pimentel-Santos, J.C. Branco, H.M. Santos, C. Lodeiro, J.L. Capelo, Unparalleled Sample Treatment Throughput for Proteomics Workflows Relying on Ultrasonic Energy, Talanta. (2017). doi:10.1016/j.talanta.2017.07.079.
[19] D. Glez-Peña, M. Reboiro-Jato, P. Maia, M. Rocha, F. Díaz, F. Fdez-Riverola, AIBench: a rapid application development framework for translational research in biomedicine, Comput. Methods Programs Biomed. 98 (2010) 191–203. doi:10.1016/j.cmpb.2009.12.003.
[20] H. López-Fernández, M. Reboiro-Jato, D. Glez-Peña, F. Aparicio, D. Gachet, M. Buenaga, F. Fdez-Riverola, BioAnnote: A software platform for annotating biomedical documents with application in medical learning environments, Comput. Methods Programs Biomed. 111 (2013) 139–147. doi:10.1016/j.cmpb.2013.03.007.
[21] H. López-Fernández, M. Reboiro-Jato, D. Glez-Peña, J.R. Méndez Reboredo, H.M. Santos, R.J. Carreira, J.L. Capelo-Martínez, F. Fdez-Riverola, Rapid development of Proteomic applications with the AIBench framework, J. Integr. Bioinforma. 8 (2011) 171. doi:10.2390/biecoll-jib-2011-171.
[22] M. Reboiro-Jato, D. Glez-Peña, J.R. Méndez-Reboredo, H.M. Santos, R.J. Carreira, J.L. Capelo, F. Fdez-Riverola, Building proteomics applications with the aibench application framework, 2011.
[23] G. Pérez-Rodríguez, D. Glez-Peña, N.F. Azevedo, M.O. Pereira, F. Fdez-Riverola, A. Lourenço, Enabling systematic, harmonised and large-scale biofilms data computation: The Biofilms Experiment Workbench, Comput. Methods Programs Biomed. 118 (2015) 309–321. doi:10.1016/j.cmpb.2014.12.005.
[24] H. Barsnes, M. Vaudel, L. Martens, JSparklines: Making tabular proteomics data come alive, PROTEOMICS. 15 (2015) 1428–1431. doi:10.1002/pmic.201400356.
[25] Y. Ishihama, Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, M. Mann, Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein, Mol. Cell. Proteomics. 4 (2005) 1265–1272. doi:10.1074/mcp.M500061-MCP200.
[26] A. Hmmier, P. Dowling, DIGE Analysis Software and Protein Identification Approaches, in: K. Ohlendieck (Ed.), Differ. Gel Electrophor., Springer New York, New York, NY, 2018: pp. 41–50. doi:10.1007/978-1-4939-7268-5_4.
[27] M.H. Maurer, Two-Dimensional Gel Electrophoresis Image Analysis via Dedicated Software Packages, in: E. Marengo, E. Robotti (Eds.), 2- PAGE Map Anal., Springer New York, New York, NY, 2016: pp. 55–65. doi:10.1007/978-1-4939-3255-9_3.
[28] R.S. Sengar, A.K. Upadhyay, M. Singh, V.M. Gadre, Analysis of 2D-gel images for detection of protein spots using a novel non-separable wavelet based

method, Biomed. Signal Process. Control. 25 (2016) 62–75. doi:10.1016/j.bspc.2015.10.013.