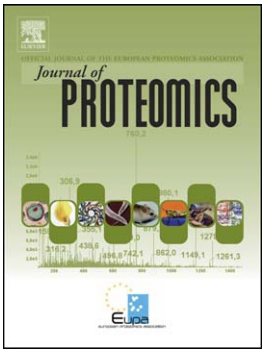
Parkinson's disease plasma biomarkers: An automated literature analysis followed by experimental validation



Parkinson's Disease plasma biomarkers: An automated literature analysis followed by experimental validation

Tiziana Alberio, Enrico M. Bucci, Massimo Natale, Dario Bonino, Marco Di Giovanni, Edo Bottacchi, Mauro Fasano

PII: S1874-3919(13)00055-9
DOI: doi: 10.1016/j.jprot.2013.01.025
Reference: JPROT 1302

To appear in: *Journal of Proteomics*

Received date: 15 November 2012
Accepted date: 22 January 2013

Please cite this article as: Alberio Tiziana, Bucci Enrico M., Natale Massimo, Bonino Dario, Giovanni Marco Di, Bottacchi Edo, Fasano Mauro, Parkinson's Disease plasma biomarkers: An automated literature analysis followed by experimental validation, *Journal of Proteomics* (2013), doi: 10.1016/j.jprot.2013.01.025

Parkinson's Disease plasma biomarkers: An automated literature analysis followed by experimental validation

Tiziana Alberio[a,b,1], Enrico M. Bucci[c,1], Massimo Natale[c,d], Dario Bonino[c], Marco Di Giovanni[e], Edo Bottacchi[e], Mauro Fasano [a,b,*]

[a] Division of Biomedical Research, Department of Theoretical and Applied Sciences, University of Insubria, Busto Arsizio, Italy.

[b] Center of Neuroscience, University of Insubria, Busto Arsizio, Italy.

[c] BioDigitalValley s.r.l., Pont Saint Martin (AO).

[d] Department of Control and Computer Engineering, Politecnico di Torino, Torino, Italy.

[e] Department of Neurology, Ospedale Regionale di Aosta, Aosta, Italy.

[1] These authors contributed equally.

[*] *Corresponding author at*: Division of Biomedical Sciences, Department of Theoretical and Applied Sciences, University of Insubria, via Alberto da Giussano 12, I-21052 Busto Arsizio, Italy. Tel: +39 0331 339450; fax: +39 0331 339459.

E-mail address: mauro.fasano@uninsubria.it (M. Fasano).

Keywords:

2-DE;

Plasma;

Parkinson’s disease;

Biomarkers;

Meta-analysis

Abstract

Diagnosis of Parkinson's disease (PD) is currently assessed by the clinical evaluation of extrapyramidal signs. The identification of specific biomarkers would be advisable, however most studies stop at the discovery phase, with no biomarkers reaching clinical exploitation. To this purpose, we developed an automated literature analysis procedure to retrieve all the background knowledge available in public databases. The bioinformatic platform allowed us to analyze more than 51000 scientific papers dealing with PD, containing information on 4121 proteins. Out of these, we could track back 35 PD-related proteins as present in at least two published 2-DE maps of human plasma. Then, 9 different proteins (haptoglobin, transthyretin, apolipoprotein A-1, serum amyloid P component, apolipoprotein E, complement factor H, fibrinogen γ, thrombin, complement C3) split into 32 spots were identified as a potential diagnostic pattern. Eventually, we compared the collected literature data to experimental gels from 90 subjects (45 PD patients, 45 non-neurodegenerative control subjects) to experimentally verify their potential as plasma biomarkers of PD.

1. Introduction

The identification of specific biomarkers for neurodegenerative disorders is one of the main goal of the current clinical research. Diagnosis of the second most common neurological disorder, Parkinson's disease (PD), is currently assessed by the clinical evaluation of extrapyramidal signs, such as tremor, rigidity and bradykinesia, when the degeneration of dopaminergic nigral neurons has raised over 70% [1],[2]. The identification of peripheral PD biomarkers would be critical for the differential diagnosis between PD and other neurodegenerative diseases that share some clinical features with PD. Moreover, early diagnosis remains a primary aim, needed also for the assessment of disease-modifying drugs [3],[4],[5]. Eventually, biomarkers of disease progression would allow a better patients follow-up and an objective measurement in clinical trials [6]. Non-motor signs frequently precede the onset of typical PD motor dysfunctions but they lack the specificity to be informative for the clinician; on the other hand, instrumental investigations (polysomnograpy or functional imaging) are characterized by high costs and the use of radioactive tracers. For these reasons, a biochemical marker to be evaluated in the periphery would be highly preferred [3],[4],[5].

Several PD biomarker candidates were discovered using unbiased proteomic approaches [4],[7]. 2-DE is able to resolve thousands of spots simultaneously and can visualize protein processing and post-translational modifications [8]. For these reasons, it has been widely used in biomarker discovery studies. In regards to the tissue where to look for biomarkers, human plasma was broadly exploited because potentially informative about almost any disease state and very easy to obtain with little discomfort for patients. Moreover, plasma seems to be suitable for proteomic analyses to be applied

in the clinical practice [9]. Nevertheless, the proteomic investigation of human plasma, especially by 2-DE, has proved to be challenging, because of the very complex nature of the sample and the presence of few highly-abundant proteins [10].

Meta-analysis of proteomics data may represent a valuable tool to extract information from datasets present in the literature, thus increasing the sample size with respect to a single proteomics study [11]. In particular, meta-analysis of 2-DE data available in the literature has proved to have a great potential to retrieve valuable information related to Parkinson's disease, which otherwise would remain hidden [12]. Plenty of information is available online in public domain databases. However, these datasets need to be filtered in order to sensibly and specifically find proteins that are associated to PD as biomarkers [13]. Here again, meta-analysis of 2-DE experiments present in the scientific literature or in pertinent datasets can help to identify biologically relevant biomarkers by filtering out proteins involved in generic processes (e.g., cell-cycle control, signal transduction, and stress responses) [14]. Eventually, the data obtained from the analysis of the literature should be further validated in an independent manner in a cohort large enough to ensure statistical significance.

In this study, we developed a bioinformatic platform for the automatic analysis of proteomics literature data (Parkinson Informative System, ParIS, http://www.biodigitalvalley.com/researchprojects.html). We used a meta-analysis approach to filter literature data and identify a list of potential PD biomarkers, proposed by different authors and visible in 2-DE maps of human plasma. We further verified these markers in a cohort of 45 PD patients compared to 45 control subjects with neurological, non neurodegenerative disorders. Some of the markers were confirmed, while others did not reach statistical significance. Eventually, we built a predictive

model able to significantly stratify PD patients by the identification of all the spots whose level was significantly different in the two groups.

2. Methods

2.1. 2-DE gel database and proteomic meta-analysis

In order to extract human plasma 2-DE images from the literature, we parsed PDF files of selected articles using the BFO Java library (http://bfo.com) as previously described [12]. Briefly, the 2-DE images extracted from proteomic papers were annotated for all the protein spots, according to identifications provided by the article authors, matching the symbols and labels on the gels to the information given in the tables or in the text of the paper. Spots were further associated to their pixel cartesian position and to pI and MW data. To get putative spot identification, we aligned the 22 literature-retrieved 2-DE gels to the human plasma reference 2-DE map (PLASMA_HUMAN) from Expasy SWISS-2DPAGE (http://world-2dpage.expasy.org/swiss-2dpage/). To store all the retrieved gels and their annotations, a human plasma 2-DE gel database was built using a MySQL environment on a Red-Hat server.

2.2. Subjects

Ninety subjects were enrolled by the Department of Neurology at the Aosta Regional Hospital, after approval by the local Institutional Review Board. All patients signed an informed consent before being recruited for the present study, according to the *Declaration of Helsinki*. Every subject was associated to an alphanumeric code by the medical personnel involved in this study, to ensure that his/her identity was not apparent to investigators. Among the enrolled individuals, 45 subjects were idiopathic PD patients, varied in terms of age, age at onset, pharmacological treatment, with no

familiarity and no other co-occurring neurological disease. Beside personal details, for each subject affected by Parkinson's disease we registered the presumptive onset year, the Hoehn & Yahr score and the dopaminergic therapy (Table 1). The 45 control subjects were recruited by the same Hospital Department among neurological patients not affected by neurodegeneration or multiple sclerosis. Gender and age distributions were similar in different groups (Table 1). Plasma samples were collected according to the protocol of the Parkinson's progression markers initiative (http://www.ppmi-info.org/wp-content/uploads/2012/03/PPMI-Biologics-Manual.pdf). Immediately after the blood collection the tubes were gently mixed and within 30 minutes the samples were centrifuged at 1500×g at 4°C for 15 minutes. Plasma was collected into 1.5 ml aliquots and immediately stored at -80°C.

2.3. Two-dimensional electrophoresis and image analysis

Protein concentration in plasma samples was determined according to Bradford. An aliquot of 400 µg of proteins was mixed with 10 µl of a solution containing SDS (10% w/v) and DTT (2.3% w/v). The sample was heated to 95°C for 5 minutes and then diluted to 250 µl with a solution containing 7 M urea, 2 M thiourea, 4% CHAPS, 1% IPG buffer 4-7 (GE Healthcare, Uppsala, Sweden), 2% Amidosulfobetaine-14 (ASB-14; Sigma-Aldrich, St. Louis, MO, USA) and a trace of bromophenol blue. Total proteins were separated through 2-DE using 13 cm IPG DryStrips with a 4–7 pH gradient (GE Healthcare) followed by 12.5%T SDS-PAGE. The resulting maps were stained with Ru(II) tris(bathophenanthroline disulfonate) (Serva, Heidelberg, Germany). Images were acquired (12 bit grayscale) with the GelDoc-It Imaging System (UVP, Upland, CA, USA) and analyzed with Melanie 7.0 (GeneBio, Geneva, Switzerland); spots showing significant variation among subject groups (Wilcoxon test, $p<0.05$) and not recognized

by matching with the database created as described in paragraph 2.1 were excised from gels and the corresponding proteins identified by LC-MS/MS fragmentation.

2.4. Statistical analysis

Spot volumes were normalized by the total volume of a subset of spots common to all gels. Spots that were missing in more than 25% of gels were not taken into account. Missing spot values in less than 25% of gels were replaced by the mean value of the spot volume of the group or, if the mean was lower than the 98th percentile, by the minimum value observed in the group [15]. Relative volumes were analyzed by the non-parametric Wilcoxon test to find significant differences ($p < 0.05$) in patients with PD with respect to control subjects and in PD patients treated or not with dopamine agonists [16],[17].

The Pearson linear correlation coefficient *r* was evaluated according to Eq. 1, where x is the independent variable (*i.e.*, age, daily L-DOPA dose), y is the relative spot volume, $\bar{x}$ is the mean x value, $\bar{y}$ is the mean y value, xy is the x × y product, $\overline{xy}$ is the product of $\bar{x}$ and $\bar{y}$ mean values.

$$r = \frac{\frac{\sum xy}{n} - \overline{xy}}{\sqrt{\frac{\sum x^2}{n} - \bar{x}^{-2}} \sqrt{\frac{\sum y^2}{n} - \bar{y}^{-2}}} \quad (1)$$

Predictive models for the classification of PD patients with respect to control subjects were built by linear discriminant analysis (LDA) of the spots identified as described above [16]. In this case, missing values were set to half of the minimum value observed in the gel. A likelihood score was assigned to each subject by linear combination of relative spot volumes according to Eq. 2.

$$\text{PD}\,Score = \sum_i c_i Vol_i \tag{2}$$

A simplified model was obtained by removing spots with lower discriminating weight (*W*), calculated according to Eq. 3,

$$W = \left| c_i \left(\overline{Vol_{i,CO}} - \overline{Vol_{i,PD}} \right) \right| \tag{3}$$

where c_i is the LDA coefficient for spot *i* and $\left| \left(\overline{Vol_{i,CO}} - \overline{Vol_{i,PD}} \right) \right|$ is the absolute separation of the mean values of spot *i* in control subjects (CO) and PD. Predictive models have been tested with the leave-one-out method and their performance quantified by measuring the area under the receiver operating characteristic (ROC) curve [16].

Aggregative nesting of spots included in the predictive model was based on pairwise Pearson correlation (Eq. 1), where x and y are relative spot volumes.

The minimum number of subjects to be included in validation studies was calculated according to Eq. 4,

$$n = \frac{\left(N_{CO} - 1\right) \cdot \sigma_{CO}^2 + \left(N_{PD} - 1\right) \cdot \sigma_{PD}^2}{\left(N_{CO} + N_{PD} - 2\right) \cdot \left(\frac{\mu_{CO} - \mu_{PD}}{2.2} \right)^2} \tag{4}$$

where σ_{CO} and σ_{PD} are standard deviations of relative spot volumes in CO and PD groups, N_{CO} and N_{PD} are numbers of subjects in the groups, and μ_{CO} and μ_{PD} are mean values of relative spot volumes in the groups [16].

All procedures for data analysis and graphics were written in ***R***, an open-source environment for statistical computing [18] and are available at http://dx.doi.org/10.1038/protex.2013.001.

2.5. In-gel digestion, mass spectrometry and protein identification

Spots were manually excised and destained (50% ethanol), dehydrated with acetonitrile (Sigma-Aldrich) (2 × 100 µl, 20 min) and then dried at 37°C by vacuum centrifugation. Protein spots were trypsin digested as already described [19].

Protein identification by LC-MS/MS was performed by Primm (Milan, Italy). Peptides from each sample were separated by HPLC 1100 online-coupled with a LC/MSD XCT Ultra Ion Trap (Agilent Technologies, Santa Clara, CA, USA), equipped with a HPLC 1100 and a chip cube (Agilent Technologies). Protein identification was performed by searching the National Center for Biotechnology Information non-redundant database using the Mascot MS/MS Ion Search program (http://www.matrixscience.com) with the built-in decoy option. The following parameters were set: ESI TRAP instrument, specific trypsin digestion, up to one missed cleavage, complete carbamidomethylation of cysteines, partial oxidation of methionines, partial protein N-terminal acetylation, peptide mass tolerance ± 600 ppm, fragment mass tolerance ± 0.6 Da, 2+ to 3+ peptide charge, species restriction to human. All identified proteins had a MOWSE score corresponding to a statistically significant ($p < 0.05$) confident identification. At least 2 different peptides had to be assigned.

3. Results

3.1 Meta-analysis reveals nine candidate PD biomarkers

First of all, as described in the methods section we have populated a database with 22 2-DE gel images of whole human plasma, extracted from proteomics papers or from the Swiss 2D page database (http://world-2dpage.expasy.org). The gel spots were annotated

as described in paragraph 2.1. Within the selected 22 images, we have assigned 927 different spots corresponding to a total of 188 independent proteins.

We filtered these proteins in order to consider only those having at least two citations in papers related to PD. To this aim, we used ProteinQuest (http://www.proteinquest.com), a web-based platform for the mining of Medline papers, which allows the identification of the co-occurrence of biological concepts (e.g., proteins, disease names, drugs). In particular, we used ProteinQuest to analyze the co-occurrence between PD-related terminology and the 188 proteins annotated in our 2-DE gel database. As a result, 84 of these proteins occurred in at least two papers related to PD. Among these proteins, we selected only proteins that were already defined as potential PD biomarkers (e.g., by ignoring proteins related to the pathogenesis) and were detectable in 2-DE plasma gels (thus excluding proteins with low abundance). In this way, we identified 9 different proteins (haptoglobin (HP), transthyretin (TTR), apolipoprotein A-1 (APOA1), serum amyloid P component (SAP), apolipoprotein E (APOE), complement factor H (CFH), fibrinogen γ (FGG), thrombin (F2) and complement c3 (C3)) that satisfy all the requirements, i.e., to be a protein visible on a plasma 2-DE gel and at the same time to be known as a potential biomarker of PD. The nine potential biomarker proteins, split into 32 spots, were identified as suitable source for a diagnostic pattern and were considered for the subsequent analysis; the increase or decrease of protein concentration as expected from the analyzed literature is reported in Supplementary Table 1.

3.2 Two-dimensional electrophoresis profiling of plasma proteins

To perform the biomarker validation analysis, we obtained plasma protein expression profiles from the enrolled groups by 2-DE. First, we performed technical replicates for

14 PD patients and 17 control subjects in order to evaluate the linear correlation of spot volumes as a reproducibility benchmark of the maps. Most of the gel pairs display linear correlation of spot volumes with Pearson correlation coefficient greater than 0.8 (Fig. 1A). Moreover, we calculated the intra-group standard deviation of spot volumes to verify if they were normally or log-normally distributed. As commonly observed in biological variables, standard deviations of spot volumes after logarithmic transformation were constant in both groups (Fig. 1B). For this reason, we decided to rely on the non-parametric Wilcoxon test [16]. Moreover, we compared pairs of 2-DE gels from different subjects by Pearson linear correlation and analyzed the distribution of fitting residuals to assess the correctness of a pairwise comparison. Supplementary Fig. 1 shows the studentized residuals against ranked magnitudes (QQ) plots [18]. On this basis, a direct proportionality between gels from different subjects may be assumed.

Once tested the quality of 2-DE maps, we screened the profiles to identify proteins or protein modifications whose changes were linked to confounding factors such as therapy and age. Spots showing linear Pearson correlation with age (evaluated in control subjects only) or daily L-DOPA dose, or showing significant differences between patients treated or not with dopamine agonists (Wilcoxon test, $p<0.05$) were excluded from the further analysis. By comparing 2-DE maps from the 45 PD patients to the 45 control subjects, we selected 26 protein spots showing significantly different levels in the two groups (Wilcoxon test, $p<0.05$) (Fig. 2 and Table 2). For spots accurately matching with the 2-DE annotated map of human plasma obtained as described in paragraph 2.1 (no. 22, 23, 24, 25, 34, 38, 43, 44, 46, 61, 62, 92, 97, 113, 117, 132, 141, 147, 150 and 151), the protein identity was assumed to be unambiguous. Other spots (no.

26, 41, 45, 66, 89 and 105) were identified by LC-MS/MS (see Table 3 and Supplementary Table 2 for details on protein identification by mass spectrometry).

3.3 Validation of candidate biomarkers

First of all, we verified the presence of the nine candidate biomarkers identified as described in paragraph 3.1 among the proteins that showed significant changes in our experimental data. Haptoglobin, transthyretin, apolipoprotein A-1, apolipoprotein E, complement factor H and a complement c3 fragment were confirmed to be markers of the disease. Indeed, we observed altered levels in spot no. 22-26, 38, 43, 92, 97, 113, 117, 132, 141, 151 (for spot-protein correspondence see Table 2).

On the contrary, serum amyloid P component, fibrinogen γ and thrombin were excluded from the diagnostic pattern because not confirmed by our experimental data. Interestingly, FGG was excluded because the level of spots corresponding to this protein correlated with L-DOPA daily dose and age.

Eventually, dealing with a large cohort of subjects, we observed significant level changes in spots that were not included in the results of the meta-analysis procedure (Spots no. 34, 41, 44, 45, 46, 61, 62, 66, 105, 147 and 150), thus allowing us to enrich the list of biomarker candidates found in plasma by 2-DE (for spot-protein correspondence see Table 2).

3.4 Linear discriminant analysis of selected spot levels

We analyzed all spots (n = 25) showing significantly different levels in PD patients by LDA, so to select the spots with the higher weight. The 10 spots showing the worst contribution (weight < 0.07) were discarded. The remaining 15 spots were clustered by aggregative nesting to detect the extent of correlation among spots (Fig. 3A). These spots were analyzed again by LDA. Likelihood scores (PD Score, obtained as a linear

combination of relative spot volumes) were significantly different in PD patients with respect to control subjects (Wilcoxon test, $p < 10^{-16}$) (Fig. 3B). To effectively test the performance of the model, each subject was iteratively excluded from the training set and predicted on the basis of the other subjects. The "leave-one-out" cross-validation procedure of the model allowed us to obtain 83% sensitivity and 82% specificity. Predictions obtained so far were used to build a ROC curve with an area under curve of 0.886 (Fig. 3C).

To test the power of our analysis we evaluated the intra-group variance and the difference of mean values for each spot listed in Table 2. The minimum number of subjects required for a significant verification is always lower than 45 with the only exception of spots 38, 44 and 117 (Supplementary Table 3).

4. Discussion

Here we show a method for the validation of candidate biomarkers as reported by the literature. Focusing on plasma biomarker discovery studies for Parkinson's disease diagnosis, we identified a subset of proteins that at the same time have been associated to Parkinson's disease and have been detected by 2-DE. Actually, the experimental result of 2-DE is a digital image that can be retrieved in an automatic way. The creation of a human plasma 2-DE gels database and the analysis of PD biomarkers papers in Medline with the application of suitable filters allowed us to focus on a reduced set of proteins to be further validated as PD biomarkers.

The choice of plasma as the source of biomarkers has been imposed by the high number of papers where biomarker discovery was performed in this biofluid. However, plasma may represent a challenging matrix for proteomics studies, especially for those

based on 2-DE [10]. Indeed, the use of 2-DE as the separating step in differential proteomics is controversial, since the technique is usually considered to show an intrinsically low reproducibility and low sensitivity. Moreover, in the case of plasma 2-DE separation, the detection of biomarkers is limited to the most abundant proteins [7]. Nevertheless, 2-DE images represent a direct display of protein expression in a tissue or biofluid.

In order to evaluate the effective suitability of candidate biomarkers by 2-DE, we first measured technical reproducibility in our set of 121 gel images. Noticeably, we observed a very low dispersion of Pearson correlation coefficients for 31 pairs of technical replicates, with values $r > 0.8$. Moreover, the normal distribution of linear fitting residuals when 2-DE gels from different subjects are compared ensures that a pairwise gel comparison is methodologically correct (Supplementary Fig. 1). Eventually, our experimental design included a group of non-healthy control subjects in order to eliminate false candidate biomarkers linked to a general state of inflammation present in PD patients. Indeed, different pieces of evidence suggest a possible implication of inflammation in the degeneration of dopaminergic neurons [20].

Despite the precautions to avoid possible common biases of biomarkers discovery studies, our approach verified six out of nine candidate biomarkers, namely haptoglobin, transthyretin, apolipoprotein A-1, apolipoprotein E, complement factor H and complement c3. In particular, HP and CFH showed the highest significance.

HP is an acute phase protein synthesized by liver cells, which has been widely proposed as a PD biomarker [21],[22]. HP probably plays a modulatory and protective role on autoimmune inflammation of the CNS [23] and on the integrity of the nigro-striatal dopaminergic system [24]. Therefore, its altered levels could mirror the

neurodegenerative process at the central level. Notwithstanding this, a single isoform of HP was excluded from the panel because of its correlation with the assumption of dopamine agonists. The advantage of 2-DE, in this case, is the possibility to analyze the different forms of the same protein separately. CFH is a cofactor in the inactivation of C3b by factor I and also increases the rate of dissociation of the C3bBb complex (C3 convertase) and the (C3b)NBB complex (C5 convertase) in the alternative complement pathway. Our results support the hypotheses of a dysregulation of the complement pathways in neurodegenerative disorders and in PD in particular [7] and are in agreement with a significant reduction in factor H levels observed in the CSF of PD patients [25]. A single fragment of C3 was found to change in our cohort, but we cannot exclude that other fragments beyond our limit of detection are effectively changing. In any case, the power analysis revealed that this spot is characterized by a high variability and that the number of subjects to be analyzed should be higher. Also TTR would require a higher number of subjects to be considered a real PD biomarker.

On the other hand, we were not able to confirm the validity of thrombin (F2), gamma fibrinogen (FGG) and serum amyloid P component (SAP). Thrombin exerts physiological and pathological functions in the central nervous system, and it has been documented that the level of thrombin increases in human brain of patients with Parkinson's disease [26],[27]. Nevertheless, the differences in concentration may not be detected at the peripheral level. Additionally, higher F2 levels have been associated to ischemic injury [28], a finding that could reduce the significance of this candidate biomarker when control subjects include stroke patients as it is our case. SAP was not significantly different between patients and control subjects, probably because PD patients were compared with neurological control subjects, who might have altered

levels of acute phase proteins. Moreover, even if FGG was correlated to PD, we excluded it from the panel because of the correlation of its levels with the therapy assumed by PD patients. In verification studies, it is particularly important to eliminate confounding effects such as the pharmacological treatment, that intrinsically discriminate between patients and control subjects [17].

Furthermore, the experimental validation procedure allowed us to identify a set of additional candidate biomarkers that were not present in the list of proteins to be validated in the present study. Some of them have been already associated to PD, however they were not identified by the automated literature search because they did not fulfill the stringency parameters imposed in the analysis process. For instance, apolipoprotein A-IV has been proposed as a plasma PD biomarker only once [22]. Therefore, it was filtered out during the automated analysis of literature.

In other cases, we observed a significant variation in proteins whose change in association with PD was not reported previously or is controversial. For instance, we observed that α2-macroglobulin level was lower in plasma of PD patients with respect to neurological, non-neurodegenerative control subjects. This protein has been originally proposed as a PD marker [29], but other studies contradict the original hypothesis [30]. As a matter of fact, like many other inflammatory markers, α2-macroglobulin displays constant levels in PD patients with different motor and cognitive impairment [31]. This finding highlights again the importance of the choice of the appropriate control group.

Eventually, a linear discriminant analysis was applied to identify the set of 15 spots that performed at best in discriminating PD patients from control subjects. A hierarchical clustering of these spots showed a major contribution of HP isoforms,

segregated in a single cluster of the dendrogram. The only other cluster contains other proteins, with a main contribution of apolipoproteins. By considering proteins identified both by the literature analysis and the experimental assessment, it was possible to build a PD predictive model. The ROC curve obtained by the leave-one-out cross validation demonstrates a quite good predictive power, with AUC= 0.886. Remarkably, a pattern recognition tool would allow to calculate a PD score from the plasma 2-DE map of a subject, independently from proteins identity.

5. Conclusions

The present study shows a possible pipeline to validate candidate PD biomarkers. Starting from a broad automated analysis of the literature, we selected few proposed biomarkers in plasma and verified them in a cohort of 90 subjects. Some of the candidates, that arose from the literature analysis, were confirmed. At the same time, we identified other proteins whose level was different between PD patients and control subjects. Putting all the results together, we performed a LDA and assigned a weight to each protein. In this way, we selected the 15 most relevant contributors to a PD predictive model. Taken altogether, these data suggest that an automated analysis of literature data provides a useful tool to biomarkers validation studies.

Acknowledgement

This work was partially funded under the “Legge 84” regional funding program of the Valle d’Aosta Region (ParIS project). Authors gratefully acknowledge Mr. Gianluca D'Agostino, Mr. Andrea Ruffino and Mr. Moreno Cornaz for technical assistance and Dr. Cristina Cereda for helpful discussion.

References

[1] Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008;79:368-76.

[2] Shulman JM, De Jager PL, Feany MB. Parkinson's disease: genetics and pathogenesis. Annu Rev Pathol 2011; 6:193-222.

[3] Morgan JC, Mehta SH, Sethi KD. Biomarkers in Parkinson's disease. Curr Neurol Neurosci Rep 2010;10:423-30.

[4] Alberio T, Fasano M. Proteomics in Parkinson's disease: An unbiased approach towards peripheral biomarkers and new therapies. J Biotechnol 2011;156:325-37.

[5] Gerlach M, Maetzler W, Broich K, Hampel H, Rems L, Reum T, Riederer P, Stöffler A, Streffer J, Berg D. Biomarker candidates of neurodegeneration in Parkinson's disease for the evaluation of disease-modifying therapeutics. J Neural Transm 2012;119:39-52.

[6] Frasier M, Chowdhury S, Eberling J, Sherer T. Biomarkers in Parkinson's disease: a funder's perspective. Biomark Med 2010;4:723-9.

[7] Sheta EA, Appel SH, Goldknopf IL. 2D gel blood serum biomarkers reveal differential clinical proteomics of the neurodegenerative diseases. Expert Rev Proteomics 2006;3:45-62.

[8] Jacob AM, Turck CW. Detection of post-translational modifications by fluorescent staining of two-dimensional gels. Methods Mol Biol 2008;446:21-32.

[9] Jacobs JM, Adkins JN, Qian WJ, Liu T, Shen Y, Camp DG 2nd, Smith RD. Utilizing human blood plasma for proteomic biomarker discovery. J Proteome Res 2005;4:1073-85.

[10] Surinova S, Schiess R, Hüttenhain R, Cerciello F, Wollscheid B, Aebersold R. On the development of plasma protein biomarkers. J Proteome Res 2011;10:5-16.

[11] Rosenberg LH, Franzén B, Auer G, Lehtiö J, Forshed J. Multivariate meta-analysis of proteomics data from human prostate and colon tumours. BMC Bioinformatics 2010;11:468.

[12] Natale M, Bonino D, Consoli P, Alberio T, Ravid RG, Fasano M, Bucci EM. A meta-analysis of two-dimensional electrophoresis pattern of the Parkinson's disease-related protein DJ-1. Bioinformatics 2010;26:946-52.

[13] Ioannidis JPA, Khoury MJ. Improving Validation Practices in "Omics" Research. Science 2011;334:1230-32.

[14] Griss J, Haudek-Prinz V, Gerner C. GPDE: A biological proteomic database for biomarker discovery and evaluation. Proteomics 2011;11:1000-4.

[15] Albrecht D, Kniemeyer O, Brakhage AA, Guthke R. Missing values in gel-based proteomics. Proteomics 2010;10:1202-11.

[16] McDonald JH. Handbook of Biological Statistics. 2nd ed. Baltimore: Sparky House Publishing; 2009.

[17] Alberio T, Pippione AC, Comi C, Olgiati S, Cecconi D, Zibetti M, Lopiano L, Fasano M. Dopaminergic therapies modulate the T-CELL proteome of patients with Parkinson's disease. IUBMB Life 2012;64:846-52.

[18] R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2009.

[19] Alberio T, Bossi AM, Milli A, Parma E, Gariboldi MB, Tosi G, Lopiano L, Fasano M. Proteomic analysis of dopamine and α-synuclein interplay in a cellular model of Parkinson's disease pathogenesis. FEBS J 2010;277:4909-19.

[20] Tansey MG, McCoy MK, Frank-Cannon TC. Neuroinflammatory mechanisms in Parkinson's disease: potential environmental triggers, pathways, and targets for early therapeutic intervention. Exp Neurol 2007; 208:1-25.

[21] Argüelles S, Venero JL, García-Rodriguez S, Tomas-Camardiel M, Ayala A, Cano J, Machado A. Use of haptoglobin and transthyretin as potential biomarkers for the preclinical diagnosis of Parkinson's disease. Neurochem Int 2010;57:227-34.

[22] Goldknopf IL, Bryson JK, Strelets I, Quintero S, Sheta EA, Mosqueda M, Park HR, Appel SH, Shill H, Sabbagh M, Chase B, Kaldjian E, Markopoulou K. Abnormal serum concentrations of proteins in Parkinson's disease. Biochem Biophys Res Commun 2009; 389:321-7.

[23] Galicia G, Maes W, Verbinnen B, Kasran A, Bullens D, Arredouani M, Ceuppens JL. Haptoglobin deficiency facilitates the development of autoimmune inflammation. Eur J Immunol 2009;39:3404-12.

[24] Costa-Mallen P, Checkoway H, Zabeti A, Edenfield MJ, Swanson PD, Longstreth WT Jr, Franklin GM, Smith-Weller T, Sadrzadeh SM. The functional polymorphism of the hemoglobin-binding protein haptoglobin influences susceptibility to idiopathic Parkinson's disease. Am J Med Genet B Neuropsychiatr Genet 2008;147B:216-22.

[25] Finehout EJ, Franck Z, Lee KH. Complement protein isoforms in CSF as possible biomarkers for neurodegenerative disease. Dis Markers 2005;21:93-101.

[26] Cannon JR, Hua Y, Richardson RJ, Xi G, Keep RF, Schallert T. The effect of thrombin on a 6-hydroxydopamine model of Parkinson's disease depends on timing. Behav Brain Res 2007;183:161-8.

[27] Sokolova E, Reiser G. Prothrombin/thrombin and the thrombin receptors PAR-1 and PAR-4 in the brain: localization, expression and participation in neurodegenerative diseases. Thromb Haemost 2008;100:576-81.

[28] Riek-Burchardt M, Striggow F, Henrich-Noack P, Reiser G, Reymann KG. Increase of prothrombin-mRNA after global cerebral ischemia in rats, with constant expression of protease nexin-1 and protease-activated receptors. Neurosci Lett 2002;329:181-4.

[29] Hu YQ, Liu BJ, Dluzen DE, Koo PH. Alteration of dopamine release by rat caudate putamen tissues superfused with alpha 2-macroglobulin. J Neurosci Res 1996;43:71-7.

[30] Nicoletti G, Annesi G, Tomaino C, Spadafora P, Pasqua AA, Annesi F, Serra P, Caracciolo M, Messina D, Zappia M, Quattrone A. No evidence of association between the alpha-2 macroglobulin gene and Parkinson's disease in a case-control sample. Neurosci Lett. 2002;328(1):65-7.

[31] Dufek M, Hamanová M, Lokaj J, Goldemund D, Rektorová I, Michálková Z, Sheardová K, Rektor I. Serum inflammatory biomarkers in Parkinson's disease. Parkinsonism Relat Disord 2009;15:318-20.

Table 1: Summary of enrolled subjects

	Controls (n = 45)	*PD (n = 45)*
Anamnestics		
Age ± SD (years)	69.1 ± 8.7	70.3 ± 7.7
Gender (males)	23 (51%)	26 (58%)
PD duration ± SD (years)		7.3 ± 4.3
Medications		
Unmedicated		0
L-DOPA		5
DA agonists		12
L-DOPA + DA agonists		28
Hoehn and Yahr stage		
1		14 (31%)
2		14 (31%)
3		13 (29%)
4		3 (7%)
5		1 (2%)

Table 2: Identity of protein spots

Match ID [a]	Protein name	Gene Symbol [b]	pI [c]	MW [c]	p Value [d]	Fold of Change [e]
22	Haptoglobin Beta Chain	**HP**	5.08	41412	7.5×10^{-05}	0.654
23	Haptoglobin Beta Chain	**HP**	4.97	42114	5.9×10^{-03}	0.715
24	Haptoglobin Beta Chain	**HP**	4.81	44464	2.8×10^{-07}	0.630
25	Haptoglobin Beta Chain	**HP**	5.21	41329	3.9×10^{-04}	0.715
26	Serum albumin	ALB	5.43	41412	1.4×10^{-02}	1.162
34	Ig mu chain C region	IGHM	6.05	48501	1.9×10^{-02}	1.145
38	Transthyretin	**TTR**	5.52	13800	1.7×10^{-02}	1.187
41	Tetranectin	CLEC3B	5.38	19391	1.6×10^{-04}	1.316
43	Haptoglobin Alpha Chain	**HP**	6.07	16882	5.6×10^{-04}	0.697
44	Alpha-1-microglobulin	AMBP	5.07	30920	4.0×10^{-02}	0.871
45	Serum albumin	ALB	6.06	58451	2.4×10^{-02}	1.223
46	Fibrinogen beta chain	FGB	6.41	55252	4.4×10^{-02}	0.869

61	Hemopexin	HPX	5.43	74551	4.7×10^{-02}	0.848
62	Hemopexin	HPX	5.53	73703	4.5×10^{-02}	0.863
66	Complement C4-A or B	C4	5.9	41412	1.1×10^{-02}	0.783
92	Haptoglobin Beta Chain	**HP**	5.38	38749	6.3×10^{-04}	0.772
97	Complement factor H	**CFH**	5.68	99064	5.0×10^{-04}	1.402
105	Alpha-2-macroglobulin	A2M	5.44	143064	3.3×10^{-02}	0.750
113	Apolipoprotein A-I	**APOA1**	5.22	23000	3.1×10^{-03}	1.119
117	Complement C3	**C3**	4.84	40915	2.0×10^{-02}	0,911
132	Haptoglobin Beta Chain	**HP**	4.88	43156	4.5×10^{-04}	0.672
141	Apolipoprotein E	**APOE**	5.53	33925	3.9×10^{-03}	0.699
147	Apolipoprotein A-IV	APOA4	5.16	43627	8.7×10^{-04}	1.306
150	Immunoglobulin light chain	IGHL	7.03	26973	3.3×10^{-02}	1,200
151	Haptoglobin Alpha Chain	**HP**	5.4	16882	2.3×10^{-04}	0.766

[a] Spots are numbered according to map detection and matching.

[b] Official symbol from http://www.ncbi.nlm.nih.gov/gene. Symbols in bold refer to gene products retrieved by the automated literature analysis.

[c] Predicted pI and Mr according to protein sequence as computed by the Compute pI/Mw tool (http://web.expasy.org/compute_pi/).

[d] Based on Wilcoxon test.

[e] Expressed as PD/control spot volume ratios.

Table 3 Mass Spectrometry identification of selected spots

Match ID	Protein name	Uniprot ID [a]	Mr. (kDa) theor [b]	pI theor [b]	Mr. (kDa) exp [c]	pI exp [c]	No. of peptides identified	MOWSE score [d]	Sequence Coverage (%)
26	Serum albumin	P02768	71	5.92	41	5.43	11	965	35%
41	Tetranectin	P05452	23	5.52	20	5.3	2	80	16%
45	Serum albumin	P02768	68	5.67	59	6.06	12	766	53%
66	Complement C4	P0C0L4	48	5.78	40	6.0-6.5	4	175	20%
105	Alpha-2-macroglobulin	P01023	162	5.95	120	5.5	19	1075	35%

[a] http://www.uniprot.org.

[b] Predicted pI and Mr according to protein sequence as computed by Mascot search results (http://www.matrixscience.com).

[c] See [12].

[d] http://www.matrixscience.com.

Caption to the Figures

Figure 1: Statistical assessment of the quality of 2-DE maps. Spot volumes have linear correlation coefficients greater than 0.8 when technical replicates of the same specimen are compared, as indicated by the box-and-whiskers plot (A). Spot volumes show log-normal distributions both in control subjects (filled circles) and in PD patients (open circles) (B).

Figure 2: Representative 2-DE map of human plasma. Spots showing significantly different abundance are contoured in green. Gene symbols are reported only for spots belonging to the nine candidates list. HPα and HPβ refers to the α and β chains of haptoglobin.

Figure 3: Discriminant analysis of the 15 spots that best discriminate PD patients. Hierarchical clustering of spot intensities shows two outgroups. One of them groups together all HP isoforms, while the other includes all the other spots and is enriched in apolipoproteins (A). Panels B and C report prediction results by the "leave-one-out" cross-validation procedure in terms of PD likelihood score distribution in the two groups (B) and classification performance, represented by the ROC curve (C).

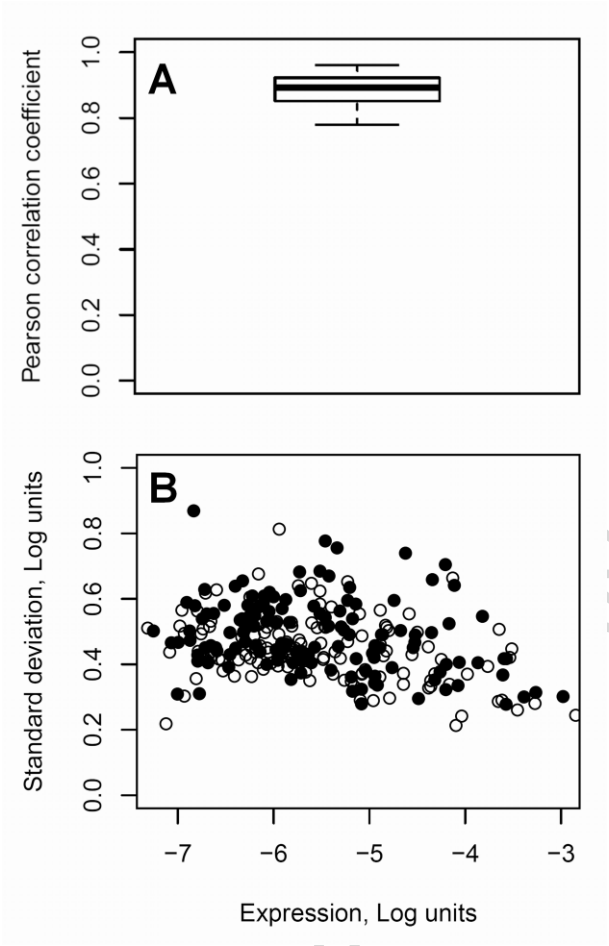


Figure 1

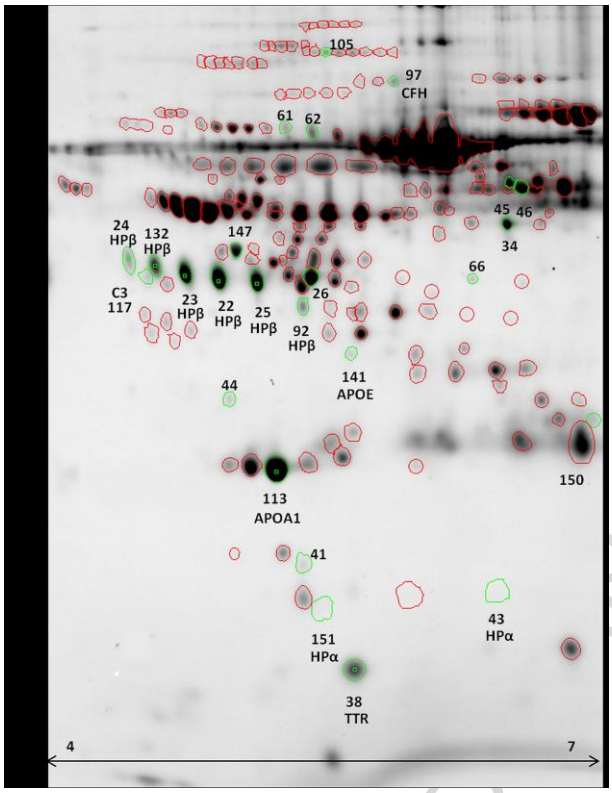


Figure 2

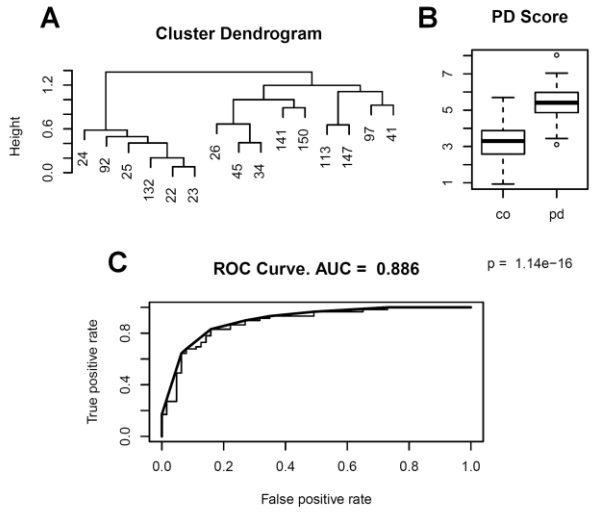
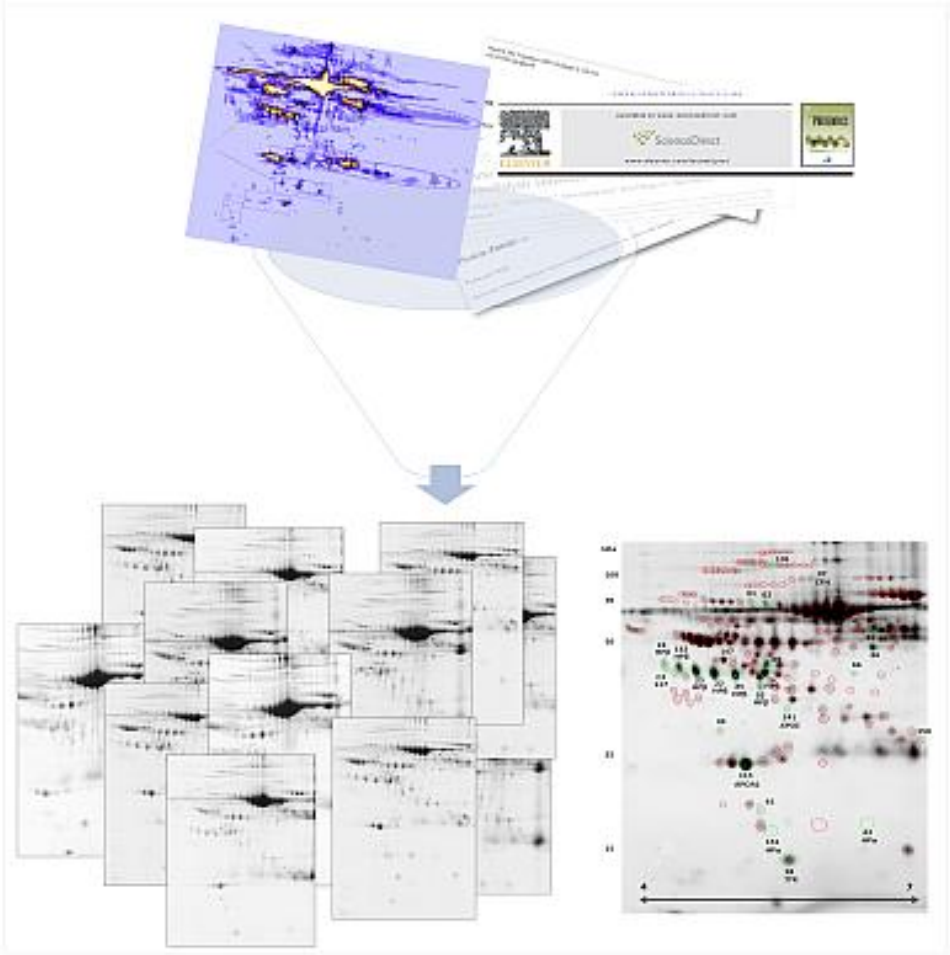


Figure 3



Graphical Abstract

Highlights

- An automated literature analysis procedure retrieved plasma PD biomarkers.
- 9 different proteins were identified as a potential PD diagnostic pattern.
- We verified some candidate biomarkers in 2-DE plasma maps of 90 subjects.