Clinical Application of Proteomics in Breast Cancer: State of the Art and Perspectives

Anthony Goncalves a–c   François Bertucci b–d

a Centre de Recherche en Cancérologie de Marseille, Département de Pharmacologie Moléculaire, Institut Paoli-Calmettes et UMR599 Inserm, b Département d’Oncologie Médicale, Institut Paoli-Calmettes, c Faculté de Médecine, Université de la Méditerranée, et d Département d’Oncologie Moléculaire, Institut Paoli-Calmettes, Marseille, France

Introduction

Completion of the human genome project and technological progress are revolutionizing translational medicine. Today, high-throughput molecular technologies (‘omics’) permit investigators to systematically interrogate the genome, transcriptome, and proteome of cancer cells. They provide real molecular portraits of biological samples, offering an unprecedented opportunity to discover new diagnostic, prognostic or therapeutic targets. Since the mid nineties of the last century, large-scale RNA expression profiling has been the most developed approach. In particular, DNA microarrays have been applied to clinical breast cancer samples [1], and have rapidly confirmed their potential clinical significance through the identification of biologically and clinically relevant molecular subtypes [luminal A and B, basal, human epidermal growth factor receptor 2 (HER2)-overexpressing, and normal-like] and of new prognostic subclasses, unidentifiable by conventional means [2]. Today, two prognostic gene expression signatures (Oncotype DX® and Mammaprint®) derived from these studies are being tested in prospective phase III trials [3, 4], the ultimate phase before their diffusion into routine clinical practice [5].
In this ‘omics context’, proteomics represents one of the latest technological developments. Similar to studies aimed at deciphering transcriptional changes important in cancer, proteomic approaches allow global and comparative studies of proteins to be performed on normal and pathological samples. Proteomics has the potential to complement the information generated by genomics for several reasons. mRNA levels do not necessarily correlate with the abundance of the corresponding protein [6–8]; additional complexity of proteins is conferred by post-translational modifications such as phosphorylation, acetylation, and glycosylation, or protein cleavage, which are not detectable at the mRNA level, but play significant roles in protein functions [9]; since proteins bridge the gap between RNA and cellular physiology, their potential as clinically relevant biomarkers is theoretically superior; finally, proteins represent more accessible and relevant therapeutic targets than nucleic acids. Many recent technical advances in mass spectrometry (MS) and microarrays have enabled high-throughput proteome analysis.

In this review, we describe the proteomics-based methods that have been applied to date to clinical breast cancer samples for diagnostic and prognostic purposes. These studies were performed on tumor tissues and biological fluids. The most important results as well as the current limitations and perspectives are summarized and discussed. For reasons of space, we will not address the issues of tissue microarrays and immunohistochemistry, which are exhaustively described elsewhere [10].

The proteomic technologies (table 1) may be roughly distinguished according to the use or not of MS. Typically, non-MS-based approaches correspond to microarray-based techniques, and require a priori hypotheses or selection of proteins to be tested, whereas MS-based techniques do not require any previous biological foreknowledge, and allow examination and quantification of a large number of initially unknown protein parameters. However, some MS-based approaches may also be hypothesis based (e.g. immunoprecipitation MS), and an alternative distinction can separate methods as candidate-based or non-candidate-based methods. Candidate-based methods require antibodies or other affinity reagents against specific candidates, whereas non-candidate-based methods are typical screening approaches.

### Table 1. Proteomic tools applied to biomarker research in mammmary oncology

<table>
<thead>
<tr>
<th>Non-MS-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tissue microarrays and immunochemistry</td>
</tr>
<tr>
<td>Protein arrays</td>
</tr>
<tr>
<td>Antibody microarrays</td>
</tr>
<tr>
<td>Reverse-phase protein arrays</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MS-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein separation tools</td>
</tr>
<tr>
<td>2D electrophoresis</td>
</tr>
<tr>
<td>Liquid chromatography</td>
</tr>
<tr>
<td>Protein-profiling tools</td>
</tr>
<tr>
<td>ProteinChips/SELDI-TOF MS (BioRad)</td>
</tr>
<tr>
<td>ClinProt/MALDI-TOF MS (Brucker)</td>
</tr>
<tr>
<td>Protein identification tools</td>
</tr>
<tr>
<td>MALDI-TOF MS</td>
</tr>
<tr>
<td>ESI-TOF MS</td>
</tr>
<tr>
<td>Q-TOF MS/MS</td>
</tr>
<tr>
<td>Ion trap MS/MS</td>
</tr>
<tr>
<td>Fourier transform MS</td>
</tr>
<tr>
<td>Protein quantification</td>
</tr>
<tr>
<td>SILAC</td>
</tr>
<tr>
<td>ICAT</td>
</tr>
<tr>
<td>iTRAQ</td>
</tr>
</tbody>
</table>

MS = Mass spectrometry; SELDI-TOF = surface-enhanced laser desorption ionization-time of flight; MALDI-TOF = matrix-assisted laser desorption ionization-time of flight; ESI-TOF = electrospray ionization-time of flight; Q-TOF = quadruple time of flight; MS/MS = tandem mass spectrometry; SILAC = stable isotope labeling with amino acids in cell culture; ICAT = isotope-coded affinity tag; iTRAQ = isotope tagging for relative and absolute quantification.

In this 'omics context', proteomics represents one of the latest technological developments. Similar to studies aimed at deciphering transcriptional changes important in cancer, proteomic approaches allow global and comparative studies of proteins to be performed on normal and pathological samples. Proteomics has the potential to complement the information generated by genomics for several reasons. mRNA levels do not necessarily correlate with the abundance of the corresponding protein [6–8]; additional complexity of proteins is conferred by post-translational modifications such as phosphorylation, acetylation, and glycosylation, or protein cleavage, which are not detectable at the mRNA level, but play significant roles in protein functions [9]; since proteins bridge the gap between RNA and cellular physiology, their potential as clinically relevant biomarkers is theoretically superior; finally, proteins represent more accessible and relevant therapeutic targets than nucleic acids. Many recent technical advances in mass spectrometry (MS) and microarrays have enabled high-throughput proteome analysis.

In this review, we describe the proteomics-based methods that have been applied to date to clinical breast cancer samples for diagnostic and prognostic purposes. These studies were performed on tumor tissues and biological fluids. The most important results as well as the current limitations and perspectives are summarized and discussed. For reasons of space, we will not address the issues of tissue microarrays and immunohistochemistry, which are exhaustively described elsewhere [10].

The proteomic technologies (table 1) may be roughly distinguished according to the use or not of MS. Typically, non-MS-based approaches correspond to microarray-based techniques, and require a priori hypotheses or selection of proteins to be tested, whereas MS-based techniques do not require any previous biological foreknowledge, and allow examination and quantification of a large number of initially unknown protein parameters. However, some MS-based approaches may also be hypothesis based (e.g. immunoprecipitation MS), and an alternative distinction can separate methods as candidate-based or non-candidate-based methods. Candidate-based methods require antibodies or other affinity reagents against specific candidates, whereas non-candidate-based methods are typical screening approaches.

### Protein Arrays

As protein probes are arrayed on solid supports and their specific interactions with proteins present in complex biological samples tested, protein arrays are theoretic protein counterparts of DNA microarrays [11]. In antibody microarrays, one of the most developed formats, spotted specific antibodies are confronted with samples along with various labeling and detection systems. Technological limitations include the availability of antibodies, the space available on the array, and the fact that the recommended experimental conditions may not be optimal for all arrayed antibodies, which have specific affinities for their ligand. Biological fluids such as serum or plasma as well as tumor tissue itself can be analyzed. For example, the expression of 378 proteins was measured in matched malignant and adjacent normal breast tissues using antibody microarrays [12]. The authors found increased expression levels of various proteins, such as casein kinase 1e, p53, annexin XI, CDC25C, eIF-4E and MAP kinase 7, in malignant breast tissues,
whereas the expression of other proteins, such as the multifunctional regulator 14–3–3e, was decreased. Importantly, the results were confirmed by immunohistochemistry, indicating the reliability of the process.

Antibody microarrays may be specifically designed to evaluate particular biological processes, such as cytokine expression. This was used in cell line models, but also in serum or tumor tissue from breast cancer patients [13–16]. For example, cytokine expression levels were screened in culture medium from breast cancer cell lines [15], identifying interleukin 8 as a factor related to estrogen receptor status and metastatic potential. A similar approach applied to MCF7 cells overexpressing HER2 or not generated a ‘HER2-induced cytokine’ signature including CXCL8 (interleukin 8), CXCL1 and growth-related onco-genes. Interestingly, the latter were also found to be up-regulated in sera from HER2-positive metastatic breast cancer patients [16]. Recently, a large-scale (129 analytes) recombinant scFv cytokine-based antibody microarray was used in an attempt to classify patients with metastatic breast cancers versus healthy controls based on differential protein expression profiling of whole serum samples. Breast cancers could be classified with 85% specificity and sensitivity. By adopting a condensed signature composed of the top 9 differentially expressed proteins, the authors [17] distinguished cancer from healthy serum proteomes with 95% sensitivity and specificity.

Instead of planar-based multiplexing, bead-based multiplexing may be used. In Luminex systems (Luminex Corp., Austin, Tex., USA), different fluorescent and spectrally resolvable beads coated with distinct antibodies to various cytokines, such as interleukins, interferons, granulocyte colony-stimulating factor and tumor necrosis factor, are incubated with a biological sample. The mixture is then incubated with a mixture of detection antibodies, each corresponding to one of the capture antibodies. The detection antibodies are tagged for detection of fluorescence. The beads are passed through a flow cytometer, and each bead is detected by two lasers: one for the measurement of fluorescence and one for detecting particular epitopes, including phosphorylated regions. Subsequently, labeled secondary antibodies are applied and analyzed with a detection/amplification system [22]. Thus, several samples and the activation state of various members of a given intracellular pathway may be examined in a single experiment [23]. However, in contrast to tissue microarrays, it is difficult, using RPPA, to characterize the subcellular location of proteins whose expression is being measured even though subcellular fractionation may be performed prior to binding arrays. Boyd et al. [24] used RPPA to examine the phosphorylation status of 100 proteins in 30 breast cancer cell lines that had previously been characterized at the transcriptional level. The technology was demonstrated to be analytically reproducible (Pearson correlation, r = 0.94, between independently derived lysates from the same cell
line) and specific pathway activation events were identified in the different molecular subtypes of breast cancer. For example, basal-like cells had low levels of pPTEN and high levels of total EGFR, Pyk2 (Y402) and PKC-α (S567); luminal-like cells had higher levels of phosphorylation of pp70S6K (S244) and pα-RAF (S259), and HER2-amplified cells had high levels of pERBB3, pFAK and pEGFR (Y1173). Importantly, the pretreatment levels of specific phosphorylated proteins correlated with in vitro response to potent and selective PI3K and MEK inhibitors. For example, a correlation existed between sensitivity to MEK inhibitors and total EGFR and phosphorylated ERK1/2 (the direct targets of MEK phosphorylation) expression, suggesting that constitutive signaling via the EGFR/RAS/MEK/ERK axis may constitute a positive predictive factor for MEK inhibition. Conversely, phosphorylation of components of Akt and mTOR signaling such as PDK1 and p70S6K was negatively associated with response. In contrast, sensitivity to the PI3K/mTOR inhibitor PI-103 was correlated with elevated phosphorylation at key nodes in the PI3K/Akt/mTOR pathway, including pAkt (T308 and S473), PRAS40 (T246) and FKHR (T24), suggesting that high levels of signaling via the pathway may be indicative of pathway addiction and predictive of response to a targeted PI3K inhibitor. RPPA has also been applied to breast cancers whose PIK3CA mutational status and PTEN protein expression level were measured [25]. As expected, Akt phosphorylation at Thr308 and Ser473 was present at higher levels in low-PTEN compared with high-PTEN tumors, as was mTOR and p70S6K phosphorylation. There were no significant differences in Akt, GSK3, mTOR or p70S6K phosphorylation between mutant PIK3CA versus wild-type PIK3CA tumors, suggesting that PTEN loss and PIK3CA mutations have markedly different functional effects on the PI3K pathway in breast cancer. Of note, the interactome of Akt has also been described using MS-based proteomic technologies [26, 27].

In another study [28], RPPA has been used to evaluate the protein expression of cyclins B1, D1 and E1 in a series of breast cancers in parallel with other technologies, such as immunohistochemistry (estrogen receptor, progesterone receptor, HER2), FISH (HER2), transcriptional profiling (molecular subtypes), gene copy number by array comparative genomic hybridization and mutation status (PIK3CA, P53). The expression levels of cyclins B1, D1 and E1 were differentially altered in the different subtypes, suggesting distinct roles according to the subtype. Cyclin B1 was the cyclin most strongly associated with poor prognosis in patients with hormone receptor-positive tumors. Its gene expression was correlated with that of PLK, CENPE, and AURKB, which encode proteins targeted by drugs under development that may be of particular interest in the treatment of these tumors.

**MS-Based Approaches**

In contrast to microarray-based approaches, the MS-based technologies are exempt of a priori assumptions, i.e. they do not require any biological foreknowledge, and permit the examination and quantification of a large number of initially unknown protein parameters.

After appropriate extraction and separation of proteins from given biological samples, MS-based approaches may be used in two complementary ways in order to identify clinically relevant biomarkers: (1) actual protein quantification and identification by increasingly sophisticated MS, and (2) protein profiling to generate a multiprotein signature. These technologies are still beset by significant limitations, such as overall sensitivity, reproducibility in a high-throughput context and difficult access to the low-abundance proteome; however, in contrast to protein arrays, they are not limited by the necessity of a priori biological knowledge and, therefore, have the potential to discover previously unrecognized protein biomarkers.

**MS-Based Protein Identification and Quantification**

Here MS is used to actually identify peptides and/or proteins, either intact or after enzymatic digestion, in a sample. In matrix-assisted laser desorption ionization-based (MALDI) MS (MALDI-MS), ionized peptides are generated from samples in a solid phase (after cocrystallization with a chemical matrix), whereas in electrospray ionization (ESI)-based MS, ionized peptides are produced from a liquid phase and are sprayed as highly charged thin droplets. Thereafter, MALDI- and ESI-generated peptides are classically analyzed by measurement of their time-of-flight (TOF-based MS), which is tightly correlated with their mass (mass to charge or m/z ratio) [29, 30]. Most frequently, MS analysis is applied to samples following enzymatic digestion (bottom-up proteomics), and protein identities may be obtained in two ways: (1) peptide mass fingerprint, where experimentally measured masses are compared with theoretical masses in the protein databases, and (2) tandem MS (MS/MS) sequencing, where ionized, digested and MS-measured peptides are then subsequently picked up for a fragmentation step and a second MS-based measurement, providing almost complete peptide sequences [31].
In both cases, a protein separation step is necessary before identification. The most famous separation procedure is two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), which allows the simultaneous study of up to several thousands of proteins on a single gel. In 2D-PAGE, extracted proteins are first separated according to their isoelectric point (pH gradient) and then according to their mass (sodium dodecyl sulfate polyacrylamide gel electrophoresis) [32]. A large number of 2D-PAGE-based procedures have been applied to preclinical models of breast cancer, but relatively few data have been generated from clinical samples, essentially comparing in situ or invasive breast cancers to normal tissues [33–42] or analyzing biological fluids [43, 44]. Recently, 77 adenocarcinomas from different primary sites, including 9 breast cancers, were analyzed [45]. Some tumor site-specific proteins were identified by MS/MS and contributed to the construction of an algorithm allowing discrimination of the distinct primaries. Niméus et al. [46], using 2D-PAGE, identified differentially expressed proteins according to the presence or absence of a metastatic relapse in 20 patients with node-positive early breast cancer after adjuvant chemotherapy with cyclophosphamide, methotrexate and fluorouracil. Another study compared the proteome extracted from primary cultures of epithelial cells obtained from 23 surgically resected breast cancers with or without subsequent metastatic relapse [47]. Three protein spots differed between the metastatic and non-metastatic groups and were identified by ESI-MS/MS. Nucleophosmin was increased in the group with metastases, whereas 2,3-trans-enoyl-CoA isomerase and glutathione peroxidase 1 were decreased. 2D-PAGE was also used to compare 13 pairs of matched primary breast cancers and axillary lymph node metastases [48]. The quantification of 135 well-resolved and MALDI-TOF MS-identified proteins revealed striking similarities in their overall expression patterns, although distinct changes in the levels of individual proteins were also observed in some sample pairs. In this study, a similar comparison was performed at the genomic level using array comparative genomic hybridization and confirmed high similarities between the matched primary tumors and the lymph node metastases. These data suggested that key biological characteristics of the primary breast tumor are maintained in the corresponding lymph node metastases.

Although still being considered as a major procedure, 2D-PAGE has limited usefulness for analysis of extreme-range molecular weights or pHs as well as hydrophobic or...
low-abundance proteins. Its major disadvantage lies in its low reproducibility and limited high-throughput capability. To improve the sensitivity, reproducibility and rapidity of conventional 2D-PAGE, two-dimensional differential gel electrophoresis was introduced [49]. Here protein extracts are first labeled by fluorochromes (cyanines: Cy2, Cy3 or Cy 5), then identical amounts of the samples to be compared are mixed and simultaneously resolved by 2D-PAGE. The pattern of migration is then visualized by a fluorescence imager after exciting the gel using 2 specific wavelengths (Cy2 and Cy3 specific). Images are analyzed and each spot is quantitatively compared between samples (fig. 1). This procedure drastically reduces the intergel variability that may be observed with 2D-PAGE and improves the accuracy of the quantification. This approach was recently applied to breast cancer cells derived from micrometastasis and revealed a stem cell protein profile [50].

Liquid chromatography (LC), where peptides in solution are injected through chromatographic columns, represents an alternative to 2D-PAGE for separating proteins. Retained species are dependent on the chromatographic properties of the column; they are then eluted and evaluated by MS instruments thereafter [29]. For example, a conventional high-performance liquid chromatography (HPLC) procedure was coupled to MALDI-MS to identify differentially expressed peptides from estrogen receptor-positive (n = 39) and estrogen receptor-negative (n = 41) invasive breast cancers [51]. In addition, HPLC may include several sequential steps of complementary chromatographic procedures, as in multidimensional protein identification technology [52]. Two steps of HPLC (cation exchange and reverse phase) are coupled to MS/MS and database-searching algorithms and allow rapid analysis of complex mixtures with direct identification of the generated peptide sequences. This technique has recently been associated with enzyme activity profiling in human breast tumors for generating functional signatures associated with the molecular subtypes [53].

To improve the quantitative capabilities and reproducibility of MS-based procedures, labeling-based technologies have recently been proposed, i.e. isotope-coded affinity tag (ICAT), stable isotope labeling with amino acids in cell culture (SILAC) and isotope tagging for relative and absolute quantification (iTRAQ). In ICAT-based technology, paired complex protein samples are isotopically labeled with tags, e.g. $^{12}$C (light) and $^{13}$C (heavy) that covalently bind to cysteine residues. Then samples are mixed, digested with trypsin, separated by HPLC and identified by MS. The tags used are similar in structure and chemical properties, but are different in mass. ICAT profiles the relative amounts of cysteine-containing peptides derived from tryptic digests of paired protein extracts [54]. This method was used for comparing the proteome of nipple aspirate fluids from cancerous breast tissues and from noncancerous contralateral breast tissues [55].

In the SILAC procedure, cells to be compared are metabolically labeled during the culture phase using supplementation of their medium with distinct isotopic forms of the same amino acids, i.e. ‘heavy’ (for example, arginine and lysine $^{13}$C$_6$) or ‘light’ (arginine and lysine $^{12}$C$_6$) forms. After labeling, cell lysates are mixed and subjected to LC-MS/MS. Peptides with incorporated heavy arginine and/or heavy lysine have a 6-Da higher mass of per amino acid relative to peptides with light isotopic arginine and/or lysine. The intensity ratios of the heavy and light peptide peaks, respectively, in a given mass spectrum indicate the relative abundance of proteins in both cell types. This approach was used to compare two isogenic human breast cancer cell lines that are equally tumorigenic in mice, but with very different metastatic capacities [56]. Following membrane purification, a SILAC-based comparative quantitative LC-MS/MS proteomics approach identified 13 overexpressed and 3 underexpressed membrane proteins in the metastatic versus nonmetastatic cell line among a total of 1,919 identified protein entries. These results were validated by flow cytometry, Western blotting, and immunocyto- and immunohistochemistry. Analysis of clinical breast cancer biopsies demonstrated a correlation between high expression of some of these proteins (ecto-5’-nucleotidase and integrin B1) and poor outcome.

The iTRAQ approach involves labeling of the primary amines of tryptic peptides with tags of varying mass. The samples to be compared are then pooled and usually fractionated by nano-LC and analyzed by MS/MS. This approach offers a robust and reproducible means for simultaneous identification and quantification of all peptides, including those with posttranslational modifications. This approach possesses the key advantage of enabling simultaneous analysis of up to 8 different biological specimens thereby increasing the analysis throughput while reducing experimental errors. Additionally, the resulting fragmented peptide product ions subjected to MS/MS are indistinguishable for the same protein originating from the different biological specimens. This leads to an improved signal-to-noise mass-spectral response for the peptide precursor (MS) and product (MS/MS) ions leading to more confident peptide identification, to an extent
not feasible with ICAT or SILAC. The relative quantification is achieved via the differences in the abundances of the reporter product ions (i.e. m/z ratio 114, 115, 116, 117) [57]. The iTRAQ approach has been applied in various studies, including the study of phosphoproteomic changes during breast cancer development [58]. Recently, iTRAQ was used to compare 3 human breast cancer samples with distinct metastatic phenotypes (2 primary breast tumors associated with metastatic lymph nodes or not, and 1 lymph node metastasis). The study resulted in the reproducible identification of 605 nonredundant proteins [59]. A quantitative comparison revealed 6 proteins with altered levels in metastatic versus nonmetastatic primary tumors and 19 proteins with altered levels in lymph node metastasis compared to nonmetastatic primary tumors. Changes in selected differentially expressed proteins were validated using real-time quantitative reverse transcription polymerase chain reaction.

Recently, there have been marked improvements in MS performance owing to improved instrumental design, such as ion trap [60], quadruple TOF or Fourier transform ion cyclotron resonance MS [61]. These more accurate and sensitive systems are expected to identify a significantly larger fraction of the proteome as compared with the conventional techniques and thus should become increasingly used for protein biomarker discovery. In addition, they may allow the analysis and accurate protein identification from entire proteins (top-down proteomics) without initial proteolytic digestion. A recent study used nano-LC coupled with Fourier transform ion cyclotron resonance MS to compare proteomes from recurrent breast tumors with distinct sensitivity to tamoxifen [62]. Analysis was performed on 25,000 pooled tumor cells obtained by laser capture microdissection from two independently processed data sets (n = 24 and n = 27) containing both tamoxifen-sensitive and tamoxifen-resistant tumors. A total of 17,263 unique peptides were identified, corresponding to 2,556 nonredundant proteins identified with more than 2 peptides. The 1,713 proteins common to the 2 data sets were selected for further analysis. Comparative analysis revealed 100 proteins differentially abundant between tamoxifen-sensitive and tamoxifen-resistant tumors, 47 of which were further verified by targeted LC-MS/MS on an independent data set. The most differentially expressed protein, extracellular matrix metalloproteinase inducer, was also validated using immunohistochemistry on tissue microarrays in an independent cohort (n = 156) with expression associated with a shorter time to progression on tamoxifen.

However, the major part of these achievements concerns higher sensitivity, resolution and mass accuracy, but the procedures still require very significant preanalytical work, which needs to be accommodated to the high-throughput setting often required in biomarker discovery. In addition, these very accurate techniques are not necessarily the most powerful in terms of quantification and may require additional external or internal additives to allow effective, reliable and reproducible quantitative comparisons between large numbers of clinical samples.

**SELDI- and MALDI-Based Protein Profiling of Clinical Samples**

MS-based approaches are also used as protein-profiling tools, before the actual identification of relevant proteins. Here biological samples (whole or fractionated) are, more or less directly, separated by miniaturized chromatography and subjected to TOF-based MS, generating for each sample a large list of protein peaks that are analyzed in terms of correlation with a clinically or biologically relevant phenotype.

Thanks to the relative simplicity of sample preparation, high analytical sensitivity and speed of data acquisition, surface-enhanced laser desorption/ionization (SELDI)-TOF MS, a variant of MALDI-TOF MS, was developed especially for clinical purposes [63]. In the SELDI procedure, biological samples (e.g. serum or tissue lysate) are cocrystallized with an energy-absorbing matrix on a solid support, comprised of various chromatographic surfaces, enabling an active role in sample fractionation. Subsequent irradiation with brief laser pulses sublimates and ionizes the proteins out of their crystalline matrix. Then the charged proteins migrate in an electric field to the TOF mass analyzer. Herein proteins are separated based on their mass, as the time to detector impact is proportional to the protein mass per charge. An adaptation of the SELDI concept has been recently proposed as well; it associates derivatized magnetic bead surfaces to conventional MALDI-TOF MS instruments. Like SELDI planar surfaces, these beads can be derivatized with multiple chemical affinity properties or biological capture components such as antibodies.

From a given biological sample, both technologies enable us to generate a protein profile composed of a list of protein peaks that are characterized by their masses and their relative intensities. Mass spectra are then processed by bioinformatic algorithms and analyzed using univariate and multivariate biostatistical tools to yield a single-marker or a multimarker protein profile that can accu-
Proteomics of Breast Cancer

In breast cancer, SELDI-TOF MS was used to investigate serum/plasma, nipple aspirate fluids and tumor tissues as a potential source for diagnostic, prognostic or predictive biomarkers. Some studies searched for serum biomarkers that differentiate breast cancer patients from those with benign disease and/or healthy subjects [72, 73]. Enrolling between 133 and 356 patients, these studies have identified diagnostic protein profiles with sensitivities and specificities of 76–96% and 85–93%, respectively. While attempts at independent validation failed for some of them [74, 75], one study [76] reported excellent performance of a 7-peak classifier (not structurally identified) in an independent validation set analyzed 14 months after the initial identification. In addition, 2 previously identified biomarkers (8.1 and 8.9) [77] were further validated in an independent cohort, and identified by MS/MS, as complement component C3a$_{\text{desArg}}$ and a C-terminal-truncated form of C3a$_{\text{desArg}}$. Another 4.3-fragment, identified as inter-alpha-inhibitor heavy chain 4 (ITIH4) fragment, was not validated. This fragment had been found to be increased in breast cancer [78], while similar ITIH4 fragments were found to be either increased [79] or devoid of discriminative power [80]. Given the inconsistent regulation observed across multiple studies, the definitive value of the different ITIH4 fragments, C3a$_{\text{desArg}}$ and C3a$_{\text{desArgD8}}$ in the diagnosis of breast cancer cannot be determined yet. SELDI has been applied to plasma from HER2-positive breast cancer patients and control patients [81]; a set of 7 biomarkers distinguished disease from healthy samples, and was validated in an independent data set. Another serum-based study with diagnostic purpose [82] compared serum samples from patients with $BRCA1$ mutations who either developed cancer ($n = 15$) or remained cancer-free ($n = 15$). It revealed 23 markers associated with specimens with $BRCA1$ mutations that classified the 2 groups with 87% sensitivity and specificity. This study also compared the serum profiles of patients with $BRCA1$-mutated cancer and with sporadic breast cancer. Various proteins were upregulated in the $BRCA1$ specimens and classified the 2 groups with 94% sensitivity and 100% specificity. However, no protein identity was reported, and the limited sample size precluded independent validation. Serum- and plasma-based diagnostic studies by SELDI-TOF MS have yielded numerous protein peaks with significantly different expression between breast cancers and healthy controls. However, although elucidation of protein identity is essential for insights into the pathophysiology of disease, so far, only a small percentage of reported peaks has been structurally identified. Moreover, since most studies did not investigate other cancer types or patients with benign breast disease, the specificity of the reported markers for breast cancer remains to be established. Furthermore, although of pivotal importance, only a few potential markers have been validated in independent sample sets. As these studies generally yielded contradictory results, further research is needed to determine the actual potential of identified markers in breast cancer diagnosis [83].

Serum-based protein profiling was also investigated as a prognostic tool, aiming to predict patient outcome following a given treatment. We retrospectively analyzed postoperative serum samples from 81 early breast cancer patients receiving postoperative anthracycline-based chemotherapy [84]. Proteins differentially expressed according to metastatic outcome were selected, from which combined supervised partial least-squares projection and logistic regression analysis generated a 40-protein model that correctly predicted the clinical outcome in 83% of patients and identified 2 classes with different survival. This model was validated using leave-one-out cross-validation. Some of its components were identified: haptoglobin-$\alpha_2$ and transferrin levels were upregulated, whereas C3a complement fraction was downregulated in serum from patients with a high probability of relapse. All these proteins are not tumor specific and likely derive from host response. However, some of them are involved in biological processes such as angiogenesis and immune...
response, and are known to influence the metastatic risk. However, these results still require independent validation and must be considered as exploratory. A recent similar study [85] found the haptoglobin-α₁ chain to be associated with recurrence-free survival, but in an opposite way (downregulated in patients with the probability of relapse). However, those results were not confirmed following validation by haptoglobin phenotyping of a 6-fold larger sample set (n = 371).

Besides serum or plasma, nipple aspirate fluids and ductal lavage fluids collected from cancer and noncancer patients have been compared [86–88]. In nipple aspirate fluids, the protein profiles were heterogeneous among different patients, whereas the similarity between nipple aspirate fluids from cancerous and normal contralateral breast from a given patient was surprisingly high. However, some protein peaks were found to be differentially expressed between cancer and noncancer samples, some of them being correlated with axillary node involvement and the extent of the disease [89].

SELDI-based profiling approaches have also been applied to breast tissues for identifying proteins differentially expressed between normal and cancer cells or tissues [51, 90–92]. Four studies have examined the prognostic value of SELDI-based protein profiling in breast cancer samples. Ricolleau et al. [93] analyzed protein cytosolic extracts from 60 node-negative early breast cancers; they found high levels of ubiquitin and/or low levels of ferritin light chain to be associated with a good prognosis. These data were verified by immunohistochemistry and Western blot. Another study examined laser-micromass dissected invasive breast cancer cells and identified 2 proteins whose expression correlated with lymph node involvement. Interestingly, one of them (with low expression in node-positive disease) had an m/z ratio of 8.5 kDa, consistent with ubiquitin [94]. We have recently profiled the cytosol of 27 human breast cancer cell lines [95]. Comparison of basal and luminal cell lines – as defined by genomics – generated a list of potential biomarkers. The MALDI-based approach identified ubiquitin as upregulated in ‘luminal-like’ cells (fig. 2). Interestingly, the peak corresponded to a posttranslational alteration of ubiquitin (namely, C-terminal truncation, which was demonstrated by differential enzymatic digestion followed MALDI-TOF MS measurements of the obtained peptides), rather than the full-length form. This was further validated in a small-sized sample of frozen luminal or basal breast cancers, suggesting the clinical pertinence of this information. In the same protein list, we identified S100A9 as a potential basal biomarker, in agreement with the presence of the corresponding gene in the transcriptionally defined basal gene cluster. Further validation came from immunohistochemistry on a large tissue microarray of early breast cancers: we showed that S100A9 expression was associated with other basal markers and shorter survival. Recently, a study profiled protein lysates from 105 breast cancers using IMAC-Cooper ProteinChip arrays and SELDI-based analysis [59]. When protein pro-
files obtained by SELDI were subjected to an unsupervised hierarchical clustering procedure, different groups of breast cancer were identified, and correlated with tumor type, nuclear grade and expression of hormonal receptors, mucin 1 and cytokeratins 5/6 or 14. These tumor groups closely resembled luminal A and B, basal and HER2-like carcinomas. Some discriminating proteins were identified, including annexin V and HSP27. Altogether, these results suggest that SELDI profiling of breast cancer identifies clinically relevant groups of patients similar to those previously defined by genomics.

Although some of them have been considered promising, the results from SELDI- or MALDI-based protein profiling have mostly been disappointing. Very few candidate protein peaks have been structurally identified yet and even fewer have been validated in large, independent series, as illustrated in table 2. For fluid-based approaches, the candidate markers that have been identified constitute normal cellular proteins and highly abundant blood proteins involved in coagulation and acute-phase response, indicating a general lack of tumor specificity [99]. In fact, as tumor-specific proteins are expected to be among the least abundant proteins, they could be below the detection limit of conventional SELDI- and MALDI-based methods, but could be revealed by more innovative MS-based devices or additional prefractionation methods. Thus, SELDI- or MALDI-based protein profiling of biological fluids without additional preprocessing methods must now be considered as largely infraoptimal for biomarker discovery; other, more recent technologies should be deployed.

### Discussion and Perspectives

This retrospective survey of pioneering retrospective studies shows that proteomics represents a promising domain of research in the diagnosis and prognosis of breast cancer. The potential biomarkers identified, consisting of a single or several proteins, represent a bar code signature of the phenotype of interest. Of course, whether they are directly associated with the phenotype in a biological sense or reflect another associated phenomenon remains to be explored. Due to their preliminary nature, the reported results raise some criticisms and highlight challenges that should be addressed in the near future to allow proteomics to reach its optimal yield and translate into tangible benefits for patients.

The sample size is relatively small and the results must be validated, refined and extended in larger series. In this context, immunohistochemistry on tissue microarrays is the most interesting approach, provided that the appro-

<table>
<thead>
<tr>
<th>Objectives</th>
<th>Ref.</th>
<th>Learning phase</th>
<th>Validation phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample type</td>
<td>patient/sample population, n</td>
<td>platform</td>
<td>sample type, n</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>[91]</td>
<td>LCM tissue invasive BC vs. NB (62 vs. 84) MALDI</td>
<td>[91]</td>
</tr>
<tr>
<td></td>
<td>[76]</td>
<td>serum invasive BC vs. HC (155 vs. 155) SELDI</td>
<td>[76]</td>
</tr>
<tr>
<td></td>
<td>[96]</td>
<td>serum BC vs. HC and BD (103 vs. 41 vs. 25) SELDI</td>
<td>[77]</td>
</tr>
<tr>
<td></td>
<td>[96]</td>
<td>serum BC vs. HC and BD (103 vs. 41 vs. 25) SELDI</td>
<td>[77]</td>
</tr>
<tr>
<td>Prognosis</td>
<td>[95]</td>
<td>plasma HER2+ BC vs. HC (61 vs. 61) SELDI</td>
<td>[95]</td>
</tr>
</tbody>
</table>

LCM = Laser capture microdissected; BC = breast cancer; NB = normal breast; HC = healthy control; FFPE = formalin-fixed paraffin-embedded; TMA = tissue microarrays.
appropriate antibody is available or can be made. As shown in table 2, only a few studies have seen their results validated in independent series, and there is no consistency of data between studies. Several aspects may explain such discrepancies. There is great interstudy heterogeneity regarding patient populations (histoclinical features, treatment and follow-up), the primary end-point of prognostic studies and the method of analysis of survival (either as a dichotomous variable but with different cut-offs, or as a continuous variable). Other sources of heterogeneity are technological, with the use of different technological platforms, different protein sets and different methods of data analysis. Combined with the heterogeneity of breast cancer, all these differences may explain the discrepant results reported in some studies. As shown in table 2, only a few studies have included an independent validation set – although crucial in high-throughput analyses due to the risk of ‘overfitting’ [100] – and multivariate analyses testing the predictive independence of the biomarker when compared with classical histoclinical features. Finally, no molecular target, identified in poor-prognosis patients that could provide an alternative therapeutic solution, has clearly emerged as yet.

Several challenges will have to be addressed in the future. The accessibility, more or less restricted to some companies and academic facilities during the past years, is increasing rapidly, and the availability, complexity and cost of other proteomics tools will likely improve with the next-generation tools. A major challenge appears to lie in the complexity and enormity of the proteome and its dynamic state as well as the molecular complexity of breast cancer. Although being an extremely rapidly moving field of technological innovations, proteomics approaches still have to achieve some qualitative revolutions before identifying biomarkers that will reach routine clinical application. Even though current MS-based approaches are increasingly sensitive, they are still failing to achieve large access to low-abundance proteomes in complex biological samples such as serum or plasma. However, low-abundance proteins – within the microgram/milliliter range – likely represent many of the potential tumor-specific biomarkers. This limitation is essentially due to the high dynamic range of techniques, precluding simultaneous measurement of protein quantities that differ by a factor of up to $10^{10}$. Strategies aiming to overcome this drawback have been developed based on depletion/enrichment of a subset of the sample proteome [101], equalization [102, 103], or subfractionation for analyzing, for example, the phosphoproteome, glyco-proteome or subcellular locations [104, 105]. As observed with all ‘omics’ tools, the handling of data for multiprotein experiments is a major bottleneck. Progress needs to be accomplished for producing, storing, analyzing, visualizing and interpreting data, and previous work on gene profiling data constitutes a precious base [106]. Efforts are also required to make results publicly available. This will facilitate not only external validation of results and meta-analyses, but also the development of analytics tools.

Another major issue is the all too pressing urge to have a high-quality sample bank (tumors and fluids) linked to a searchable database containing all histoclinical parameters of tumors as well as documenting patient consent and ensuring confidentiality. The careful and uniform collection of samples and data should become a component of all future clinical trials. As proteins are very sensitive to degradation, it is crucial to guarantee the quality of biological sampling and storage in carefully designed prospective collections.

Technological improvements are ongoing that may allow MS-based tools to be applied to formalin-fixed paraffin-embedded tissues [107, 108], which still constitute the main source of tumor material in pathological depositories. The amount of material required for experiments is another important issue because of the small size of clinical specimens and the absence of techniques for amplification of proteins. Improvements in microtechniques and nanotechnologies may soon provide ways to circumvent this obstacle.

Another ongoing major challenge is to develop kits that can simultaneously extract DNA, RNA and proteins from frozen samples.

Finally, it is clear that data from proteomics must be combined with data from other ‘omics’ approaches in an attempt to investigate mechanistic pathways in more detail as well as providing diagnostic or prognostic information in the clinical setting. This may represent a novel field of application of systems biology in order to combine multiple sources of data (e.g. genomics, proteomics and metabolomics) into a single unified hypothesis to explain and/or shed light on the biological processes involved in oncogenesis.

**Conclusion**

Clinical proteomics of breast cancer is still in its infancy and remains in the research field. Even if promising results have been reported, there has been no breakthrough that is close to becoming clinically applicable. A
number of putative diagnostic and prognostic markers or signatures have been found, but their use in the clinical routine is premature, and no new targeted therapy has yet emerged. There is no doubt that many other potential markers will be identified in the next years. The challenge will be to perform correctly designed large retrospective studies including independent validation sets, followed by prospective validation studies to demonstrate the clinical benefits for patients. Even if a number of issues remain to be addressed before proteomics can enter clinical practice, it is anticipated that in the future such a promising approach will affect the current generation of physicians in charge of breast cancer patients.

It is too early to define the format of the proteomic tool that will be applied in clinical practice. One example includes protein arrays that combine quantitative and functional information, by detecting the activation of kinases or signaling pathways to be targeted by specific antitumor therapies. Characterization of discriminator proteins will provide new markers for screening, diagnosis, prognosis and follow-up, and will help developing new molecularly targeted anticancer drugs. Faced with the complexity of cancer proteomics and its great clinical potential, a global coordination of ongoing efforts appears to be crucial.

References


