In-depth Exploration of Cerebrospinal Fluid by Combining Peptide Ligand Library Treatment and Label-free Protein Quantification*

Emmanuelle Mouton-Barbosa, a,b,c Florence Roux-Dalvai, a,b,c David Bouyssié, a,b François Berger, d Eric Schmidt, e Pier Giorgio Righetti, f,g Luc Guerrier, h Egisto Boschetti, h Odile Burlet-Schiltz, a,b Bernard Monsarrat, a,b,j and Anne Gonzalez de Peredo a,b,j

Cerebrospinal fluid (CSF) is the biological fluid in closest contact with the brain and thus contains proteins of neural cell origin. Hence, CSF is a biochemical window into the brain and is particularly attractive for the search for biomarkers of neurological diseases. However, as in the case of other biological fluids, one of the main analytical challenges in proteomic characterization of the CSF is the very wide concentration range of proteins, largely exceeding the dynamic range of current analytical approaches. Here, we used the combinatorial peptide ligand library technology (ProteoMiner) to reduce the dynamic range of protein concentration in CSF and unmask previously undetected proteins by nano-LC-MS/MS analysis on an LTQ-Orbitrap mass spectrometer. This method was first applied on a large pool of CSF from different sources with the aim to better characterize the protein content of this fluid, especially for the low abundance components. We were able to identify 1212 proteins in CSF, and among these, 745 were only detected after peptide library treatment. However, additional difficulties for clinical studies of CSF are the low protein concentration of this fluid and the low volumes typically obtained after lumbar puncture, precluding the conventional use of ProteoMiner with large volume columns for treatment of patient samples. The method has thus been optimized to be compatible with low volume samples. We could show that the treatment is still efficient with this miniaturized protocol and that the dynamic range of protein concentration is actually reduced even with small amounts of beads, leading to an increase of more than 100% of the number of identified proteins in one LC-MS/MS run. Moreover, using a dedicated bioinformatics analytical work flow, we found that the method is reproducible and applicable for label-free quantification of series of samples processed in parallel. Molecular & Cellular Proteomics 9:1006–1021, 2010.

The identification of biological markers heralding a pathological condition represents a major challenge in clinical proteomics. In this context, body fluids are a particularly interesting material source because their collection is minimally invasive compared with tissue biopsies. Indeed, in the discovery phase, the selection of a fluid in close proximity to a diseased organ may increase the probability of finding a biomarker originating from a pathological tissue. In that respect, cerebrospinal fluid (CSF), which is in closest contact with the brain, represents a useful reservoir of potential clinically relevant biomarkers for neurological diseases. Although acquisition of CSF via lumbar puncture is an invasive procedure, it is less invasive and more readily obtainable than a brain biopsy.

CSF has several functions, including buoyancy, protection of the brain against pressure gradients, transport of active biological substances for normal maintenance of the brain, and excretion of toxic and waste substances (1, 2). It surrounds the exterior of the central nervous system, filling also the four large ventricular cavities inside the brain, the spinal canal, and subarachnoid spaces. CSF is produced mainly in the choroid plexus, a structure present in the four ventricles, consisting of a single layer of epithelial cells surrounding a core of connective tissue and blood capillaries. These epithelial cells have tight junctions on the side facing the ventricle that prevent the majority of blood substances from crossing the cell layer into the CSF and form what is known as the blood-CSF barrier. They produce CSF through a secretary

1 The abbreviations used are: CSF, cerebrospinal fluid; ECF, extracellular fluid; HOS, hydro-organic solution; PAI, protein abundance index; FDR, false discovery rate; XIC, extracted ion chromatogram; RT, retention time; 2D, two-dimensional; 1D, one-dimensional; nano-LC, liquid nanochromatography; LTQ, linear trap quadrupole; IPI, International Protein Index; MFPaQ, Mascot File Parsing and Quantification; SLITRK, SLIT and NTRK-like proteins; NTRK, Neurotrophic Tyrosine Kinase Receptor.
process based on active ion transport mechanisms that allow maintenance of a constant ionic composition in the CSF and thus create a stable ionic environment for the brain independent of plasma ion concentration that can fluctuate significantly (3). The protein content of the CSF produced in the choroid plexus originates mainly from blood proteins, which slowly cross the epithelial layer by a diffusion mechanism dependent on molecular size (1, 4). Because of the presence of this barrier, however, the final protein concentration in CSF is around 200 times lower than in plasma. Some CSF proteins, on the other hand, are known to be synthesized and actively secreted by the epithelial cells of the choroid plexus, such as transthyretin or cystatin C (5).

In addition, a small amount of the CSF is not produced by the choroid plexus but originates from the extracellular fluid (ECF) that fills the extracellular space of the brain. This ECF is connected with the CFS via the perivascular space also known as Virchow-Robin space. Indeed, CSF also extends into the sulci and the depth of the cerebral cortex along the blood vessels in these perivascular spaces in which small molecules diffuse freely. ECF also flows into the ventricles by crossing the ependyma cell barrier between brain and CSF. ECF carries proteins derived directly from the brain that have been estimated to represent about 20% of the final CSF protein content (6). Therefore, there is a continuous movement of metabolites from deep parenchyma to cortical subarachnoid space and ventricular system. In that respect, CSF is also a biochemical window into the brain and constitutes an interesting source of potential neurological biomarkers.

Protein analysis in CSF has been actively performed during the past years to discover species relevant to a disease. Many validated clinical assays, each directed toward a particular protein and mostly based on antibody detection methods, have been developed to investigate protein candidates as potential markers of several pathologies. These focused measurements (for a review, see Ref. 1) allowed a precise quantification of the selected proteins in CSF and have shown that the dynamic range of protein concentrations spans at least 8 orders of magnitude between the most abundant protein (albumin at 130–350 mg/liter, representing alone around 45% of CSF protein content) and the lowest detected proteins in such assays in the ng/liter range. This very large dynamic range represents, as for other body fluids, a major analytical challenge and has largely hampered comprehensive studies aiming at global CSF proteomic characterization. The first studies were based on 2D gels, which provided a useful tool to establish CSF maps, analyze protein modifications, and perform differential profiling (7–10). However, the relatively low dynamic range of the technique has usually limited the total number of detectable unique proteins to around 40 abundant species. Liquid nanochromatography (nano-LC)-MS-based technologies allowed a significant increase of this number during the past few years, and different studies reported several hundreds of unique proteins confidently identified in CSF (11–14). To reach this number, extensive sample fractionation was usually applied at the protein or peptide level to lower the concentration dynamic range of each individual fraction as well as to increase the MS/MS sequencing time and analytical coverage of the sample. Most of these studies, however, remained largely qualitative or restricted to the quantitative differential analysis of a small number of isotopically labeled samples. Only recently, the development of label-free approaches, based on the direct comparison of MS peptide signals across different nano-LC-MS runs using appropriate bioinformatics tools, has opened the way to the quantitative profiling of large series of patient samples for clinical studies with mass spectrometry techniques (15). Thanks to the high resolution of recent mass spectrometers, such strategies seem now more accurate and promising for fluids proteomic profiling. However, a compromise has yet to be found between a labor-intensive prefractionation of the samples and the number of patients included in the study, which should be ideally high for statistical significance.

Here, we performed a proteomic characterization of CSF using the combinatorial peptide ligand library technology, which has been described recently as an efficient approach to decrease the concentration dynamic range of complex protein mixtures (16–19). This method is based on the treatment of the sample with combinatorial libraries of peptide ligands bound to porous beads. Each bead contains a unique hexapeptide ligand distributed throughout its porous structure, and any protein in the starting material can theoretically interact with one or a few beads among the wide diversity (10–206) of ligand beads from the library. Once the most abundant protein species have saturated their binding sites, the remaining molecules are washed away in the flow-through, whereas minor protein species are progressively enriched on their corresponding beads. This strategy has been efficiently applied to capture a very large population of previously undetected proteins in several types of samples, such as urine (20), serum (21), and platelets (22), and we used it recently for the extensive characterization of the red blood cell proteome (23). Using a large pool of CSF from different sources, this method, associated with fractionation and nano-LC-MS/MS analysis, allowed here the detection of more than a thousand proteins in CSF. Moreover, we evaluated several technical features relevant for its use in clinical studies, such as the conservation of relative quantitative information among parallel sample treatments, and the applicability on low CSF volumes typically obtained from lumbar puncture. We found that the miniaturized protocol is efficient and reproducible and allows a significant increase of the number of quantified proteins in the peptide library-treated, but unfractionated sample. It could thus offer an interesting alternative for the rapid and in-depth profiling of large series of CSF patient samples, avoiding the use of extensive prefractionation operations. It provides one of the first clinical compatible proteomics strat-
egies targeting the deep proteome for CSF biomarker discovery, which should be a crucial biomedical issue for the exploration and management of brain diseases.

**EXPERIMENTAL PROCEDURES**

*Materials*—The solid-phase combinatorial peptide library called ProteoMiner™ (NH₂-Library) and its carboxylated version (COOH-Library) are both from Bio-Rad. The former was purchased as a commercial product; the second is not commercially available and was a gift of the company. Complete protease inhibitor mixture tablets were from Roche Diagnostics. Sequencing grade trypsin was from Promega (Madison, WI), N-Ethylmaleimide, urea, thiourea, CHAPS, isopropanol, acetonitrile, trifluoroacetic acid, and sodium dodecyl sulfate were all from Sigma-Aldrich. All other chemicals were also from Aldrich and were of analytical grade. C₁₉ precolumns and analytical columns were from Dionex (Amsterdam, The Netherlands). Equipment and reagents for mono- and bidimensional electrophoresis were both from Bio-Rad.

*CSF Collection*—Human CSF used in this study was a pool obtained from various samples collected in two different hospital centers. Some patients were from the University Hospital of Grenoble, France where collection followed the French regulation and was approved at the national level (molecular neurosurgery bank) with approval number AC-2007-23, and others were from the University Hospital Purpan, Toulouse, France where collection was approved by the local ethical review board (Comité de Protection des Personnes Sud Ouest II) and declared at the national level (Direction Générale de la Recherche et de l’Innovation) under the reference DC-2008-483.

For the large CSF study (treatment on 1-ml ligand bead columns), 1290 ml of CSF were gathered from a pool of 250 patients (Pool 1). Most of the patients underwent a lumbar puncture in the context of headache and/or suspicion of a neurological disease not confirmed by magnetic resonance imaging and CSF study (n = 199). Volumes of samples obtained from these patients were 0.2 ml (n = 100) or 0.5 ml (n = 99). Another group of patients suffered from hydrocephalus, i.e., an excess of CSF (n = 49), and a tap test was performed to withdraw CSF via lumbar puncture. Volumes of the samples obtained in this way were 0.5 (n = 1), 1 (n = 8), 3 (n = 30), 6 (n = 5), and 12 ml (n = 5). In addition, large volumes were obtained for samples collected using an overnight CSF drainage protocol for two patients suffering from hydrocephalus (108, 150, 425, and 350 ml). All samples came from the CSF discarded after biochemical evaluation. For the assays on low volumes of beads, another pool of CSF samples was used (Pool 2) from 40 patients suffering from headache and investigated by lumbar puncture for a potential neurological disease not confirmed by magnetic resonance imaging and CSF study. The volume of samples collected in this way was 0.5 ml per individual. In every case, we checked the absence of cellular abnormalities in CSF. We also checked the absence of an increased total proteomic content to eliminate pathologies with inflammatory reaction as well as blood-brain barrier abnormalities. Infectious diseases as well as tumor formations and meningitis were eliminated. All samples were taken with the informed consent of the patients and were frozen immediately after clinical laboratory analysis.

*Treatment of CSF Proteins with Hexapeptide Ligand Libraries*—About 1290 ml of frozen pooled human CSF (CSF Pool 1; 0.42 mg/ml protein concentration) was thawed at 4 °C, added with two tablets of protease inhibitor mixture, and dialyzed against 50 mM ammonium bicarbonate overnight at +4 °C. The dialyzed material was then centrifuged at 5000 × g at 4 °C for 30 min to obtain a clear protein solution and lyophilized. CSF proteins were solubilized in 70 ml of physiological PBS and then loaded on a column containing 1 ml (6.6-mm inner diameter × 32 mm in length) of NH₂-Library at a flow rate of 0.25 ml/min. The column effluent was directly injected in a second column of the same dimensions packed with COOH-Library. The columns connected in series were then washed with PBS until UV base line at 280 nm of the effluent of the second column. After the wash, each individual column was independently subjected to three distinct elutions of captured proteins using, respectively, TUC solution (2 n thiourea, 7 M urea, 2% CHAPS), UCA solution (9 M urea, citric acid up to pH 3.3) and a hydro-organic solution (HOS) composed of 6% (v/v) acetonitrile, 12% (v/v) isopropanol, 10% (v/v) ammonia at 20%, and 72% (v/v) water. The six eluates were immediately neutralized (second and third eluates), dialyzed against 20 mM ammonium bicarbonate (cutoff of dialysis membrane was 1000 Da), and then lyophilized. Dry protein samples were then stored at −20 °C waiting for further analysis. Small aliquots were taken for protein assay. The amount of protein obtained from each fraction was: 2352, 7342, and 440 μg from NH₂-Library for TUC, acidic urea, and hydro-organic eluates, respectively, and 1422, 3572, and 88 μg from COOH-Library for TUC, acidic urea, and hydro-organic eluates, respectively.

**1D SDS-PAGE Fractionation and Nano-LC-MS/MS Analysis of CSF Proteins**—One hundred and fifty micrograms of each elution fraction from the two library columns as well as 150 μg of the initial non-treated human CSF and 150 μg of the flow-through were diluted in Laemml buffer and boiled for 5 min before being separated on a 12% acrylamide SDS-PAGE gel. Proteins were visualized by Coomassie Blue staining. Each lane was cut into 20 homogenous slices that were washed in 100 mM ammonium bicarbonate for 15 min at 37 °C followed by a second wash in 100 mM ammonium bicarbonate, acetonitrile (1:1) for 15 min at 37 °C. Reduction and alkylation of cysteine residues were performed by mixing the gel pieces in 10 mM DTT for 35 min at 56 °C followed by 55 mM iodoacetamide for 30 min at room temperature in the dark. An additional cycle of washes in ammonium bicarbonate and ammonium bicarbonate/acetonitrile was then performed. Proteins were digested by incubating each gel slice with 0.6 μg of modified sequencing grade trypsin in 50 mM ammonium bicarbonate overnight at 37 °C. The resulting peptides were extracted from the gel by three steps: a first incubation in 50 mM ammonium bicarbonate for 15 min at 37 °C and two incubations in 10% formic acid, acetonitrile (1:1) for 15 min at 37 °C. The three collected extractions were pooled with the initial digestion supernatant, dried in a SpeedVac, and resuspended with 14 μl of 5% acetonitrile, 0.05% trifluoroacetic acid.

The peptides mixtures were analyzed by nano-LC-MS/MS using an Ultimate3000 system (Dionex) coupled to an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Five microliters of each sample were loaded on a C₁₉ precolumn (300-μm inner diameter × 5 mm; Dionex) at 20 μl/min in 5% acetonitrile, 0.05% trifluoroacetic acid. After 5 min of desalting, the precolumn was switched on line with the analytical C₁₉ column (75-μm inner diameter × 15 cm; PepMap C₁₉, Dionex) equilibrated in 95% solvent A (5% acetonitrile, 0.2% formic acid) and 5% solvent B (80% acetonitrile, 0.2% formic acid). Peptides were eluted using a 5–50% gradient of solvent B during 80 min at a 300 nl/min flow rate. The LTQ-Orbitrap was operated in data-dependent acquisition mode with the Xcalibur software. Survey scan MS spectra were acquired in the Orbitrap on the 300–2000 m/z range with the resolution set to a value of 60,000. The five most intense ions per survey scan were selected for CID fragmentation, and the resulting fragments were analyzed in the linear trap (LTQ). Dynamic exclusion was used within 60 s to prevent repetitive selection of the same peptide.

**Quantitative Measurements of Proteins Spiked in Serum after Peptide Library Treatment**—Human serum samples (5 ml) were spiked with four proteins (bovine β-lactoglobulin, bovine β-casein, bovine κ-casein, and rabbit phosphorylase b, all purchased from Sigma-Aldrich) at four different final concentrations (20 nm, 200 nm, 2 μM, and 20 μM). This was done in triplicate for each spiked protein amount. Each of the 12 samples was then loaded on a spin column containing...
0.5 ml of ProteoMiner beads and incubated with rotation for 2 h. The beads were then washed three times with PBS. After the wash, each individual column was independently subjected to elution with 1.25 ml of 2× Laemmli sample buffer (80 mM Tris-HCl, pH 6.8, 20% glycerol, 4% SDS, 50 mM DTT). Aliquots of 50 μl each of the sample were loaded on a gel for 1D SDS-PAGE and migrated through the stacking gel until the top of the separating gel, and proteins were concentrated into one gel band. Tryptic peptides were prepared and analyzed in one nano-LC-MS/MS run as described above. To quantify the spiked proteins after peptide library treatment, the extracted ion chromatogram (XIC) signal of some well-characterized tryptic peptide ions from these proteins was manually extracted from the MS survey of nano-LC-MS/MS raw files using the Xcalibur software. XIC areas were integrated in Xcalibur under the QualBrowser interface using the ICIS algorithm. Mean values and S.D. were calculated for triplicate measurements.

**Treatment of CSF Samples on Small Volumes of Peptide Library Beads**—A pooled CSF sample of 20 ml (CSF Pool 2) was divided into 2-ml aliquots that were added to either 5 (four replicates) or 2 μl (four replicates) of PBS-equilibrated ProteoMiner beads in 2-ml centrifugation tubes. The CSF-bead suspensions were gently stirred at 4 °C for 5 h, and then the beads were collected by centrifugation and washed twice with PBS. The proteins captured by the beads were eluted by addition of 15 μl of Laemmli sample buffer for electrophoresis and boiling 2 min at 95 °C. Proteins were loaded on a gel for 1D SDS-PAGE, and the whole protein mixture was concentrated into a unique gel band. Tryptic peptides were prepared and analyzed in one nano-LC-MS/MS run as described above. For comparison, an aliquot of crude CSF (10 μl; Pool 2) was prepared, digested, and analyzed in the same way.

**Database Search and Data Analysis**—The Mascot Daemon software (version 2.2.0, Matrix Science, London, UK) was used to perform database searches in batch mode with all the raw files acquired on each sample. To automatically extract peak lists from Xcalibur raw files, the Extract.msn.exe macro provided with Xcalibur (version 2.0 SR2, Thermo Fisher Scientific) was used through the Mascot Daemon interface. The following parameters were set for creation of the peak lists: parent ions in the mass range 400–4500, no grouping of MS/MS scans, and threshold at 1000. A peak list was created for each analyzed fraction (i.e., gel slice), and individual Mascot searches were performed for each fraction. Data were searched against all entries in the IPI human v3.61 protein database (82,631 sequences). Carbamidomethylation of cysteines was set as a fixed modification, and oxidation of methionine and protein N-terminal acetylation were set as a variable modifications for all Mascot searches. Specificity of trypsin digestion was set for cleavage after Lys or Arg except before Pro, and one missed trypsin cleavage site was allowed. The mass tolerances in MS and MS/MS were set to 5 ppm and 0.6 Da, respectively, and the instrument setting was specified as “ESI-Trap.” Mascot results were parsed with the in-house developed software Mascot File Parsing and Quantification (MFPaQ) version 4.0 (24), and protein hits were automatically validated if they satisfied one of the following criteria: identification with at least one top ranking peptide of a minimal length of 5 amino acids and with a Mascot score higher than the identity threshold at p = 0.001 (99.9% probability) or at least two top ranking peptides each of a minimal length of 5 amino acids and with a Mascot score higher than the identity threshold at p = 0.05 (95% probability). All the annotated MS/MS spectra corresponding to proteins identified with a unique peptide are shown in supplemental Data 5 and 6. To calculate the false discovery rate (FDR), the search was performed using the “decoy” option in Mascot, and MFPaQ used the same criteria to validate decoy and target hits. The FDR was calculated at the protein level (FDR = number of validated decoy hits/(number of validated target hits + number of validated decoy hits) × 100), and using the specified validation criteria, it ranked between 0 and 0.8% for all the samples analyzed with an average value of 0.4%. When several proteins matched exactly the same set of peptides, only one member of the protein was reported in final protein lists in supplemental Data 2 and 3 for more clarity (the one returned by Mascot in the protein summary list), but detailed protein groups are displayed in an additional sheet. Moreover, in each Mascot result file, the MFPaQ software detected highly homologous Mascot protein hits, i.e., proteins identified with top ranking MS/MS queries also assigned to another protein hit of higher score. These homologous protein hits were validated and included in the final list only if they were additionally assigned a specific top ranking peptide with Mascot score higher than the identity threshold at p = 0.001. In the case of sample fractionation by 1D SDS-PAGE, MFPaQ was used to create a unique non-redundant protein list from the identification results of each fraction (i.e., gel slice). Clustering of proteins was performed based on peptide sharing by grouping together all protein sequences matching the same set of peptides (only top ranking peptides with a Mascot score higher than the identity threshold at p = 0.05 were considered). To merge or compare lists of protein groups from different samples, clustering was performed based on accession number sharing (clusters of protein groups are created if they have one common member).

**Label-free Quantification**—Quantification of proteins was performed using the label-free module implemented in the MFPaQ v4.0.0 software (SourceForge). For each sample, the software uses the validated identification results (using the validation criteria described above) and XICs of the identified peptides in the corresponding raw nano-LC-MS files based on their experimentally measured retention time (RT) and monoisotopic m/z values. Only top ranking peptides with a Mascot score higher than the identity threshold at p = 0.05 were selected for quantification. If some peptide ions were sequenced by MS/MS and validated only in some of the samples to be compared, their XIC signal was extracted in the nano-LC-MS raw file of the other samples using a predicted RT value calculated as follows. The software selects all the peptide ions that are identified by MS/MS in all the LC-MS/MS runs that are to be aligned. It extracts the XIC signal of each of these peptide ions in each run (within a time interval around the MS/MS sequencing time), defines their retention time as the apex of their elution peak, and stores all these time values in a calibration matrix. Then, for all peptide ions that were identified by MS/MS in only some of the runs, this matrix can be used to predict their retention time in the other runs by a linear interpolation method. Quantification of peptide ions was performed based on calculated XIC areas values. For pairwise comparison, the software computed peptide ratios (defined as the average of the ratios from peptide ions of different charge state) and protein ratios (defined as the average of the ratios of the peptides assigned to the protein). These mean values were calculated from an arithmetic averaging of the ratios, after applying a transformation to the asymmetric distribution of ratio values, to make it symmetric and centered around 0: x – 1 for ratios higher than 1 and 1 – (1/x) for ratios smaller than 1. The arithmetic average was then computed, and the final mean ratio was obtained by applying the reverse transformation to this value. Systematic experimental errors were corrected by dividing each protein ratio by the median of all the ratios. When multiple replicate analyses were performed, the coefficient of variation of peptide ion XIC areas was calculated over the different replicates after normalization of the area value for each nano-LC-MS run (normalization was based on the total sum of all the XIC areas extracted by the software in the run). To compare abundances of different proteins or to represent the abundance profile of one protein in different samples, a protein abundance index was calculated, defined as the average of XIC area values for three intense reference tryptic peptides identified for this protein (the three peptides exhibiting the highest intensities across the different replicates are...
selected as reference peptides and used to compute the protein abundance index (PAI) of this protein in each replicate).

To optimize the number of quantified proteins in replicate nano-LC-MS/MS analyses, quantification was also performed with the use of previously built identification databases, containing m/z and RT values associated with peptide sequences, which were subsequently used to extract MS signals in individual runs. For each quantitative experiment, a small database was thus created from the same sample that was to be analyzed and quantified using the label-free module of MFPaQ. This module was first tested on replicate nano-LC-MS runs of the first eluate from the NH2-Library (obtained from treatment of CSF Pool 1), and a database was created from an SDS-PAGE shotgun analysis of this sample (150 μg fractionated into 20 gel slices). To this aim, the fractions were analyzed by nano-LC-MS/MS as described above, the sequenced and validated peptide ions were stored in MFPaQ, and this database was used to quantify proteins identified in triplicate nano-LC-MS/MS analysis of 4 μg of either crude or treated CSF (TUC eluate from NH2-Library). Another database was built in the same way to check the reproducibility of the miniaturized protocol and quantify proteins detected in nano-LC-MS runs of replicate samples from CSF Pool 2 treated on small volumes of beads. This was done by shotgun analysis of a 2-ml CSF aliquot (Pool 2) treated on 5 μL of beads and fractionated into 20 gel slices.

RESULTS

Treatment of Large Volume of Pooled CSF Samples with Peptide Libraries—To assess the efficiency of ProteoMiner treatment on CSF, an extensive proteomics study was performed using an experimental scheme previously described for other biological fluids (22, 23) (Fig. 1). The sample was loaded on a sequence of two columns containing two different peptide ligand libraries, one composed of an N-terminal collection of hexapeptides synthesized by combinatorial chemistry (NH2-Library), and the other one containing the same peptides in which the N terminus has been modified to a carboxylate group (COOH-Library). The use of this second library has been shown before to be beneficial for capturing different species present in the sample compared with the NH2-Library and for finally increasing the number of identified proteins (22, 23). Moreover, to optimize the number of protein species captured by both combinatorial libraries, a high diversity of hexapeptide ligands is typically used to treat the samples. Thus, a volume of 1 ml of ligand beads was used here in each column as described in previous reports. As a large overloading of total bead capacity has to be performed to saturate abundant proteins and obtain a sufficient enrichment of low abundance species, large protein amounts are generally necessary when using this classical protocol. This was achieved in the case of CSF by pooling samples from different sources to yield a total amount of 770 mg of CSF proteins. The captured proteins were then collected by eluting each column sequentially with TUC, UCA, and HOS buffers. The six resulting eluates were analyzed by SDS-PAGE as were the non-treated CSF and the flow-through from the columns (150 μg from each sample). As shown in Fig. 2A, a very intense band of albumin is observed in the latter two samples. Very different patterns were obtained for the column eluates, which showed a large decrease of the albumin band and the apparition of many new protein bands. This was confirmed by analysis of the samples using two-
dimensional electrophoresis. Although the 2D pattern of initial untreated CSF shows a relatively low number of protein spots (mainly albumin as well as light and heavy chains of immunoglobulins), a high number of new protein spots emerge in the patterns from fractions of library-treated samples (supplemental Data 1).

The lanes of the 1D gel corresponding to the six elution fractions and to the crude CSF were then each cut into 20 gel slices, digested with trypsin, and analyzed by nano-LC-MS/MS on an LTQ-Orbitrap mass spectrometer. Proteins identified by Mascot were validated using stringent criteria to yield a false discovery rate around 0.4% at the protein level, and concatenated lists of unique protein groups were generated using the MFPaQ software for each gel lane. Fig. 2A shows the numbers of protein groups identified for the six eluates and for the untreated CSF. There is a significant effect of the peptide library treatment as the number of unique species identified in the richest eluate (696 unique protein groups in the TUC elution from the NH₂-Library) is increased by 46% compared with the crude material (476 protein groups in non-treated CSF). Moreover, the analysis of the six fractions from the two libraries yielded a global number of 1149 unique protein groups identified in CSF after ProteoMiner treatment. More than 80% of the proteins identified in crude CSF were also found in the bead eluates where 745 new proteins could be additionally identified (Fig. 2B). Most of the 1149 protein groups identified after ProteoMiner treatment were found in both peptide ligand libraries with a major contribution of the NH₂-Library. However, the analysis of the fractions from the COOH-Library allowed detection of about 12% additional protein species (Fig. 2B). Detailed concatenated lists corresponding to proteins identified in either untreated CSF or peptide library-treated CSF as well as a global list containing all proteins identified in this study (non-treated CSF and all eluate fractions), i.e. 1212 non-redundant protein groups, are given in supplemental Data 2. These results indicate that treatment of the sample with the two peptide libraries significantly enhances the number of proteins identified by nano-LC-MS/MS. Of course, we compare the results obtained on 150 µg of non-treated CSF with those obtained on

Fig. 2. **CSF treatment on two peptide ligand libraries and nano-LC-MS/MS analysis.** A large pool of CSF samples was treated with 1 ml of beads from the NH₂-Library and COOH-Library, and both columns were sequentially eluted with TUC, UCA, and HOS buffers. A, analysis of non-treated CSF, flow-through (FT), and elution fractions by monodimensional SDS-PAGE. The number of non-redundant protein groups identified by nano-LC-MS/MS in each gel lane is indicated. B, Venn diagrams giving the extent of overlap of protein groups found by LC-MS/MS analysis between non-treated CSF and all the eluates from both libraries and between all the eluates from the NH₂-Library and all the eluates from the COOH-Library. C, number of unique protein groups identified in each 1D gel fraction over the lanes of either crude CSF (dotted line) or the first eluate from the NH₂-Library (solid line) and the albumin abundance index (calculated as the mean XIC area value for the three most intense peptides of the protein) in each 1D gel fraction over the lane of crude CSF.
150 μg of the elution fractions from the library columns, but it must be noted that a much larger amount of starting CSF material was necessary to perform the treatment. Indeed, using the peptide ligand library strategy, the more material that is loaded, the more low abundance proteins get progressively enriched, whereas major species are not. Conversely, a straightforward fractionation of the crude sample by SDS-PAGE is limited by the capacity of the 1D gel. Even a massive overloading of the 1D gel provides a relatively small increase in the number of identified proteins (23) as it mainly results in loading more of the major proteins without providing a real enrichment of minor species. On the contrary, such proteins can actually be enriched by overloading the peptide ligand beads, which results in a reduction of protein dynamic range and a better MS/MS sampling of the ions detected by nano-LC-MS/MS. This effect is particularly strong for the gel fractions containing mainly albumin in the crude CSF: as this protein is largely decreased in the corresponding gel slices of the ProteoMiner eluates, many new proteins are unmasked in these bands and largely account for the number of newly identified proteins after treatment (Fig. 2C).

To evaluate the origin and function of the proteins revealed by the treatment, we classified the global list of protein groups identified in the study according to their gene ontology terms using the GoMiner software. An example of a subclass of proteins of probable neural origin, associated with the gene ontology term “neurogenesis,” is shown in Table I. Proteins are ranked by decreasing abundance (reflected by their total number of MS/MS counts), and those identified only after ProteoMiner treatment are displayed in bold (no protein specific to the crude material was present in this category). It clearly appears that the newly identified species are mostly low abundance proteins, and many of them are involved in very specific neuronal functions such as control of axonal extension and dendrite outgrowth. This is for example the case for Semaphorin-3B, Meteorin, and proteins of the SLITRK family, which have been shown before to be overexpressed in some astrocytic brain tumors (25). These data

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>Best score</th>
<th>Peptides</th>
<th>Total MS/MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPI00021842</td>
<td>Apolipoprotein E</td>
<td>2269</td>
<td>38</td>
<td>3317</td>
</tr>
<tr>
<td>IPI00291262</td>
<td>Clusterin</td>
<td>1499</td>
<td>26</td>
<td>2366</td>
</tr>
<tr>
<td>IPI00061114</td>
<td>Pigment epithelium-derived factor</td>
<td>1470</td>
<td>25</td>
<td>880</td>
</tr>
<tr>
<td>IPI00032220</td>
<td>Angiotensinogen</td>
<td>1402</td>
<td>16</td>
<td>877</td>
</tr>
<tr>
<td>IPI00022246</td>
<td>Azurocidin</td>
<td>864</td>
<td>14</td>
<td>361</td>
</tr>
<tr>
<td>IPI00220642</td>
<td>14-3-3 protein γ</td>
<td>882</td>
<td>19</td>
<td>219</td>
</tr>
<tr>
<td>IPI00216319</td>
<td>14-3-3 protein η</td>
<td>530</td>
<td>19</td>
<td>106</td>
</tr>
<tr>
<td>IPI00333778</td>
<td>NRCAM protein</td>
<td>549</td>
<td>26</td>
<td>85</td>
</tr>
<tr>
<td>IPI00374563</td>
<td>Agrin</td>
<td>622</td>
<td>28</td>
<td>78</td>
</tr>
<tr>
<td>IPI00290959</td>
<td>Neural cell adhesion molecule L1-like protein</td>
<td>565</td>
<td>27</td>
<td>65</td>
</tr>
<tr>
<td>IPI00027166</td>
<td>Metalloproteinase inhibitor 2</td>
<td>214</td>
<td>10</td>
<td>41</td>
</tr>
<tr>
<td>IPI00024966</td>
<td>Contactin-2</td>
<td>355</td>
<td>21</td>
<td>39</td>
</tr>
<tr>
<td>IPI0013976</td>
<td>Laminin subunit β-1</td>
<td>313</td>
<td>12</td>
<td>39</td>
</tr>
<tr>
<td>IPI00442299</td>
<td>Neurexin-1α</td>
<td>200</td>
<td>10</td>
<td>33</td>
</tr>
<tr>
<td>IPI00160552</td>
<td>Tenasin-R</td>
<td>158</td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>IPI00027463</td>
<td>Protein S100-A6</td>
<td>225</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>IPI00029693</td>
<td>Neurexin-1</td>
<td>122</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>IPI00179330</td>
<td>Ubiquitin and ribosomal protein S27a</td>
<td>105</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>IPI00384225</td>
<td>Meteorin</td>
<td>245</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>IPI00387168</td>
<td>Proprotein convertase subtilisin/kexin type 9</td>
<td>290</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>IPI00012283</td>
<td>Semaphorin-3B</td>
<td>243</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>IPI00102543</td>
<td>SLIT and NTRK-like protein 1</td>
<td>57</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>IPI00165438</td>
<td>Muscle type neurelin 1</td>
<td>73</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>IPI00293939</td>
<td>Protein S100-B</td>
<td>165</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>IPI00179589</td>
<td>Myotrophin</td>
<td>80</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>IPI00333140</td>
<td>Delta and Notch-like epidermal growth factor-related receptor</td>
<td>101</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>IPI00217507</td>
<td>Neurofilament medium polypeptide</td>
<td>121</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>IPI00034588</td>
<td>Neurexin-3β</td>
<td>36</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>IPI00305975</td>
<td>Spondin-2</td>
<td>74</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>IPI00024212</td>
<td>Palmitoyl-protein thioesterase 1</td>
<td>60</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>IPI00217146</td>
<td>SLIT and NTRK-like protein 4</td>
<td>77</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>IPI00009253</td>
<td>α-Soluble NSF attachment protein</td>
<td>65</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Quantitative Analysis of CSF after ProteoMiner Treatment
indicate that the treatment of CSF with the peptide libraries is able to unmask minor components originating from brain that could constitute biologically relevant biomarkers and that its use could be interesting for clinical studies on CSF.

Assessment of Saturation of Peptide Library Beads for Highly Concentrated Proteins—As the very principle of this strategy is to modify the concentration of the proteins in the starting material, its potential use for quantitative proteomic profiling studies on CSF remained to be assessed. We have previously performed quantitative experiments using red blood cell lysate samples spiked with growing amounts of a yeast protein and treated in triplicate on peptide library beads. We found (i) a good reproducibility of the MS signal of the spiked protein between replicate experiments and (ii) a linear response of the MS signal with increasing concentrations of this protein (23). We concluded that relative quantitative comparison of a protein in different samples was still possible after the treatment as long as the protein did not saturate the beads. Indeed, high abundance proteins that tend to saturate their binding sites during the capture stage would be impossible to quantify because they asymptotically approach a certain value of saturation that does not significantly change with an additional load. For such proteins, any differential expression ratios between samples would probably not be practically exploitable after treatment. To evaluate this saturation effect and roughly assess at which order of concentration it would occur, we performed a test on a panel of proteins spiked at very high amounts in serum samples (up to 20 μM; i.e. at the level of some major serum proteins) that were afterward treated in triplicate on ProteoMiner beads (Fig. 1). Fig. 3 shows the mean of the MS signal response (XIC area) of peptides originating from these proteins as a function of the final concentration of the protein spiked in the sample. Interestingly, we obtained a constantly growing signal over a wide range of concentrations, globally linear at intermediate concentrations and flattening only between 2 and 20 μM, probably due to the progressive saturation of the protein baits on the ligand library. These data suggest that quantification of proteins after the treatment may be possible over a concentration range spanning at least 3 orders of magnitude albeit with a compression of differential ratios for highly abundant proteins.

Label-free Quantification of Proteins in CSF Samples Using MFPaQ Software—To evaluate the quantitative reproducibility of the peptide library treatment on a more global scale, a bioinformatics tool able to automatically extract the MS signal of all detected peptides was necessary. We previously described the MFPaQ software, which was designed to parse and validate protein identifications from Mascot result files and quantify the identified proteins when using isotopic labeling strategies such as ICAT, stable isotope labeling with amino acids in cell culture, or 14N/15N labeling (24, 26). The quantification module of this software starts from validated peptides lists and uses the m/z value and retention time associated to peptide ions (retrieved from Mascot result files) to extract the XIC signal of these ions in the corresponding raw files. In the case of isotopic labeling strategies, if only one member of the light/heavy peptide pair is identified by MS/MS, the software calculates the theoretical m/z value of its co-peptide and extracts its XIC signal in the same MS survey scans as the two peptides are expected to co-elute in the same run. In MFPaQ version 4.0, the quantification module was upgraded so that it could handle label-free quantification using a similar approach. Thus, it was used here to extract the XIC signal for the same m/z value in two or several LC-MS/MS runs, corresponding to the same unlabeled peptide detected in different parallel analyses.

To assess the accuracy of the bioinformatics label-free quantification and the technical reproducibility of the MS measurement, replicate analyses were performed on either crude CSF or ProteoMiner-treated CSF (first eluate from the NH2-Library), and identified proteins were quantified using this tool (Fig. 1). For these assays, the samples were not fractionated, and the whole protein mixture was digested and analyzed in one LC-MS/MS run. As shown in Fig. 4A, the reproducibility of the nano-LC-MS runs on the Orbitrap mass spectrometer was found to be fairly good between replicate injections of either crude or treated CSF. The peptides XIC area values showed a coefficient of variation for triplicate nano-LC-MS runs typically around 5–10%. The vast majority of the protein population was correctly quantified by the soft-
ware with a protein ratio showing little deviation to the expected value of 1 between replicate injections and few outlier values due to incorrect signal extraction (Fig. 4B). As expected, the number of proteins identified and quantified in treated CSF was higher than in crude CSF (238 versus 140; Fig. 4D). These numbers indicate that the treatment with the combinatorial ligand library is even more beneficial when the sample is analyzed in only one LC-MS/MS run without fractionation: indeed, in that case, the huge prevalence of albumin in the global mixture hampers the identification of all others proteins in crude CSF, whereas this effect is much reduced in the treated sample, leading to a final increase of 70% of the number of identified proteins.

Contrary to pattern-based strategies in which LC-MS features have to be defined from the analysis of peptide elution and isotopic profiles in LC-MS maps, the approach used in MFPaQ, based on extracted ion chromatograms of identified peptides, is driven by experimentally measured RT and by monoisotopic m/z values validated from MS/MS sequencing. Thus, it allows performing peak detection in a quick and accurate way. A drawback of this method, however, is that only peptides that have been identified by MS/MS in at least one of the LC-MS/MS runs can be quantified. Although the peptide library treatment allows an increase of the number of identified proteins in CSF, this number remains relatively small in a one-run analysis because of MS/MS undersampling of the highly complex peptide mixture. To circumvent this problem, we implemented a strategy based on an identification database containing sequences of previously identified peptides along with their m/z and retention time associated values. After fractionation of treated CSF on a 1D SDS gel and LC-MS/MS analysis of 20 gel slices, a database containing 5482 peptide ions, matching to 720 protein groups, was built. Extraction of XICs for all the peptide ions of the database was then performed in replicate analytical runs of the unfractonated sample (either crude or treated CSF). Because of the limited dynamic range of the instrument, not all of the peptides from this database could be retrieved in the individual

---

**Fig. 4. Label-free quantification of CSF proteins using MFPaQ.** A, reproducibility and accuracy of peptide quantification over triplicate LC-MS/MS analyses of crude CSF sample tryptic digests: correlation of XIC area values in pairwise LC-MS run comparisons (left panels) and distribution of peptide ions according to their coefficient of variation (CV) for the three replicate measurements (right panel). B, C, and D, quantification of proteins over two replicate LC-MS/MS runs of either crude or ProteoMiner-treated CSF. The distribution of protein groups was plotted according to their relative abundance ratio calculated by MFPaQ either by direct comparison of the two LC-MS/MS runs (B) or using a previously created identification database (C). The total numbers of correctly quantified protein groups are displayed for both samples using the two methods (D).
runs. However, using such an approach, the MS/MS under-sampling problem could be largely overcome, and the number of proteins correctly quantified in replicate runs of the individual samples significantly increased (Fig. 4C). Best results were obtained with treated CSF in which up to 445 proteins could be quantified in one run (Fig. 4D). Thus, the bioinformatics work flow developed in the MFPaQ software provided an efficient quantitative solution for profiling several hundreds of proteins in treated CSF samples.

Efficiency and Reproducibility of ProteoMiner Treatment on Low CSF Volumes—The experiment described above, designed to provide an extensive and qualitative characterization of the CSF proteome, was performed by treating a large volume of biological fluid on 1-ml peptide bead columns containing a very large panel of the two peptide ligand libraries. Moreover, differential elutions were performed as well as extensive fractionation of each eluate on a 1D gel to reach a better analytical coverage of the sample. Clearly, such an experimental scheme cannot be applied for routine clinical studies. A lumbar puncture typically yields only 1–2 ml of CSF, and because of the very low protein concentration in this fluid, this corresponds to less than 1 mg of material. With such amounts, the bead volume must be decreased to only a few microliters to keep a reasonable overloading ratio of starting material versus bead capacity (about 10 mg/ml). Moreover, if the aim of the study is to profile in a reproducible way large numbers of CSF patient samples, fractionation will be very difficult to implement. Thus, we miniaturized the treatment protocol to be compatible with such a clinical application model. A pooled CSF sample was prepared from several lumbar punctures (CSF Pool 2), and 2-ml aliquots were treated in replicate experiments by incubation on either 2- or 5-μl batch volumes of NH₂-Library beads (Fig. 1). Elution was performed by boiling the beads in Laemmli buffer (28), samples were loaded on a gel for 1D SDS-PAGE, and the whole protein mixture was concentrated in one gel band. After tryptic digestion, the resulting peptides were analyzed in one LC-MS/MS run. With such low volumes of beads, both the efficiency and the reproducibility of the treatment had to be assessed. Fig. 5 shows the statistics for sequencing events and identification numbers in LC-MS/MS analytical runs corresponding to either untreated CSF or replicate treated CSF using a small volume of peptide library beads. Although the number of MS/MS queries for all samples is fairly similar, many more peptides and proteins are confidently identified in treated CSF than in initial untreated CSF. Slightly better results are obtained with a 5-μl than with a 2-μl volume of beads: the final number of protein groups identified increases by ~120 and 100% with these two respective treatments. Thus, this miniaturized protocol is still efficient and significantly improves the number of proteins identified in single run LC-MS/MS analysis of CSF samples. To check the reproducibility of the protocol, label-free quantification of the proteins identified in four replicate experiments was performed using MFPaQ. To evaluate the abundance level of each protein in the samples, we plotted a PAI calculated as the average of signal intensity for the three most abundant tryptic peptides of each protein. In the four LC-MS/MS analyses of the replicate ProteoMiner experiments, the global number of unique protein groups identified and quantified by the software was 283 (5-μl treatment) and 250 (2-μl treatment) when the four runs were compared directly (without database). Although not all of these proteins were identified by MS/MS in every LC-MS/MS run, MS signal could be extracted for all of them in the four replicates for both conditions. As shown in Fig. 6A, the PAI profiles for the 283 proteins identified after 5-μl bead treatment were relatively constant over the four replicates. The median of the coefficients of variation for all peptide intensity values over four replicates was 9.2% (5-μl treatment) and 11.7% (2-μl treatment) (Fig. 6B), showing a low variability of the experimental protocol. To increase the number of quantified proteins, we applied the bioinformatics work flow based on an identification database. This was done by nano-LC-MS/MS shotgun analysis of one sample from CSF Pool 2 treated on 5 μl of beads fractionated by SDS-PAGE into 20 gel bands, which allowed building a database containing 5015 peptide ions.
matching to 568 protein groups. In this way, 530 proteins could be quantified with reproducible PAI values over the four replicates for the 5-μl treatment, and 491 proteins could be quantified for the 2-μl treatment (Fig. 6B and supplemental Data 3). Coefficients of variation for the quantified peptides on this population of proteins were only slightly higher (10.4 and 12.4%, respectively, for the 5- and 2-μl treatments), indicating that the quantification remained fairly accurate even on lower abundance proteins. Together, these data indicate that the miniaturized protocol is efficient and reproducible and that several hundred proteins can be accurately quantified in 2-h-long analytical runs by using this treatment on low CSF volumes.

DISCUSSION

In this study, we tried to evaluate the benefits of combinatorial peptide libraries for proteomic characterization of CSF and their potential use for clinical profiling studies by nano-LC-MS. In the first place, an experiment was performed on a large pool of CSF samples with two columns (NH2- and COOH-Library) each containing 1 ml of beads and sequentially eluted with different buffers; each eluate was then analyzed by 1D gel fractionation and nano-LC-MS/MS. The aim of this experimental design was to optimize the number of proteins detected in CSF by (i) the use of the maximal diversity of hexapeptide ligands to favor the binding of many different species and (ii) the extensive fractionation of the proteins recovered from the columns to improve the analytical coverage of the sample. By using this work flow, we could confidently identify 1149 unique protein groups in the elution fractions from the libraries versus 476 protein groups in non-treated CSF. Clearly, there was a positive effect of peptide library beads for the reduction of the dynamic range of the protein mixture by strongly cutting down the high abundance
proteins while enriching the low abundance species. This effect was particularly visible in the gel bands containing albumin in which many more proteins were identified in the treated sample than in the starting material. Indeed, as this protein alone represents about 45% of the total protein amount, it is largely responsible of the masking of low abundance species in CSF as in other fluids such as serum or urine. Moreover, the increased number of proteins detectable in the treated sample was noticeable in almost all regions of the 1D gel migration lane, indicating that the enrichment of minor species was effective on the global protein population. This was also clearly visible from the 2D maps of the untreated CSF and the eluates from the libraries. As reported previously for other fluids (20, 22, 23), a few proteins detected in the starting material were not found in the treated sample (66 proteins in the case of CSF), corresponding to proteins that may not find a bait ligand to form a stable complex. Despite this drawback, the benefit of using the ligand libraries was quite substantial with 745 newly identified proteins.

Together, this study allowed the identification of 1212 unique protein groups in CSF, which represents an in-depth characterization of this fluid by nano-LC-MS/MS. Other data sets of CSF proteins have been published in the past. For instance, using 1D gel fractionation and LC-MS/MS analysis on high resolution mass spectrometers (Orbitrap or FTICR), Zougman et al. (13) reported the identification of 798 protein groups by analyzing individual or pooled CSF samples. In a previous report, Pan et al. (11) gathered all the data of several CSF studies performed in their group using a variety of protein fractionation techniques (1D SDS-PAGE, ACN precipitation, glycoprotein affinity chromatography, and ICAT labeling), peptide separation methodologies (strong cation exchange and reverse-phase HPLC), and mass spectrometric platforms (ion trap, FTICR, and MALDI-TOF/TOF) (12, 14, 29–31) and merged the results of all these analytical strategies to yield a list of 2594 protein accession numbers. When clustered at 98% homology and compared using the Protein Center software (Proxeon), the three protein lists (this study, 1148 protein groups; Pan et al. (13), 761 protein groups; Pan et al. (11), 2093 protein groups) showed a certain extent of overlap with roughly around 500 protein groups shared between two lists in pairwise comparisons and a core set of about 400 protein groups found in common between all three data sets (supplemental Data 4). However, each list showed a relatively high number of specific proteins not found in the two other lists. In the case of the present study, 564 protein groups not reported in the previous publications were identified, most of them found thanks to the sample treatment with peptide libraries. Many reasons may account for the significant difference in protein identifications between these investigations, related either to the analytical techniques used or to the CSF sample itself. Indeed, because of the depth and diversity of the CSF proteome, none of the lists are probably exhaustive, and different fractionation methods or sample treatments may give access to separate categories of proteins. In this respect, the application of selective affinity purification methods (e.g. glycoprotein capture or ICAT labeling) versus more general enrichment strategies such as combinatorial peptide ligand libraries likely resulted in the detection of different low abundance species. On the other hand, the molecular composition of CSF is highly dynamic and may vary significantly due to individual phenotype fluctuations, age of the patients, sample collection (lumbar puncture versus ventricular withdrawal), volume of CSF extracted, pathologies, and CSF bulk flow rate. In particular, the level of blood-derived proteins in CSF is highly dependent on blood-CSF barrier function and dysfunction. Moving from the concept of a morphological blood-CSF “leakage” model, this barrier function is now more widely interpreted using a model connecting molecular flux of blood proteins with CSF flow rate (4, 32). Basically, the CSF flow rate is considered as the main factor modulating protein concentration in the fluid, and a reduced flow rate is for example sufficient to explain the non-linear increase of blood-derived CSF protein concentrations measured in neurological diseases (33). The variation of CSF flow rate with age and pathologies of the individuals will thus largely influence the relative concentrations of blood- or brain-derived proteins in the sample analyzed and may largely account for the heterogeneity of CSF proteomics reports so far.

In an attempt to determine the portion of the CSF proteome originating from blood, Zougman et al. (13) compared the list of proteins identified in their study with the Human Proteome Organisation plasma proteome (data set of 889 high confidence plasma proteins) and found that the overlap was quite small as only 24% of the proteins identified in their CSF study were found in the plasma list. When we performed the same kind of comparison with our data, we found a slightly higher overlap when the list of proteins identified in crude CSF was used (39% of the proteins found in the plasma list), which may well reflect the different nature of the samples analyzed (large pool of CSF from disease patients in our study versus five diagnostically normal individuals). However, when the comparison was performed with the more extensive list obtained from analysis of ProteoMiner-treated CSF, only 19% of the identified proteins were shared with the plasma data set. Similarly, 18% of the CSF data set from Pan et al. (11) overlapped with the plasma list. Thus, as discussed previously, it appears from these numbers that a very large portion of the CSF proteome is constituted by intrinsic species that may not be derived from blood but rather from brain. On the other hand, it is usually described that only about 20% of the protein content of CSF is predominantly brain-derived (1, 6, 34). However, this percentage relates to total protein amounts, whereas the different qualitative proteomics studies evaluate the diversity of the CSF composition only in terms of protein numbers. To get a more precise description of the CSF proteome, we quantified the proteins identified in our study by extracting MS intensity signal for each identified
peptide and calculated for each protein a PAI defined as the average MS signal response for its three most intense tryptic peptides (35). When summed together, the PAI of the 175 proteins shared between the crude CSF list and the plasma list represents around 75% of the total PAI sum, a value close to what is generally described.

Despite the high abundance of blood-derived proteins in CSF, the peptide library technology used here was found to be an efficient approach to unravel low abundance brain species that could not be detected before sample treatment. Interestingly, in CSF from patients without any neurological disease, it permits the detection of proteins related to different cell compartments inside the brain and to different crucial metabolic pathways, sometimes known to be overexpressed in particular diseases. For example, among the proteins involved in neurogenesis displayed in Table I, many were identified only after CSF treatment. This is the case for Semaphorin3B and neuropilin-1 and -2, which play a role in axonal guidance and cell migration (36–38). Detection of these proteins may be of particular interest when studying tumoral brain pathologies such as glioma as they have been proposed to be involved in local invasion and migration of tumor cells, which are pivotal mechanisms in glioma progression (39, 40). Similarly, the 14-3-3 η isoform and the SLIT and NTRK-like protein 1, identified only in treated CSF, have also been shown to be expressed in astrocytoma and glioma cells (25, 41). Other classical brain markers were detected after peptide library treatment, such as neuron-specific enolase (γ-enolase) and the astrocytic protein S100B, which are released in the CSF following brain cell degradation and are often used to detect pathologies involving brain damage such as Creutzfeldt-Jakob disease (42) or to predict the severity of central nervous system injury in cases of cerebral hypoxia, brain infarction, or trauma (43–45). The detection of other minor astrocytic proteins such as the phosphoprotein PEA-15 (49) are potential indicators of the neurochemical activity of neuronal cells. Neurotrophin receptors were also detected in treated CSF, such as TrkB (receptor of the brain-derived neurotrophic factor), which plays a role in cognition, learning, and memory formation by modulating synapse plasticity and neuronal survival and thus represents a critical molecule in neurodegenerative diseases such as Alzheimer disease (50, 51). Similarly, the detection of the PARK7/DJ-1 protein, one of the Parkinson disease-associated proteins (52), as well as different actors of the proteasome and protease families is interesting in the perspective of biomarker investigation in the context of neurodegenerative diseases (53). Finally, several enzymes involved in major metabolisms were detected, such as the oxidative mitochondrial metabolism (pyruvate kinase isozymes M1/M2, isocitrate dehydrogenase, dihydrolipoyl dehydrogenase, acetyl-CoA acetyltransferase, and malate dehydrogenase), which is important for the exploration of several neurological diseases (54).

Although the peptide ligand libraries appear to be a powerful tool to qualitatively map the CSF proteome, its use for clinical studies was yet to be evaluated. The quantitative aspects and reproducibility of the method were important features to be checked to validate its use for differential proteomics studies. The present work shows quantitative data obtained by measuring, after nano-LC-MS analysis, the signal intensities of peptides derived from growing amounts of exogenous proteins spiked in serum samples subsequently treated with ProteoMiner beads. These spike tests were performed over a wide range of concentrations, up to high final values, to evaluate the saturation effect and check whether relative quantification was still possible even for high or medium abundance proteins. We observed for all proteins a linear MS response up to the μM range, indicating that relative quantification of proteins should be accurate up to such concentration. This is in agreement with previous observations (23) and with a recent study in which quantitative aspects of peptide library treatment were evaluated by 2D electrophoresis (55). Indeed, protein baits captured by the ligand library behave under the law of mass action where relative concentration of species and their affinity constant for a given peptide play the most important role. Binding of a protein to a given ligand takes place proportionally to the concentration of the protein, but as a limited number of molecules of this peptide ligand are available at the surface of the bead, the protein tends to saturate the bead at high concentrations. This saturation effect was visible at μM concentrations for most of the proteins measured. However, no clear plateau was reached probably because complex mechanisms involving multiple affinity equilibria take place at the surface of the beads: one protein does not interact only with one peptide ligand but with several ligands with different affinity constants, and competition between protein species for a given ligand takes place proportionally to their respective concentrations. Thus, a protein may tend to saturate with higher loads, but a growing MS signal is still measurable, indicating that differential expression of even relatively highly abundant proteins could potentially be measured but with an underestimation of the real ratio.

In addition to these spike experiments, reproducibility of the technology was checked on the global population of proteins detected by nano-LC-MS analysis using the label-free quantification module of the MFPaQ software. Reproducibility assays were performed with a miniaturized protocol applicable to real clinical CSF samples. Indeed, due to the high dilution of CSF proteins and to the low volumes of fluid usually available after lumbar puncture, the total protein amount contained in such samples is quite low, which necessitates performing the treatment on very low volumes of beads. The use of small
Volumes of peptide ligand beads raise an important concern, which is to assess the effect of a large undersampling of the peptide library on both the efficiency and the reproducibility of the method. Indeed, the library is a pool of beads, each bearing a unique, specific ligand, and because of the diameter of the beads, the total volume of resin necessary to have a statistical representation of almost all the hexapeptide ligands has been estimated to be around a few milliliters. Thus, when very small volumes of beads are taken out of the bulk, the final number of peptide ligands is much smaller than the total number of library diversomers, and in addition, the population of ligands sampled out for each experiment is variable. We show here that this had no major effect on the efficiency of the treatment. When small volume CSF samples were incubated with very low volumes of beads, the number of proteins detected by nano-LC-MS/MS was increased by more than 100% when the analysis was performed in one run, a result even better when one considers the initial experiment using 1 ml of beads. Moreover, label-free quantification on the global population of peptides and proteins identified in replicate treatment experiments showed that, despite the sampling of the library, all protein species were captured with comparable efficiency in all replicate experiments. Coefficients of variation calculated on MS intensity values were centered around 9–12%, a value corresponding roughly to the technical variability of the nano-LC-MS measurement. This can be explained by the fact that the statistical number of peptidomers covering the interaction needs is much reduced compared with combinatorial calculations. Basically, in terms of functionality, peptides that differ from each other because one glycine is replaced by an alanine or because isoleucine is replaced by valine have probably very similar capturing properties for a given protein. In addition, it was shown that sequences of three amino acids seemed already enough for the capture of most proteins from a crude extract (56). Actually, although with hexapeptides obtained with for example 15 amino acids the number of diversomers is 11.4 millions, with tripeptides the number of diversomers is reduced to 3375. Translated into a volume of beads, around 2 μl would be sufficient to cover 90% of tripeptides corresponding to the distal part of the ligands. This approach supports the reproducibility of the data obtained on CSF using 2 and 5 μl of beads. Naturally, the larger the number of beads, the better the coverage of the library, which seems also to be consistent with the fact that a little better reproducibility is obtained with 5 μl of beads compared with 2 μl of beads.

Ideally, a proteomics analytical method aiming at biomarker discovery shouldallow(i)thedetectionofalargenumberofproteins,(ii)theirquantificationwithagoodaccuracy, and (iii) the analysis of a large number of patient samples to provide statistically significant results. However, a compromise has to be found between these requirements. For example, although extensive 1D gel fractionation of a sample represents a useful way to identify a high number of species by MS/MS, it may not be an ideal solution to quantitatively profile large series of samples in a reproducible way. Conversely, although SELDI-TOF is a well-suited tool to perform this latter task, it may only give access to the profiling of a limited number of low mass proteins. NanoLC-MS profiling of enzymatically digested proteins using label-free quantification on high resolution mass spectrometers may provide alternative solutions. These strategies have already been used to identify new biomarkers in CSF with good results (15, 27). In such approaches, very limited or no fractionation at all is usually performed on the sample because of the large number of analyses that must be performed (number of patients and technical replicates) and the intrinsic variability that may be introduced by the fractionation procedure itself. In such an experimental scheme where the protein mixture is analyzed in one run, the treatment of CSF samples on a peptide ligand library was shown to be particularly useful to lower the dynamic range of the protein mixture and increase the number of detected and quantified proteins without fractionating the sample. Moreover, the MFPaQ software (SourceForge) was shown to be an efficient tool to accurately quantify proteins in CSF. With the use of an identification database, previously built from the analysis of 20 fractions of the sample separated by 1D SDS-PAGE, around 500 protein species could be mapped and successfully quantified in replicate nano-LC-MS/MS analyses of CSF aliquots treated on low volumes of beads. Clearly, the analytical coverage of the sample was not as large as in the first large scale, extensive proteomics analysis, which yielded a final list of 1212 proteins. However, it was obtained with an analytical time reduced by a factor 7, and the analysis was reproducible and potentially applicable to many more replicates. Moreover, the list of proteins detected using this strategy still contains many biologically relevant species derived from brain cells (e.g. the brain injury markers S100B and neuronal-specific enolase, glial fibrillary acidic protein, pentraxin-1 and -2, neuropilin-1 and -2, PARK7/DJ-1, SLITRK1, etc.), providing potentially interesting candidates in the perspective of biomarker discovery studies.

In conclusion, the analytical work flow presented here using the miniaturized ProteoMiner protocol and the bioinformatics data processing method developed in MFPaQ allowed profiling of the CSF proteome with a reasonable depth (about 500 proteins quantified) in short analytical times (less than 2 h per sample) and with good accuracy (coefficients of variation typically around 10% for replicate measurements). It could thus represent a useful approach for future clinical studies on CSF.

---

REFERENCES


