Transition from identity to bioactivity-guided proteomics for biomarker discovery with focus on the PF2D platform

Antoine Ménoret¹, Stephen J. Crocker², Annabelle Rodriguez³, Vijay A. Rathinam¹, Robert B. Clark¹ and Anthony T. Vella¹

¹ Department of Immunology, University of Connecticut Health Center, Farmington, CT, USA
² Department of Neuroscience, University of Connecticut Health Center, Farmington, CT, USA
³ Department of Cell Biology, University of Connecticut Health Center, Farmington, CT, USA

Proteomic strategies provide a valuable tool kit to identify proteins involved in diseases. With recent progress in MS technology, high throughput proteomics has accelerated protein identification for potential biomarkers. Numerous biomarker candidates have been identified in several diseases, and many are common among pathologies. An overall strategy that could complement and strengthen the search for biomarkers is combining protein identity with biological outcomes. This review describes an emerging framework of bridging bioactivity to protein identity, exploring the possibility that some biomarkers will have a mechanistic role in the disease process. A review of pulmonary, cardiovascular, and CNS biomarkers will be discussed to demonstrate the utility of combining bioactivity with identification as a means to not only find meaningful biomarkers, but also to uncover functional mediators of disease.

Keywords: Acute respiratory distress syndrome / Asthma / Atherosclerosis / Bioassay / Inflammation / Multiple sclerosis / Proteoform / PTMs / Validation assay

1 Introduction

Scrutinizing relevant biological fluids and tissues to detect biomarkers pertinent to disease has become increasingly important for clinical research. New relevant biomarkers could be used to detect the risk of acquiring a disease, the presence or the stage of pathology, and responses to therapy. Genomic and transcriptional profiling studies have shown strong potential for biomarker discovery in a broad array of diseases while proteomic strategies are still emerging [1]. Considering all possible gene expression patterns, mutations, coding, and noncoding SNPs, RNA splicing, microRNA biology, and the immensity of protein PTMs, it was estimated in human that up to 1.8 million different protein forms, recently coined “proteoforms,” can be considered [2, 3]. A proteoform designates all of the different molecular forms in which the protein product of a single gene can be found, including changes due to genetic variations, alternatively spliced RNA transcripts and PTMs [2]. This enormous diversity, the dynamic turnover of proteins and their release in various biological fluids during pathology is the basis for identification of protein biomarkers. Proteomics of various diseases, especially using approaches that can identify proteoform diversity, offer new biomarker avenues.

Given the number of organ systems and their associated pathologies that can be explored using proteomics approaches, it becomes increasingly difficult to address them all. Proteomic analysis of brain, lung, and cardiovascular systems has shown that some of the same markers of disease are often identified across different tissues. For example, an abundant protein such as actin has been identified as a marker for hypersensitivity pneumonitis, multiple sclerosis, and aortic aneurism [4–11]. Likewise, Cystatin C is simultaneously a putative biomarker for malignant pleural effusion, traumatic brain injury, and cardiac injury [12–14]. Other common protein biomarkers in various diseases, listed in Table 1, suggest that acute-phase proteins as well as proteins released...
Table 1. Biomarker proteins common to cardiac (C), neurological (N), and pulmonary (P) diseases

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Condition</th>
<th>Tissue</th>
<th>Proteomic technology</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actin</td>
<td>C</td>
<td>Artherosclerotic aorta</td>
<td>Aortic tissues</td>
<td>LC-MS/MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Aortic aneurysm</td>
<td>Aortic wall</td>
<td>2D-DIGE and MS/MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Multiple sclerosis</td>
<td>CSF</td>
<td>2D gel and LC/MS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiple sclerosis</td>
<td>CSF</td>
<td>2D-DIGE, MALDI-TOF MS and UPLC/Q-TOF MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Hypersensitivity pneumonia</td>
<td>BALF</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td>α-1 Antichymotrypsin</td>
<td>C</td>
<td>Valvular heart disease</td>
<td>Plasma</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Multiple sclerosis</td>
<td>CSF</td>
<td>2D-DIGE, MALDI-TOF MS and UPLC/Q-TOF MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>EAE</td>
<td>CSF</td>
<td>2D gel and LC/MS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiple sclerosis</td>
<td>CSF</td>
<td>2D gel and LC/MS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAE</td>
<td>CSF</td>
<td>ChipLC-QTOF MS, Orbitrap XL</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>NSCLC</td>
<td>Plasma</td>
<td>Monoclonal antibody libraries</td>
</tr>
<tr>
<td>β-2 Microglobulin</td>
<td>C</td>
<td>PAD</td>
<td>Plasma</td>
<td>SELDI-TOF MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Multiple sclerosis</td>
<td>CSF</td>
<td>2D gel and MALDI MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>CBD</td>
<td>BALF</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td>Calreticulin</td>
<td>C</td>
<td>AAA</td>
<td>Aortic tissue</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Parkinson’s disease</td>
<td>Striatum</td>
<td>2D gel and MALDI TOF-MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Lung cancer</td>
<td>BALF</td>
<td>2D gel and MALDI TOF-MS</td>
</tr>
<tr>
<td>Ceruloplasmin</td>
<td>C</td>
<td>Cardiopulmonary bypass</td>
<td>Plasma</td>
<td>2-DIGE and MALDI-TOF/TOF</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>HAD</td>
<td>Serum</td>
<td>2D gel and IB</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Lung injury</td>
<td>BALF</td>
<td>1D gel and MS</td>
</tr>
<tr>
<td>Clusterin</td>
<td>C</td>
<td>Alzheimer disease</td>
<td>Plasma</td>
<td>2D gel and LC-MS/MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Multiple sclerosis</td>
<td>Plasma</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Multiple sclerosis</td>
<td>CSF</td>
<td>MALDI TOF-MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Lung cancer</td>
<td>BALF</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td>Complement C3</td>
<td>C</td>
<td>Plasma clots</td>
<td>Plasma</td>
<td>2G gel and MALDI-TOF-MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>EAE</td>
<td>CSF</td>
<td>QTOF LC-MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>NSCLC/RILT</td>
<td>Spinal cord</td>
<td>iTRAQ MS/MS</td>
</tr>
<tr>
<td>Cystatin C</td>
<td>C</td>
<td>Cardiac injury</td>
<td>Conditioned media of cardiomyocytes</td>
<td>LC-MS/MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Traumatic brain injury</td>
<td>CSF</td>
<td>DIGE and MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Malignant pleural effusion</td>
<td>Adenocarcinoma cell lines</td>
<td>1D gel and LC-MS/MS</td>
</tr>
<tr>
<td>α-Enolase</td>
<td>C</td>
<td>CHF</td>
<td>Cardiomyocytes</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>EAE</td>
<td>Spinal cord</td>
<td>2DE and MALDI-TOF MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Multiple sclerosis</td>
<td>CSF</td>
<td>2D-DIGE, MALDI-TOF MS and UPLC/Q-TOF MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Multiple sclerosis</td>
<td>Blood cells</td>
<td>2D gel, IB, ELISA and MS</td>
</tr>
<tr>
<td>Haptoglobin</td>
<td>C</td>
<td>Myocardial infarction</td>
<td>Plasma</td>
<td>2D-DiGE and MALDI-TOF/TOF</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Ischemic stroke</td>
<td>Plasma</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Idiopathic pulmonary fibrosis</td>
<td>BALF</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td>Hemopexin</td>
<td>C</td>
<td>Cardiopulmonary bypass</td>
<td>Plasma</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Multiple sclerosis</td>
<td>CSF</td>
<td>2D gel and MALDI-MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Multiple sclerosis</td>
<td>Spinal cord</td>
<td>MALDI-MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Chemical lung exposure</td>
<td>Plasma</td>
<td>2D gel and MALDI-TOF</td>
</tr>
</tbody>
</table>

(Continued)
Table 1. Continued

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Condition</th>
<th>Tissue</th>
<th>Proteomic technology</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malate</td>
<td>C</td>
<td>Ischemia-reperfusion</td>
<td>Hearts mitochondria</td>
<td>2DE and MALDI-TOF MS</td>
</tr>
<tr>
<td>Dehydrogenase</td>
<td>N</td>
<td>EAE</td>
<td>Spinal cord</td>
<td>2D gel and MALDI-TOF MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Multiple sclerosis</td>
<td>CSF</td>
<td>LC-ESI-MS/MS</td>
</tr>
<tr>
<td>MMP-9</td>
<td>C</td>
<td>ALI</td>
<td>Lung tissue</td>
<td>2DE and MALDI-TOF MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>MACE</td>
<td>Plasma</td>
<td>ELISA</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Brain stroke</td>
<td>Infarcted brains</td>
<td>Protein array</td>
</tr>
<tr>
<td>Paraoxonase</td>
<td>C</td>
<td>ASCVD</td>
<td>Plasma</td>
<td>iTRAQ 2D LC-MS/MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>EAE</td>
<td>CSF</td>
<td>LC-MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>SCLC</td>
<td>Serum</td>
<td>Affinity purification and iTRAQ</td>
</tr>
<tr>
<td>T-kininogen</td>
<td>C</td>
<td>AMI</td>
<td>Serum</td>
<td>2DE and MALDI-TOF/TOF MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>EAE</td>
<td>CSF</td>
<td>LC-MS and QTOF LC-MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Endotoxin challenge</td>
<td>BALF</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td>Thrombospondin 1</td>
<td>C</td>
<td>Myocardial infarction.</td>
<td>Left ventricle infarct tissue</td>
<td>2D gel and MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Ischemic stroke</td>
<td>Plasma</td>
<td>1D gel and MS/MS</td>
</tr>
<tr>
<td>Vimentin</td>
<td>C</td>
<td>Malignant mesothelioma</td>
<td>Aortic valve</td>
<td>2D-DIGE and MS</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>EAE</td>
<td>Spinal cord</td>
<td>iTRAQ, MALDI-MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>PM/DM</td>
<td>Blood/CNS</td>
<td>1D, 2D gel IB, and MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Malignant mesothelioma</td>
<td>BALF</td>
<td>2D gel and LC-MS/MS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>Malignant mesothelioma</td>
<td>Lung biopsies</td>
<td>2D gel and MS</td>
</tr>
</tbody>
</table>

**Cardiac diseases.** AAA, abdominal aortic aneurysms; AMI, acute myocardial infarction; ASCVD, atherosclerotic cardiovascular disease; CHF, congestive heart failure; MACE, major adverse cardiovascular event; PAD, peripheral arterial disease; RHD, rheumatic heart disease.

**Neurological diseases.** EAE, experimental autoimmune encephalomyelitis; HAD, HIV-1 associated dementia.

**Pulmonary diseases.** ALI, acute lung injury; IPF, idiopathic pulmonary fibrosis; PM/DM, polymyositis/dermatomyositis; RILT, radiation-induced lung toxicity.

**Tissues:** BALF, bronchoalveolar lavage fluid; BCEC, brain capillary endothelial cells; CBD, chronic beryllium disease; CSF, cerebrospinal fluid; NSCLC, nonsmall cell lung cancer; SCLC, small cell lung cancer; TIF, tissue interstitial fluid.

**Proteomic technologies:** 2D-DIGE, 2-dimensional fluorescence difference in-gel electrophoresis; iTRAQ, isobaric tag for relative and absolute quantification; LC/MS, liquid chromatography tandem mass spectrometry; MALDI-TOF, matrix-assisted laser desorption/ionization coupled to time of flight; MS, mass spectrometry, MS/MS, tandem mass spectrometry; SELDI, surface-enhanced laser desorption/ionization; UPLC-MSE, nano-ultra performance liquid chromatography coupled to electrospray ionisation mass spectrometry; UPLC/Q-TOF MS, ultra-performance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry.

during cellular injury are frequently detected and not necessarily specific to a disease. Interestingly, the same biomarker can be found in various biological fluids across different diseases. For example, clusterin has been detected in bronchoalveolar lavage fluid (BALF), plasma, and cerebrospinal fluid (CSF) in patients with lung cancer, multiple sclerosis, and Alzheimer disease [5, 15–17] (Table 1). Similarly, hemopexin is present in plasma, CSF, and the spinal cord of patients with lung injury, multiple sclerosis, and cardiopulmonary bypass [15, 18–21] (Table 1). Such markers, with often excellent sensitivity but poor specificity, are not necessarily the best candidates for disease detection and diagnosis. The issue of specificity will continue to be a problem unless the proteomic approach is combined with other diagnostics tests, including disease-related bioassays.

The ultimate goal of clinical proteomics is to find meaningful biomarkers associated, as specifically as possible, with the pathology of a particular tissue, preferably in a quantifiable manner. Clinically, the ideal biomarker should be easily obtained with minimum discomfort or risk to the patient, hence be present in a clinically accessible peripheral body tissue and/or fluid (e.g., blood, semen, saliva, urine, hair, or breath). Alternatively, biomarkers found in a tissue that cannot be sampled for routine examination may still be useful if they are detectable by noninvasive imaging. A reliable biomarker should be detected with high sensitivity and high reproducibility. To achieve this goal, rare proteins, expressed, or modified in pathological situations must first be identified. The search for novel protein biomarkers in various diseases such as multiple sclerosis, asthma, acute lung injury, and atherosclerosis are listed in this review. The broad utility of these systems will be addressed as well as the importance of validation using bioactivity-guided proteomics versus proteomics strategies that solely focus on identification. The PF2D platform will be used as an example of 2D-LC
coupled to MS analysis to illustrate the need of validation and bridging bioactivity to protein identity.

2 Proteomics in clinical and basic research

2.1 Proteomics of a CNS disease, the case of multiple sclerosis

Multiple sclerosis is a chronic inflammatory disease of the CNS characterized by infiltrating auto-reactive immune cells that mediate demyelination and axonal degeneration of neurons [22]. Currently, its diagnosis is almost exclusively established by clinical examination and magnetic resonance imaging. Clinical biomarkers would help in differential diagnosis identifying disease subtypes (i.e. relapsing remitting, secondary progressive, primary progressive, progressive relapsing) with the possibility of earlier diagnosis, monitoring treatment effectiveness, and identifying new treatment targets. Human proteomic studies have used postmortem brain samples, CSF, and blood as a source to screen proteins [8,9,15,19,23,24]. Since the tissue damage in multiple sclerosis is localized to the CNS, proteomics of postmortem brain are particularly relevant in revealing putative biomarkers and mechanisms linked to advanced pathology, although they do not qualify for longitudinal studies. Collecting CSF during disease progression is a more practical approach that can be applied to longitudinal studies [25]. However, the invasive procedure necessary to collect CSF is a major drawback for routine screening. Proteins from blood samples are easier to collect, however, the complexity of the blood proteome, the masking effect of highly abundant blood proteins, such as albumin, and the low concentration of CNS proteins in the blood, has not generated enthusiasm for using blood proteomics in the long run [26]. Hence, several methods have been developed to deplete albumin and other abundant serum proteins to gain sensitivity [27–29].

The experimental autoimmune encephalomyelitis (EAE) murine model stimulates T cells to trigger demyelination of the CNS in a way that simulates multiple sclerosis [30]. Proteomic analysis of human and animal tissues have revealed a large number of putative markers for multiple sclerosis and EAE, with some common to both [26,31,32]. The differences in induction and pathogenesis between EAE and multiple sclerosis argue for caution in interpreting proteomics changes in EAE alone. One example representing successful workflow from proteomic identification to in vivo model validation, is that of activated protein C (aPC) [33]. Global profiling of brain lesions from multiple sclerosis cases identified aPC within chronic active plaque. Administered in vivo, aPC ameliorated EAE through its anti-coagulant function and suppression of NFkB signaling. Among these common markers, those with altered PTM are particularly interesting. For example, occludin and neurofilament light-heavy chain have been found as phosphorylated proteins in both EAE [34–36] and multiple sclerosis [37–40]. Myelin basic protein and glial fibrillary acidic protein can be citrullinated during EAE [34,41] and detected in human white matter of multiple sclerosis patients [42,43]. Mitochondrial hsp70 and glycerolaldehyde 3-phosphate dehydrogenase can be nitrated in EAE [44] and found in chronic inactive multiple sclerosis lesions [45,46]. These proteins characterized by the presence of PTM on abundant proteins could be specific markers of disease. Such biomarkers can theoretically be followed by imaging of the brain by using specific ligand or antibody in a similar way as recently proposed for prostate cancer [47] and Alzheimer's disease [48]. Proteomics-based studies have found many candidates for multiple sclerosis and have contributed to confirming the immunological and neurodegenerative process of the disease, yet there is no validated and consistent biomarker available for clinical use.

2.2 Proteomics of the cardiovascular system, the case of atherosclerosis

Atherosclerosis is one of the leading causes of death worldwide characterized by cholesterol deposition and chronic inflammation leading to progressive sclerosis of the arterial vessels that promote the pathophysiology of cardiovascular disease, stroke, and peripheral vascular disease [49]. The complex mechanism underlying plaque formation and disease progression involves matrix degradation, angiogenesis, oxidative stress lipid deposition and metabolism, apoptosis, and autophagy [50,51]. Advanced atherosclerotic disease can be diagnosed by clinical examination (e.g. vascular bruits) and by various cardiovascular imagining technologies [52]. Improving our knowledge about the molecular events involved at the onset of atherosclerosis and finding biomarkers of vascular atheroma would have considerable utility. Classical proteins commonly used in the clinic have been considered as biomarkers, but none of them have been implemented in the clinic due to their inconsistency [53,54]. More recent proteomics studies that used data mining techniques for biomarkers in plasma, serum, urine, atheroma, and secretome have revealed numerous potential candidates, but few have been submitted to functional validation [55]. The current proteomics approaches for atherosclerosis have produced a vast amount of experimental data [55–57]. Among them phospholipid-associated proteins: sod2, sod3, gst, hsp20, annexin A10, and fibrinogen fragment D are new candidate biomarkers. Many other proteins have been identified as putative biomarkers, but have not necessarily been confirmed by other investigators or fully validated [58–61]. On the contrary, several independent proteomics studies have shown that hsp27 plays an active role in atherosclerosis and appears to be a biomarker of myocardial ischemia [62–65]. Interestingly, 2DE and Dige showed that three isoforms of phosphorylation-dependent hsp27 are modulated during atherosclerosis [62,63]. The level of hsp27 circulating in
the blood of patients with cardiovascular disease has been measured in patients and healthy volunteers, but its role as a biomarker has not always been consistent, perhaps because the different proteoforms of hsp27 have not been considered [65–68]. Moreover, serum levels of anti-hsp27 antibodies are significantly associated with acute coronary syndrome and stroke [69,70]. Therefore, investigating the PTM of hsp27 might clarify its role as a biomarker for atherosclerosis. Soluble tumor necrosis factor-like weak inducer of apoptosis (sTWEAK) is another example of a potential candidate biomarker for atherosclerosis identified by proteomic analysis [71]. Isolation of an 18 kDa intact protein released in greater amounts by arteries of healthy patients than those with atherosclerotic plaques, was identified as sTWEAK [72]. ELISA and Western blot of secretome (culture of conditioned media from normal or pathological endarteries) validated that plaques release less than normal arteries [72]. As a functional validation, it was shown that mice treated with a TWEAK blocking monoclonal antibody and mice with genetic deletion of TWEAK are less prone to develop atherosclerosis [73]. In independent clinical studies, it was observed that circulating sTWEAK decreased in patients with coronary artery disease [74], chronic heart failure [75], and abdominal aortic aneurysms [76]. Additionally, sTWEAK has been shown to predict atherosclerosis in renal transplant patients [77]. Therefore, sTWEAK appears to match some of the requirements of a clinical biomarker, while also possessing a bioactivity role in disease.

### 2.3 Proteomics of the lung, the case of asthma, and acute respiratory distress syndrome

Asthma is a disease of the airways characterized by reversible airway obstruction, airway hyperresponsiveness, and chronic airway inflammation [78]. Acute respiratory distress syndrome (ARDS), and its experimental form, acute lung injury (ALI), are characterized by a widespread pulmonary inflammation with alveolar damage resulting in respiratory insufficiency [79]. Proteomics of asthma and ARDS/ALI have been used to pinpoint the pathophysiology of disease and to identify biomarkers. For both diseases, proteomes were analyzed using samples from BALF, sputum, lung tissue, plasma, and serum, and the findings have been reviewed in detail elsewhere [78–82]. Among all the putative biomarkers of asthma, S100A9 also called calgranulin b, is a potential candidate. Calgranulin b has been found in several proteomic studies to be upregulated in sputum of asthmatic patients [83,84]. It was also upregulated in the sputum of patients with uncontrolled asthma analyzed by 2DE and MALDI-TOF MS, and was validated by Western blot and ELISA [85]. Interestingly, proteomics has identified four forms of calgranulin b with alternate start methionine codons and phosphorylation [86]. Moreover, calgranulin b is constitutively expressed in neutrophils [87], involved in lung inflammation as a damage-associated molecular pattern (DAMP) that binds receptor of advanced glycation end-products (RAGE) and Toll-like receptor-4 [88], and has also been shown to be oxidized, nitrosylated, and glutathionylated [89,90]. Since calgranulin b is a DAMP, a bioassay could be designed to follow its bioactivity and validation of a specific proteoform of calgranulin b could be isolated using proteomics, not only as a specific biomarker, but perhaps also as a specific inflammatory agent in asthma.

Biomarkers for ARDS/ALI can be separated by factors induced during inflammation and those induced by cell injury [81]. Compared to other pathologies, a smaller number of proteomic analysis have been performed. Insulin-like growth factor has been identified in BALF of ARDS patients. Validation by Western blot and ELISA showed GFPB3 was over-expressed in BALF of ARDS patients compared to normal volunteers [91], but the opposite result was recently obtained using a much larger patient cohort [92]. Heparin-binding epidermal growth factor-like growth factor has also been identified in BALF of ARDS patients [91], and is a potential candidate because of its protective role in ALI [93]. Nevertheless, further studies are needed for its validation. Peroxiredoxin 1, cytochrome b5, and cytokeratin 17 have been identified and validated in animal models of ALI, but have not yet been investigated in human clinical studies [94,95]. Proteomic studies of pulmonary edema fluid, plasma, and BALF from ARDS/ALI patients have detected many acute phase proteins and inflammatory proteins [96,97]. Since none of the current biomarkers possess sufficient specificity and sensitivity, a new approach has highlighted a combination of biomarkers that has shown to be promising [98–100]. Specifically, due to the variety of causes of ARDS/ALI, a panel of biomarkers has a greater chance to pinpoint diagnosis compared to a single biomarker. Moreover, association of physiological markers can add accuracy to the diagnosis. For example, inflammatory cytokines like IL-6 and IL-8 are reliable predictors of oxygenation, ventilator-free days and 28-day mortality [99]. In cases of community-acquired pneumonia, levels of circulating soluble CD14 combined with a pathological scoring system enhanced the predictability of ARDS and mortality [101]. Also, there is excellent evidence suggesting metalloproteinases can fulfill a role in marking ARDS/ALI severity, especially MMP-1 and MMP-3 levels which correlated with increased mortality, disease severity, and multigorgan failure [102]. All the pathologies listed above, for different reasons, occur coincident with an inflammatory response. This facet could be exploited for biomarker mining. Multiple sclerosis is an inflammation-mediated disease characterized by an inflammatory environment in the CNS, activation of CD8 and CD4 T cells and the release of free radicals, proteolytic enzymes, and a variety of inflammatory mediators [103–105]. During ARDS/ALI, neutrophil infiltration and alveolar macrophage activation initiate the inflammatory response by producing high levels of proinflammatory cytokines, iNOS, free radicals, damage signals, and aberrant proteolysis [106–108]. Asthma is also characterized by inflammation, but classically considered a T-helper type 2 disease.
of the airways with eosinophil infiltration [109]. The inflammatory status in atherosclerosis is quite established as well, with continuous infiltration of leukocytes into the plaque, high levels of reactive oxygen species, and activation of metalloproteinases MMP2 and 9 [110, 111]. All these different types of inflammation lead to a large variety of factors able to modify proteins and creating additional proteoforms. As discussed below, the development of proteomics methods offers began to propose solutions for identifying these newly created proteoforms.

3 Evolution of proteomic approaches

3.1 Past and present proteomics

Proteomics in which intact proteins were first isolated and then identified, were the first to be used in proteomics research. One of the most popular methods was the 2DE separation of proteins, that is still used effectively today [112]. Proteins of interest, separated according to their pl and then mass, were excised from the gel and initially sequenced by Edman degradation [113]. More recently, trypsin-digest of excised spots are identified by mass spec and exact mass ions compared with databases to pinpoint protein identification. To eliminate gel-to-gel variability and allow higher sensitivity and more accurate quantification, DIGE was developed. In this newer version of 2D gel proteomics, proteins are labeled with fluorescent dyes before simultaneous separation in the same run allowing better sample comparison. Despite this advancement many limitations remain including the difficulty in detecting rare proteins, the masking effect of abundant proteins, the poor resolution of highly hydrophobic proteins, the limitation of pl range, and the lack of method automation. Recent advances in MS technologies, reviewed in greater detail elsewhere [55, 114], have bypassed some of the limitations of the gel-based systems and have advanced biomarker research. These technologies are called bottom-up proteomics or shotgun proteomics, which starts with peptide breakdown by enzymatic digestion of the entire proteome sample. Classical shotgun proteomics includes trypsin digestion of a complex mixture of proteins into a more complex mixture of peptides, followed by fractionation, fragmentation, sequencing of the peptides by MS, and identification of the proteins of origin via in silico digest protein databases. Shotgun proteomics have become broadly used and have provided immense quantities of protein biomarker potential candidates. This approach is unmatched for sensitivity and precise detection of known target peptides in a sample. However, there are certain drawbacks. Trypsin can miss proteins that are not properly denatured, proteolyze proteins into peptides too small for LC-mass spec detection, or skip a seemingly cleavable position [115]. This decreases the depth of the proteome to analyze and reduces the accuracy of the comparison with the theoretical in silico digests contained in databases used to identify the protein [116]. Also, shotgun proteomics can miss PTM that can lead to inaccurate protein quantification [117]. This limitation is unfortunate as protein PTM are frequently associated with a disease [118–121]. Furthermore, proteolytic digestion increases the complexity of the original protein sample, and prolongs the time of analysis, making high-throughput screening difficult. The vast quantity of peptides derived from the most abundant proteins, especially in the serum, limits the loading capacity of the LC-mass spec/mass spec columns and masks the presence of rare proteins. A large number of peptides increases considerably the chance of finding a significant number of false-positive matches when searched against a large database [122]. Another challenge common to many proteomics strategies, is the often underrepresentation of membrane proteins due to their low abundance and poor solubility [123]. Most importantly, and central to the concept developed in this review, trypsin digestion typically destroys protein biological function impeding validation in bioactivity assay.

One important aspect of proteomic strategies rely on accurate quantitative approaches based on stable isotope labeling, label-free statistical assessment methods and absolute quantification approaches [124]. Stable isotope labeling (e.g. iTRAQ, metabolic labeling, isotope-coded affinity tag reagents) uses stable isotopes or mass tag variants to label two different samples [125]. The proteins of a sample are tagged with a high molecular weight variant and the other one with low molecular weight variant or no tag. The samples are mixed, fractioned, and analyzed by MS. The peaks in the MS analysis indicate the relative ratio of the mass tag variants that indicate the relative abundance of the protein or peptide. Label-free statistical assessment or data-independent acquisition methods provide a sensitive approach for large-scale quantitative proteomic analyses [126]. With these methods, a sample is searched for the presence and quantity of a limited set of peptides that are specified prior to data acquisition. These kinds of targeted proteomics can precisely quantify specific proteins in many samples and are well suited for biomarker research. Finally, absolute quantification approaches, called AQUA strategy, or stable isotope dilution-MRM-MS permit absolute quantification of a given protein by spiking the sample with a peptide of interest synthesized using amino acids containing stable heavy isotopes, creating a slight increase of mass [127]. After protein digestion, the native and synthetic peptides are resolved together during the chromatography and ionized identically. By knowing the quantity of the labeled peptide, the precise abundance of the original peptide can be calculated. This approach is directed to the quantification of known biomarkers rather than the discovery of new ones [128].

3.2 Proteomics approach for intact proteins

New proteomics during which proteins are first fractionated, and sometimes totally purified and identification by MS, provides complementary approaches to shotgun proteomics that
have been reviewed in detail [112, 129, 130]. Processing of intact proteins into the mass spectrometer permits advanced characterization of PTM and bypasses the difficulties associated with peptide-based proteomics [112]. In these strategies, separation of intact proteins based on mass, charge, and/or hydrophobicity increases the separation of proteoforms. Samples containing simple protein mixtures or fully purified protein can allow for 100% sequence coverage and full characterization. The most popular ways to separate intact proteins rely on classical LC methods like ion exchange, IEF, or reverse phase. More recently, capillary electrophoresis has allowed separation of intact proteins and peptides in small capillaries under high voltage, using low sample and reagent consumption, and with high efficiency [131]. Some of these approaches, called top-down proteomics, although less employed than bottom-up proteomics, have shown their usefulness in investigating disease biomarkers [132–139]. They have been described as a complementary cost-effective approach in clinical research [140], and recently promoted by a new consortium for the comprehensive analysis of intact proteins [www.topdownproteomics.org].

### 3.3 PF 2D fractionation platform

Among the proteomics platforms with integrated proteome analysis capability, the PF2D offers a way to fractionate any kind of proteome by 2D LC. Quantitative and qualitative proteomic maps of fractions containing proteins are easily generated, which can be used to identify proteins by MS sequencing. Like other platforms [141], the PF2D isolates intact proteoforms that can be tested using bioassays before their identification by MS. Similar to other proteomic strategies, some investigators have added additional dimensions before or after 2D LC fractionation to increase separation power. Immunoaffinity depletion of abundant proteins in plasma samples has been performed before PF2D separation [142], and 1D SDS-PAGE analysis of the PF2D reverse-phase fractions has been used as a way to select protein bands before MS [142–144].

#### 3.3.1 Sample preparation

To achieve maximum proteome depth, a sample must include proteins in a buffer that allow their separation while keeping them in solution. Sample preparation for PF2D proteomics uses solubilization buffers containing 6 M urea and a nonionic detergent detergent, n-octyl-β-D-glucopyranoside, that can be largely removed by processing samples through the reverse phase chromatography. High urea concentration with detergent has the advantage of being chemically stable and denaturing, hence inhibiting most enzyme activities including proteases. Membrane proteins, known to be difficult to analyze, have been successfully identified on PF2D platforms, suggesting this technology could be used for routine analysis of membrane proteins [145, 146], and for large extracellular matrix proteins that are usually not resolved by 2DE [147]. The protease inhibitors contained in the buffer and high urea concentration prevent protein degradation and loss of PTM. Samples as diverse as eukaryote and prokaryote cell lysate, serum, plasma, mitochondria, extracellular matrix, urine, and BALF have been successfully analyzed by PF2D proteomics [95, 144, 146–154]. Conveniently, the sample can be prepared with or without reducing agents, allowing for isolation of proteins linked by disulfide bonds. This allows the study of proteins which structure and enzymatic activity is modulated by thioredoxin and other disulfide modifying proteins [155].

#### 3.3.2 First dimension IEF

Proteins are separated in a first dimension by IEF. Most investigators have used buffers in the range of pH 4 to 8.5 that cover the pl of most proteins [144–147, 150, 152, 154, 156–161], but a few studies have tried additional pH ranges. For example, proteins from the same samples have been separated using two different pH ranges (4.0–8.5 and 7.0–10.0) [154], and basic proteins from macrophages infected with Candida albicans have been successfully identified using a pH range from 8.3 to 11.3 [162]. The elution of proteins from the IEF column is performed by a decreasing pH gradient, fractions are collected in a 96-deep well plate refrigerated at 12°C as they are automatically reinjected for a second fractionation. The first dimension collects fractions, not at a constant interval of time, but every 0.3 unit of pH. This aids in comparison of pH fractionation from run to run, and increases reproducibility. In fact, the reproducibility of the PF2D platform has actually been evaluated by running 10 consecutive aliquots of urine from the same patient [153], and improved protocols extend the life of the column to maximize cost-effectiveness [150, 163].

#### 3.3.3 Second dimension, reverse phase chromatography

After the first dimension, fractions are reinjected through an automated autoloader onto a C-18 column heated at 50°C for reverse phase chromatography. Bound material is eluted through a concentration gradient of acetonitrile and peaks are detected at 214 nm. Typically, 12–20 first dimension fractions, selected from the linear part of first dimension gradient are fractionated by reverse phase chromatography and separated by hydrophobicity. Thus each first dimension run is fractionated into 36 fractions, generating a total of 432–720 fractions. Two-dimensional proteomic maps displaying pI versus hydrophobicity are constructed using the Proteoview/Deltaview software built in the PF2D platform.
3.3.4 Analysis of proteome maps

Analysis usually reveals a large majority of identical chromatographic patterns that we refer to as signatures and differential ones as fingerprints [143]. For example, Raf kinase inhibitor protein (RKIP), which was similarly expressed in resting and restimulated primed lymphocytes, was detected in a fingerprint suggesting the presence of an alternative proteoform [143]. Shotgun proteomics strategies that start with a trypsin digestion of the proteome are not well suited to find proteins that are physiologically cleaved, or proteins slightly modified and would have likely missed the differential presence of RKIP. After analysis of the proteomic maps, proteins contained in the selected fingerprint fractions can be directly identified by MS or subjected to further analysis for either purification refinement (e.g. SDS-PAGE), and validation using biological assays that will confirm the relevance of the protein.

3.3.5 Limitations

One difficulty in using the PF2D is its great sensitivity to pH variation. The first IEF dimension relies on precise fractionation every 0.3 unit of pH. Another related difficulty is the 16–18 h required to run a sample through the 2D LC. This can affect the reproducibility when a large cohort of samples are analyzed because the pH of the buffers need to be maintained at a precise value for several days. Moreover, considering the large number of fractions produced at each run there is a need for substantial –80°C storage area. A possible area of improvement is the loading capacity of the first dimension. Currently, no more than 4 mg of total protein can be analyzed per run without lose in resolution. Loading a larger quantity of protein would facilitate the detection of rare proteins.

3.4 Proteomics require validation assays

Validation has been recently described as one of the key bottlenecks contributing to the high failure rate of candidate biomarkers [164, 165]. First, validation should identify protein biomarker in the sample of interest and not in the control, or at least to lower levels in the control. This task is often performed using ELISA, Western blotting, protein microarrays as well as MS. The second goal is significantly linking the identified protein to a biological, pathological, or clinical outcome. Without stringent validation, overly simplistic selection of biomarkers results in a large number of potential candidates that fail after initial promise. One possibility to improve the success rate in biomarker research would be to develop new strategies driven more by bioactivity of the protein than by its identity. A general view of the difference and complementarity between identity-guided proteomics and bioactivity-guided proteomics is presented in Fig. 1. Identity-guided proteomics focuses first on efficient protein extraction and processing through MS-coupled system for identification, whereas bioactivity-guided proteomics is driven by the biological activity of protein(s). Consequently, validation of the protein(s) of interest occurs at the final stage of identity-guided proteomics but is built in bioactivity-guided proteomics. Hence, sequential and thorough validation performed during bioactivity-guided proteomics could improve the success rate in identifying new biomarkers.

Since PF2D fractionates intact proteins, it can also integrate different kinds of validation strategies and begin to test the feasibility of bioactivity-guided proteomics. As mentioned earlier, RKIP was targeted because of the PF2D data but was also validated using locostatin, a specific small molecule RKIP-inhibitor [143]. Cytokine inhibition was observed in locostatin-treated mouse and human T cells, and in vivo treatment with locostatin inhibited TNF-α production upon triggering antigen specific T cells. A third line of validation was obtained using a model of T cell induced systemic inflammatory response syndrome (SIRS) where RKIP deletion inhibited production of IFN-γ due to intrinsic altered T-cell receptor activity [166]. Thus, the bioactivity validation of RKIP qualifies this protein as a potential marker and druggable protein for the treatment of SIRS.

Another example of validation linked to PF2D proteomics is the identification of T-cell specific human tumor antigens [167]. Specifically, tumor lysate was fractionated by chromatofocusing, then 20 fractions were tested by ELISPOT in a cross-presentation assay using autologous T cells, and seven reactive fractions were selected for the second dimension. Subfractions displaying protein peaks were tested using the same assay and proteins were identified by MS. Tumor expression of the proteins was first validated by both RT-PCR and immunofluorescence. Further validation was obtained for Transthyretin and Calgranulin that were recognized as autologous tumor antigens by T cells from patients with brain tumors. By integrating 2D LC proteomics with an assay that tracks biological activity of proteins through the purification scheme, two new tumor antigens were identified. A similar strategy could theoretically be applied to identify antigens in autoimmune or asthma-related illnesses not only for T cells but also B cells.

Similar in principle to the T-cell screening of tumor antigens present in the PF2D fractions, antibodies have been used in technologies called PF2D-SERPA (serological proteome analysis) and immuno-PF2D-MS/MS [144, 168]. PF2D-SERPA can test the efficiency of a novel vaccine by screening sera reactivity of vaccinated individuals. Proteins from complex targets, like inactivated pathogens or opportunistic micro-organisms, separated by 2D LC and incubated with the sera revealed the specific antigens. The resulting liquid fractions can be validated further by Western blot for subsequent peptide sequencing by MS/MS. Such analyses can uncover the most immunogenic antigens expressed by the pathogen and be tested in vaccine development [169, 170]. This method can theoretically detect proteoforms (isoforms and PTM) that are valuable variations of an antigen for vaccine
A. Identity-guided proteomics

Figure 1. From identity to bioactivity-guided proteomics. (A) Identity-guided proteomics focuses on the most efficient and reliable way to identify proteins and peptides from a complex sample. Priority is given to efficient protein extraction, sometimes after several enrichment steps, and processing through a MS-coupled system for identification. Validation of the protein(s) of interest occurs at the final stage of the entire strategy. Identity-guided proteomics can theoretically identify and quantify intact or fragmented proteins from any kind of sample, but often suffers from low success rate in identifying new biomarkers due to incomplete validation. (B) Bioactivity-guided proteomics focuses on the most biologically relevant protein(s) from a complex sample. Emphasis is given to sequential validation using bioassay(s), before processing through MS-coupled system and identification. It requires intact proteins or at least biologically active ones and the development of reliable and sensitive bioassay(s). Additional validation of the protein of interest can be performed at the end of the strategy. The sequential validation performed during bioactivity-guided proteomics should improve the success rate of biomarkers usable in the clinic.

Finally, an example of identification and validation of factors associated with ALI was obtained using a combination of proteomics and in situ analysis. In a model of ALI induced by Staphylococcal Enterotoxin A inhalation, mouse BALF was separated by PF2D. Cytochrome b5 and Cytokeratin 17 were identified as putative biomarkers of ALI [95]. Validation was performed by immunohistochemistry in lung tissue sections that showed staining in the epithelial cells of the bronchioles. Although Cytochrome b5 and Cytokeratin 17 are potentially valuable biomarkers for ALI, their presence did not explain the exacerbation of lung inflammation. Nevertheless, it is likely that these biomarkers are an indicator of cell death since they are not typically secreted like cytokines or acute phase proteins. Thus, there are clear advantages to automated 2D LC fractionation but validation is still the centerpiece of proteomics.

4 Concluding remarks and perspectives

One reason proteomics of intact proteins is used less than bottom-up approaches is due to the slow development of high-throughput technology. Recent improvements have changed that perception with large-scale experiments where over 5000 proteoforms were observed [172, 173]. In general, proteomics have also recently benefited from advances in instrumentation and LC separation [172, 174, 175]. The PF2D proteomics presented in this review could also be improved with slip-flow technology. This new chromatographic technique, which uses capillaries packed with 0.47 μm silica colloidal crystal particles, have shown higher efficiency in reversed-phase separation of intact proteins [176, 177]. This could replace
the second dimension of the PF2D, currently a C-18 reverse phase chromatography step, and provide a higher resolution for the final protein fractionation. Proteomics have already demonstrated, in a diverse range of diseases, that analysis of proteoforms of intact proteins are useful and relevant for translational studies [132, 143, 178, 179]. One area particularly well suited for this approach is bioassays, where this might advance biomarker discovery in diseases associated with acute inflammation. The pathways activated during inflammation induce phosphorylation of signaling proteins, active proteases, and protein ubiquitination, nitration, sumoylation, and numerous other protein modifications. The functional significance of these new proteoforms can be tested by immunological assays and technologies currently used to study the immune system. Single-cell mass cytometry, also known as cytometry by TOF or CyTOF, is a recent example of merging flow cytometry, a technology classically used to study immune cells, with MS [180]. CyTOF uses antibodies tagged to transition metals, which potentially allow up to 100 independent measurements on a single cell. This technology can be applied to the screening of a patient’s blood sample to track, for example, a specific phospho-protein biomarker in a rare cell population [181]. Multiplexed ion beam imaging is a related technology for advanced immunohistochemistry studies [182, 183]. Antibodies tagged with transition metals have been used to stain tissue sections following routine immunohistochemistry protocols. The samples were then ablated spot by spot in a laser ablation chamber and processed through a CyTOF mass cytometer. This approach avoids sample autofluorescence, does not require an amplification step and has been used to simultaneously image 32 proteins and protein modifications at subcellular resolution [183].

Another emerging area that complements proteomics, is detection and bioactivity of enzymes in complex proteomes using activity-based probes (ABPs) [184, 185]. Click chemistry can engineer these synthetic ABPs with both a reactive and a reporting group [186]. The reactive group is designed to covalently bind a specific class of enzyme while the detection group enables its analysis and this strategy has been used for activity-based protein profiling (ABPP) [187]. ABPP has been successful in the detection and analysis of more than 80% of mammalian Serine Hydrolases [188]. ABPP has also been applied to the screening of anti-inflammatory small molecules using an in vitro macrophage bioassay [189]. In a viral infection model, Caspase, Hydrolase, and Tyrosine Phosphatase enzyme activities have been analyzed at the protein isoform level by using proteomics followed by ABPP [190]. Besides the detection of specific enzymatic activity, ABPP has also enhanced identification and characterization of unknown protein functions [191] and the targets of natural products [192].

In conclusion, there are at least two methods to enhance proteomic biomarker discovery. The first being classical validation using relevant biological samples and the second using bioactivity assays (Fig. 1). The latter is the newest challenge, but advances in automated protein fractionation and cell culture strategies should drive the field past protein identifi-

cation alone and could contribute further in understanding disease mechanism.

The authors have declared no conflict of interest.

5 References


[75] Chorianopoulos, E., Rosenberg, M., Zugck, C., Wolf, J. et al., Decreased soluble TWEAK levels predict an adverse


[138] Desiderio, C., D’Angelo, L., Rossetti, D. V., laronove, F. et al., Cerebrospinal fluid top-down proteomics evidenced the potential biomarker role of LVV- and VV-hemorphin-7 in...


