Proteomics of Breast Cancer

PRINCIPLES AND POTENTIAL CLINICAL APPLICATIONS*

François Bertucci‡§¶, Daniel Birnbaum‡, and Anthony Goncalves§¶**

Progresses in screening, early diagnosis, prediction of aggressiveness and of therapeutic response or toxicity, and identification of new targets for therapeutic will improve survival of breast cancer. These progresses will likely be accelerated by the new proteomic techniques. In this review, we describe the different techniques currently applied to clinical samples of breast cancer and the most important results obtained with the two most popular proteomic approaches in translational research (tissue microarrays and SELDI-TOF). Molecular & Cellular Proteomics 5:1772–1786, 2006.

Worldwide, breast cancer is one of the most frequent and deadly cancers. Although the survival of patients has increased over the last decades in relation with screening programs and postoperative adjuvant systemic therapies (hormone therapy and chemotherapy), many patients die from metastatic relapse. Breast cancer is a complex disease. Accumulation of numerous and often unknown molecular alterations causes cell proliferation, genetic instability, and acquisition of an increasingly invasive and resistant phenotype. The combinatorial origin and the heterogeneity of malignant cells and the variability of the host background create molecularly distinct subgroups of tumors endowed with different phenotypes and clinical outcomes. This heterogeneity is only partially apprehended by the clinicopathological parameters currently used by clinicians, hampering the selection of the most appropriate diagnostic or therapeutic strategy for each case. In addition, current cytotoxic drugs do not differentiate between cancerous and normal cells, causing sometimes disastrous adverse effects. Major efforts aim at characterizing the heterogeneity of the disease by defining more reliable diagnostic/prognostic factors and at developing molecular therapies selectively targeting the tumor cells.

For some decades, the study of molecular alterations has successfully elucidated some mechanisms of mammary oncogenesis and identified key genes such as ERBB2, TP53, CCND1, BRCA1, and BRCA2. It has also allowed considerable therapeutic progress by targeting hormonal receptors and ERBB2/HER2 receptor. Today high throughput molecular typing provides an unprecedented opportunity to tackle the combinatorial aspect and the complexity of breast cancer. A combination of markers (molecular signature) will probably be more sensitive and specific than a single molecular marker to reflect the actual heterogeneity of disease; more reliable for screening, diagnosis, prognosis, and prediction of therapeutic response; and more useful to find new therapeutic targets. The first large scale techniques applied to the cancer field were DNA microarrays for mRNA expression profiling (1). Several studies have demonstrated the clinical potential of this approach, notably by identifying new biologically relevant and prognostic subclasses of breast cancer unidentifiable by conventional means (2–14). With the recent development of similar technologies at the DNA and protein levels, there is no reason to limit profiling to RNA.

Proteomics has the potential to complement and further enlarge the wealth of information generated by genomics in breast cancer for several reasons. mRNA levels do not necessarily correlate with corresponding protein abundance (15–19). Additional complexity is conferred by protein post-translational modifications, including phosphorylations, acetylations, and glycosylations, or protein cleavages (20). These modifications are not detectable at the mRNA level but play significant roles in protein stability, localization, interactions, and functions. Finally proteins represent more accessible and relevant therapeutic targets than nucleic acids.

However, for a long time, protein measurements have been considered more difficult than DNA or RNA analyses. Composed of 21 naturally occurring amino acids, proteins can represent a virtually unlimited variety of molecules that has been estimated from 1 to 20 million in mammalian organisms. Moreover protein abundance in a biological sample may vary from few copies to several millions. Consequently an accurate quantitative estimation, most often from limited material in clinical samples, requires both high sensitivity and dynamic range. Although these requirements are only partially achieved yet, several technical strategies have been developed during the past 10 years. They allow access to the proteome in various biological samples with still preliminary but promising results in the clinical setting. Breast cancer studies have concerned tumor tissue but also biological fluids including serum, plasma, nipple aspirate, or ductal lavage material with different intents, including better understanding of mammary oncogenesis and improvements in screening.

*From the ‡Centre de Recherche en Cancérologie de Marseille, Département d’Oncologie Moléculaire, Institut Paoli-Calmettes (IPC) and UMR599 INSERM, §Département d’Oncologie Médicale and **Département de Pharmacologie Moléculaire, IPC, and ¶Faculté de Médecine, Université de la Méditerranée, 13009 Marseille, France. Received, May 24, 2006
Published, MCP Papers in Press, May 29, 2006, DOI 10.1074/mcp.R600011-MCP200

© 2006 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org
and diagnosis as well as prognosis and/or prediction of therapeutic response.

In this review, we will first present the main characteristics of proteomics-based methods that have been applied to date to breast cancer samples. Then we will discuss the preliminary results generated from clinical samples by the most common approaches and the perspectives of protein profiling.

PROTEOMICS METHODS

The proteomic techniques may be roughly distinguished according to the requirement or not for an initial biological foreknowledge. Approaches that require “a priori” hypotheses or selections include microarray-based techniques. They are essentially dominated by tissue microarrays (TMAs), which address protein expression of candidate markers on a large number of arrayed tumor tissues. Other microarray-based techniques include antibody arrays and reverse-phase protein arrays, which are promising methods but still in their infancies with respect to clinical applications. Approaches without a priori requirement, which do not suppose any previous biological foreknowledge, allow examining and quantifying a large number of initially unknown protein parameters. They are used to identify unique or multiple protein markers, which correlate with a pertinent tumor phenotype. This kind of approach is essentially based on mass spectrometry tools, most often following an initial chemical or physical separation of the analyte.

Tissue Microarrays

Increasing identification of molecular markers with potential clinical significance by high throughput techniques has boosted the development of fast approaches to validation in large series of samples. DNA or protein microarrays have some disadvantages (high cost and requirement of unfixed frozen samples) that make them not adaptable for such validation. Initially described ~20 years ago (21, 22), TMA was modernized in 1998 (23). The technique allows the simultaneous analysis at the DNA (fluorescence in situ hybridization), mRNA (in situ hybridization), or protein (immunohistochemistry (IHC)) levels of up to 1,000 tumor samples arrayed onto one microscope glass slide. Cylindrical cores as small as 600 μm in diameter are generated from formalin-fixed paraffin-embedded (FFPE) archived tumor blocks. These cores are arrayed in a new paraffin block that is then cut into thin sections (100–200/ block) for analysis. All samples from a section may be simultaneously interrogated by a specific antibody by IHC as well as morphologically analyzed. Immunostaining is scored by the pathologist. In contrast to the other methods described in this review, the molecular information is obtained in situ in the context of cell morphology and tissue architecture, and specimens are derived from FFPE archival tissues (although TMAs may also be done from frozen tissues (cryo-TMAs)). Advantages of TMA are multiple as compared with traditional IHC on whole sections. TMA is much less time-consuming, labor-intensive, and economically costly; all samples are tested in similar experimental conditions, leading to more reproducible and rigorous data; and the equipment required to make TMAs is not complicated. Finally TMA helps in the management of tissue archives by allowing preservation of tissue resources for research as well as in the testing and optimization of diagnostic tests.

TMA is so far the most used proteomic technique in oncology with more than 1,500 publications in PubMed in April 2006 with the terms “tissue microarrays” and “cancer.” The high number of samples simultaneously profiled reinforces the statistical significance of results. Analysis of individual molecular markers on large series of samples remains the main application of the technique with different objectives with regard to the spotted samples. TMA has been used for the screening of multiple tumor types (“multitumor TMAs”) (24) for the expression of a protein of interest (25, 26). The largest reported multitumor TMA, constructed at the University of Basel, contained up to 4,700 tumors representing 135 different tumor types (27). Another objective is the analysis of different stages of tumor progression for a given cancer as reported for bladder (28, 29) or prostate cancers (30, 31). If follow-up data are available, TMAs may be used to measure the prognostic significance of a given protein as reported in colorectal (32, 33), renal (34), and bladder (28) cancers. TMAs can also be constructed from cell lines (35) or other experimental material such as xenografts (31). In these applications, the tested individual markers have previously known or suspected clinical relevance or represent novel markers derived from high throughput molecular analyses. More recently, TMAs were used to identify among tens of tested proteins a multiprotein expression signature associated with a given phenotype. The multidimensionality of the dataset imposed the use of complex analytic tools similar to those used for analysis of quantitative DNA microarray data (36), but the task was complicated by the qualitative or semiquantitative nature of the IHC data. Unsupervised hierarchical clustering was first applied and allowed the identification of coherent tumor and protein clusters. Until now, supervised methods have not been frequently used to define multiprotein predictors of a given phenotype. These multiprotein approaches were reported in lymphoma (37), uterine (38), and prostate (39) cancers; melanoma (40); and sarcoma (41). Several criticisms have been formulated with regard to the use of TMAs in the cancer field. The most recurrent concerned the representativity of a small sample (0.6 mm in diameter) of

---

1 The abbreviations used are: TMA, tissue microarray; IHC, immunohistochemistry; FFPE, formalin-fixed paraffin-embedded; NAF, nipple aspirate fluid; MBC, medullary breast carcinoma; IBC, inflammatory breast cancer; NiBC, non-IBC; EBC, early breast cancer; MFS, metastasis-free survival; ER, estrogen receptor; 2D, two-dimensional; EGFR, epidermal growth factor receptor.
a potentially heterogeneous tumor as compared with a classical large section. In fact, several studies (17, 42) have shown excellent concordance for IHC when two or three samples per tumor were carefully selected by the pathologist for TMA construction. Other criticisms include those related to IHC (subjectivity and reproducibility of manual scoring; qualitative or semiquantitative scoring, which may lack clinically relevant information; inconstant availability of antibodies with sufficient quality for analysis of FFPE samples; and lack of standardization of protocols), the prior knowledge of the proteins to be tested that does not readily contribute to the discovery of new markers as well as the need for bioinformatical tools for analysis of multidimensional qualitative data, or the frequent discordance between mRNA and protein expression levels. In the future, a much more widespread use of TMA is expected and will help minimize these limitations. For example, the automated quantitative analysis of TMA has made progress, providing more rapid, reproducible, and objective results (44–47); several software packages are commercially available (48), and such automation will be a key issue in the future. Yet the first results of TMA studies are extremely encouraging, notably in breast cancer. Results are described below.

Protein Microarrays

These approaches tend to recapitulate at the protein level the mRNA expression profiling studies by arraying various protein probes onto specific surfaces and then measuring interactions with specific proteins in complex samples. The most advanced format in this setting is the antibody microarray where the bait proteins are specific antibodies printed on solid surfaces (49). Samples to analyze are either directly labeled with fluorophores or indirectly labeled with a tag subsequently detected by a labeled antibody. A potential problem is that direct protein labeling may disrupt the binding of an analyte. In addition, because all proteins are labeled, high abundance proteins may bind nonspecifically to the bait molecule, generating high background. A reference sample may be co-incubated with a test sample to normalize for variation between spots in capture antibody concentration. The assay is competitive and generates a linear response according to the concentration of the analyte. This two-color comparative fluorescence strategy has been used to compare protein levels between malignant and normal breast tissues from the same patient (50). It used a commercially available antibody microarray (Clontech, Mountain View, CA) targeting 378 different proteins. Several proteins had higher levels in the malignant tissue, which was confirmed by IHC. A similar approach using the Panorama™ Ab Microarray-Cell Signaling array (Sigma) has been applied to adipose tissue of mammary gland from high risk breast cancer patients and revealed the expression of a large number of proteins potentially involved in the regulation of the breast tumor microenvironment (51).

Alternatively labeling can be done after array incubation with a second specific antibody supposed to bind the same protein as the bait antibody but on a different epitope. This antibody can be linked to a detection system ("sandwich" method) such as fluorophores, resonance light scattering, enhanced chemiluminescence, tyramide signal amplification, and streptavidin systems. This kind of strategy has been applied to biological fluids related to breast cancer samples. Medium from human breast cancer cells in culture was incubated on cytokine antibody array membranes, revealed by biotin-tagged antibodies, horseradish peroxidase-conjugated streptavidin, and the ECL system. Interleukin-8 was associated with ER status and metastatic potential of breast cancer cells (52). Similarly the expression of various cytokines was studied using cytokine-dedicated antibody microarrays in serum, tumor, or fat tissue interstitial fluids from patients (51, 53, 54).

Protein and antibody arrays may also provide information on the post-translational modifications of targeted proteins. For example, secondary antibodies used in sandwich assays may target either total proteins or the active phosphotyrosine forms. This has been applied to breast cancer cell lysates to evaluate the activation of ERBB receptors upon ligand binding and after specific perturbation by receptor inhibitors (55).

Reverse-phase protein array is another promising approach (56). Protein samples are robotically spotted onto specific surfaces and then probed by appropriate antibodies with labeled secondary antibodies and a detection/amplification system. This method, like the TMAs, allows various samples to be analyzed in a single experiment but with a limited throughput of antibodies to test.

A number of reports have demonstrated the utility of protein arrays, but their wide application in cancer research is still limited in part because of the cost of producing antibodies and the limited availability of antibodies with high specificity and affinity for the target. As already mentioned, antibody arrays are generally hypothesis-driven approaches, which require a previous knowledge of a limited number of candidate proteins to evaluate their expression among different samples. New high throughput antibody production and screening methods could allow efficient probing of a large number of uncharacterized proteins. This has been demonstrated with phage display-based strategies where a large library of peptides can be expressed by phages and then secondarily probed with cancer patient serum to identify novel tumor antigen peptides with immunoreactivity associated to malignant disease (57) as well as autoantibodies with potential diagnostic values (58).

Mass Spectrometry-based Approaches

These approaches take advantage of the ability of mass spectrometers to separate peptides or proteins according to their m/z. Mass spectrometer measurements may serve to identify peptides after tryptic digestion of proteins separated
subjected to analysis, which is correlated to their identification of proteins separated by 2D electrophoresis. Mass spectrometry-based proteomics may include an initial phase of protein separation. The standard is 2D electrophoresis (59). In 2D electrophoresis, proteins are separated according to two independent physicochemical parameters, isoelectric focusing and molecular weight. After successful separation and appropriate gel staining, proteins of interest are isolated from the gel, digested, and subjected to mass spectrometry. Peptide mass mapping by MALDI-TOF analysis is now a routine tool for the identification of proteins separated by 2D electrophoresis (60). MALDI-TOF MS calculates the time-of-flight of peptides subjected to analysis, which is correlated to their m/z. These peptide masses are used to search sequence databases and match with a known protein. The efficiency of MS is dependent on the development of comprehensive sequence databases and expressed sequence tag databases. When the peptide mass mapping experiment does not yield an unequivocal match, fragment ion measurement by MS/MS may be done. Commercial availability of tandem mass spectrometers with MALDI sources, such as the Q-TOF and TOF-TOF, also allows the acquisition of more extensive fragmentation for one or more peptides to confirm the peptide mass map results. ESI-MS/MS, an alternative to MALDI in terms of ion production (61), is also very frequently used coupled to triple quadrupole, quadrupole ion trap, and quadrupole TOF to generate peptide fragmentation spectra. Peptide digests may also be analyzed by LC-MS/MS. To improve speed, sensitivity, and reproducibility of conventional 2D electrophoresis, 2D DIGE was developed (62). With this method, protein extracts are covalently labeled using different fluorescent dyes, e.g., cyanine (Cy2, Cy3, or Cy5) dyes or Alexa dye. Typically the test sample is labeled with Cy3, and the reference sample is labeled with Cy5. Equal concentrations of the two differentially labeled protein samples are then mixed and co-separated during the same 2D electrophoresis process. The 2D DIGE pattern is then visualized by scanning the gel at two wavelengths (for Cy3 and Cy5) using a fluorescence imager. A comparison of the two images generated allows the quantification of each spot. 2D DIGE eliminates or drastically reduces gel-to-gel variability associated with standard 2D electrophoresis and improves the accuracy of quantitative protein profiling, allowing its use to be envisioned in a high throughput research environment (63).

A large number of 2D electrophoresis-based approaches have been developed in preclinical models of breast cancer, including cell lines, but relatively few data have been generated from clinical samples and generally from a limited number of samples. Studies have described in situ and/or infiltrating ductal carcinoma and compared them with normal tissues or non-malignant tumors (64–71). These approaches have also been applied to biological fluids including serum (72) and nipple aspirate fluids (NAFs) (73). Although 2D electrophoresis has excellent resolution, the requirement for subsequent protein staining and sample handling limits the exhaustivity, sensitivity, and reproducibility of the approach. Other major drawbacks that limit its clinical application include the requirement of a large amount of starting material and its labor-intensive nature.

**Gel-based Approaches**—Mass spectrometry-based proteomics may include an initial phase of protein separation. The standard is 2D electrophoresis (59). In 2D electrophoresis, proteins are separated according to two independent physicochemical parameters, isoelectric focusing and molecular weight. After successful separation and appropriate gel staining, proteins of interest are isolated from the gel, digested, and subjected to mass spectrometry. Peptide mass mapping by MALDI-TOF analysis is now a routine tool for the identification of proteins separated by 2D electrophoresis (60). MALDI-TOF MS calculates the time-of-flight of peptides subjected to analysis, which is correlated to their m/z. These peptide masses are used to search sequence databases and match with a known protein. The efficiency of MS is dependent on the development of comprehensive sequence databases and expressed sequence tag databases. When the peptide mass mapping experiment does not yield an unequivocal match, fragment ion measurement by MS/MS may be done. Commercial availability of tandem mass spectrometers with MALDI sources, such as the Q-TOF and TOF-TOF, also allows the acquisition of more extensive fragmentation for one or more peptides to confirm the peptide mass map results. ESI-MS/MS, an alternative to MALDI in terms of ion production (61), is also very frequently used coupled to triple quadrupole, quadrupole ion trap, and quadrupole TOF to generate peptide fragmentation spectra. Peptide digests may also be analyzed by LC-MS/MS. To improve speed, sensitivity, and reproducibility of conventional 2D electrophoresis, 2D DIGE was developed (62). With this method, protein extracts are covalently labeled using different fluorescent dyes, e.g., cyanine (Cy2, Cy3, or Cy5) dyes or Alexa dye. Typically the test sample is labeled with Cy3, and the reference sample is labeled with Cy5. Equal concentrations of the two differentially labeled protein samples are then mixed and co-separated during the same 2D electrophoresis process. The 2D DIGE pattern is then visualized by scanning the gel at two wavelengths (for Cy3 and Cy5) using a fluorescence imager. A comparison of the two images generated allows the quantification of each spot. 2D DIGE eliminates or drastically reduces gel-to-gel variability associated with standard 2D electrophoresis and improves the accuracy of quantitative protein profiling, allowing its use to be envisioned in a high throughput research environment (63).

A large number of 2D electrophoresis-based approaches have been developed in preclinical models of breast cancer, including cell lines, but relatively few data have been generated from clinical samples and generally from a limited number of samples. Studies have described in situ and/or infiltrating ductal carcinoma and compared them with normal tissues or non-malignant tumors (64–71). These approaches have also been applied to biological fluids including serum (72) and nipple aspirate fluids (NAFs) (73). Although 2D electrophoresis has excellent resolution, the requirement for subsequent protein staining and sample handling limits the exhaustivity, sensitivity, and reproducibility of the approach. Other major drawbacks that limit its clinical application include the requirement of a large amount of starting material and its labor-intensive nature.

**Non-gel-based Approaches**—Other MS-based techniques use separative techniques that are not dependent on gel electrophoresis. A two-dimensional all liquid-phase method combined with MS for protein profile analysis was developed. Proteins are first fractionated by pi using IEF and then further separated according to hydrophobicity by reverse-phase HPLC (74). To improve its quantitative and high throughput capacity and its reproducibility, LC may be coupled to sample isotope labeling as in the ICAT technique (75). In this method, paired complex protein samples are isotope-labeled with tags, e.g. 15C (light) and 13C (heavy) that bind covalently 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 (63). This method has been applied to compare NAFs from tumor-bearing and contralateral disease-free breasts of patients with unilateral early stage breast cancer (EBC), identifying and quantifying differences in various specific protein expressions (76). Another non-gel-based approach is multidimensional protein identification technology (MudPIT). Two steps of HPLC (cation exchange and reverse-phase) are coupled to MS/MS and database-searching algorithms allowing the rapid analysis of complex mixtures with direct identification of the generated peptide sequences. This technique has been recently associated to enzyme activity profiling in human tumor tissues, including breast tumors, and has generated functional signatures that correlate with previously described molecular subtypes (77).

One particular approach for the discovery of high quality protein signatures involves the identification of protein markers directly in tissue biopsy and straightforward analysis of a tissue section using MALDI-MS. This technique has been recently used for tumor samples, mainly lung and brain cancers, with good results in terms of diagnostic or prognostic classification (78, 79). Although only a few breast cancer samples have been analyzed so far (80), this technique represents a promising way to generate protein biomarkers from clinical samples.

SELDI-TOF MS is currently the most widely used and advertised non-gel-based method. The technique couples protein separation directly to presentation to the mass spectrometer. The first dimension of protein separation uses chromatographic substrates (the ProteinChip® array, Ciphergen Biosystems, Fremont, CA), which include anion exchange, cation exchange, normal phase, reverse-phase, and...
IMAC. The various types of substrates have different affinity to different subsets of proteins, thereby increasing protein representation when combining various arrays. The combination of these arrays with up-front prefractionation using 96-well batch chromatography (e.g., anion exchange) allows the detection of up to 2,000 protein species from serum (81, 82). The resulting spectral masses are analyzed using univariate and multivariate statistical tools to yield a single marker or multi-marker pattern that can classify clinical samples. Discriminator protein peaks are then purified and submitted to the MS-based identification process. Knowledge of the chemistry that retains these potential biomarkers helps define the optimal purification process. Another approach may not require the identification of individual protein components but rather deals with discriminating patterns generated by sophisticated bioinformatic tools including artificial neural networks, genetic algorithms, and decision trees (83). The SELDI technique was developed to profile clinical biological fluids, notably serum/plasma, and got fame when numerous studies showed promising potential in identifying unique biomarkers or complex patterns with diagnostic value, allowing its use as a screening/early diagnosis tool in various cancers to be envisioned (84–88). One major criticism of the technique relies on the overall lack of sensitivity and capability to detect tumor-specific protein traces within a large amount of nonspecific protein species (89). This concern is particularly relevant in a screening/early diagnosis setting where tumor burden is expected to be minimal. However, although still controversial in its reproducibility and ability to detect actual specific tumor signatures, SELDI has several advantages, such as ease of use, high throughput, and relatively affordable cost, making it a very attractive technique for working with large sample cohorts in a clinical setting. The first results generated using this technique on clinical breast samples are discussed below.

**CLINICAL APPLICATIONS IN BREAST CANCER RESEARCH**

In this section, we focus on some studies that applied the two most popular proteomic technologies, TMAs and SELDI-TOF, to clinical issues in breast cancer.

**Tissue Microarray Studies**

Since the first publication in 1998 (23), breast cancer has been one of the most analyzed cancers using TMAs. Most studies focused on the clinical significance of a single or a few molecular markers individually on large series of samples. Recently tens of proteins were simultaneously tested in multivariate analysis to identify a multiprotein signature associated with a given phenotype. We present here some studies that addressed the diagnostic and prognostic issues.

**Characterization of Previously Known Diagnostic Classes—** Several studies compared tumor classes within different breast cancer classifications. The genetic classification distinguishes the sporadic forms (~95% of cases) and the hereditary forms (~5%); the latter are mainly due to mutations in BRCA1 and BRCA2 genes. Using DNA microarrays, Hedenfalk et al. (90) profiled breast cancers from BRCA1 (seven samples) and BRCA2 (eight samples) mutation carriers and identified a set of genes including CCND1 that distinguished the two genotypes. Using a TMA that included 23 BRCA1- and 17 BRCA2-mutated tumors, overexpression of CCND1 in BRCA2 tumors was confirmed. The expression of 37 proteins was assessed in 20 BRCA1, 14 BRCA2, and 59 sporadic breast cancers on TMA (91). Proteins included hormone receptors, p53, ERBB2, cell cycle regulators, apoptosis markers, and basal markers. BRCA1 and BRCA2 tumors were different with respect to hormone receptors, cell cycle, and apoptosis and basal cell markers. The most striking difference was in the expression of cell cycle proteins with up-regulation of type D (D1 and D3) cyclins and their cyclin-dependent kinase (CDK4) regulators in BRCA2 cancers with respect to BRCA1 cancers. Most BRCA1 tumors were also ER/ERBB2-negative and highly proliferative and had a basal phenotype.

The most frequent pathological type of breast cancer is ductal breast carcinoma (~80%). Medullary breast carcinoma (MBC) is a rare but enigmatic pathological type. Despite features of aggressiveness, it is associated with a favorable prognosis. Morphological diagnosis remains difficult, and very little is known about the molecular alterations involved in this disease. The expression of 18 proteins (markers of cellular origin, markers related to function and proliferation/differentiation of breast tissue, oncogenes, and tumor suppressors) was studied on 61 typical MBCs and 300 grade III non-MBCs (92). Typical MBCs were characterized by a high degree of basal/myoepithelial differentiation as shown with DNA microarrays (139). In multivariate analysis with logistic regression, typical MBC was optimally defined by high expression of P-cadherin and MIB1/Ki67, ERBB2 negativity, and p53 positivity.

Key proteins can be discovered by comparing the expression profiles of different stages of progression in mammary oncogenesis from normal to metastatic tissues. Expression of multifunctional 14-3-3-σ protein was studied in breast cancer (93). A series of 65 primary breast carcinomas was profiled by Western blotting, 2D electrophoresis, and IHC, providing evidence that, contrary to what was previously thought, 14-3-3-σ down-regulation at the protein level is not a frequent event in breast cancer. Results were validated by TMA on an independent set of 65 tumors. Another study used a “breast cancer progression TMA” including samples from 196 lymph node-negative tumors, 196 lymph node-positive tumors, and three different lymph node metastases from each node-positive tumor. The study demonstrated high concordance in ERBB2 status between primary tumors and their lymph node metastases (94).

Anatomoclinical forms of breast cancer include early forms (~85% of all cases), locally advanced forms (~10%), and metastatic forms (~5%). A rare (~3%) but often lethal form is
inflammatory breast cancer (IBC). Diagnosis, based on clinical and/or pathological criteria, may be difficult, and little is known about the underlying molecular alterations. Two TMA studies compared the protein profile of IBCs and non-IBC (NIBCs). The first study profiled 34 IBCs and 41 NIBCs and confirmed the overexpression of E-cadherin and Rhoc GTPase in IBC (95). We analyzed the expression of eight proteins (ER, progesterone receptor, E-cadherin, EGFR, ERBB2, MUC1, MIB1, and p53) in 80 IBCs and 552 NIBCs (96). All proteins were differentially expressed between IBCs and NIBCs in univariate analysis; five protein expressions were significant in multivariate analysis: high expression of E-cadherin, ER negativity, MIB1 positivity, MUC1 cytoplasmic staining, and ERBB2 positivity. The probability that a tumor with the complete protein signature was an IBC was 91% but less than 50% when three or fewer parameters were present.

**Molecular Identification of New Breast Cancer Subclasses**—Gene expression studies and clustering methods have defined biologically and clinically relevant subclasses of breast cancer. In pioneering studies, global hierarchical clustering partitioned samples in two major groups, ER-positive and ER-negative, indicating that many genes display expression positively or negatively associated to ER status (5, 97–99). Perou et al. (100) refined this classification by making reference to the epithelial cell types. By confronting expression profiles of tumors and cell lines, they identified gene clusters associated to cell types (such as luminal and basal) and others specific to genomic alterations or signaling pathways (such as ERBB2 cluster). These connections were corroborated using IHC with antibodies against luminal and basal cell cytokeratins. Subsequent analyses refined this model with the subdivision of ER-positive tumors, resulting in the definition of at least five molecular subtypes (luminal A, luminal B, basal, ERBB2-overexpressing, and normal breastlike), which correlated with relevant histoclinical data (3, 10–12).

Six proteomic studies attempted to validate this classification on paraffin-embedded tumors at the protein level. Clustering was applied to IHC data of 7–31 proteins in large series of samples, ranging from 97 to 1,944 patients. Most of the markers tested had a well established or suspected role in breast cancer or represented proteins for which the transcript was a discriminator in previous gene profiling studies. ER, progesterone receptor, ERBB2, and p53 were present in all studies; CK5/6 was present in five of six; CK8/18, cyclin E, and Ki67 were present in four; and BCL2, cyclin D1, E-cadherin, and other ERBBs were present in three studies. Hierarchical clustering revealed concordant protein clusters and tumor clusters in agreement with RNA expression studies with clear distinction between ER-negative and ER-positive tumors and existence of subgroups in each category. The biological significance of these clusters was supported by their similarity with those derived from DNA microarrays studies, the confirmation of known correlations between proteins, and the presence of luminal and basal epithelial markers. Korsch-
Proteomics of Breast Cancer

100% specific for identifying the 21 basal tumors. Tested on a TMA containing 663 breast cancers with long follow-up, such a pattern delineated a group of tumors (15% of cases) associated with a much worse survival than the other two groups (ER-positive/ERBB2-negative and ERBB2-positive).

Prognostic Studies—Most studies on breast cancer prognosis using TMAs analyzed a single protein, whereas only two studies analyzed tens of proteins to identify a multiprotein signature. Examples of single tested protein with previously known or suspected relevance in breast cancer are multiple. In a series of 1,576 tumors (109), elevated expression of COX2 was associated with unfavorable outcome. Other proteins with negative prognostic impact included the catalytic subunit of telomerase (TERT) tested in 611 tumors (110), the epithelial cellular adhesion molecule (Ep-CAM/TACSTD1) in 1,715 tumors (111), or the early placenta insulin-like growth factor (EPIIL) in 603 tumors (112). Expression of KIT, a target of imatinib, was not associated with survival in a series of 1,654 tumors (113). Conversely proteins associated with favorable clinical outcome included fragile histidine triad (114), tyrosine-phosphorylated STAT5 (115), or BCL2 (116). Thirteen proteins were first tested in 930 tumors (116). BCL2 was the only independent prognostic factor in multivariate analysis combining IHC and the Nottingham Prognostic Index. BCL2 impact was then confirmed in an independent series of 1,961 tumors.

TMAs have also been used to validate and extend large series of sample results from gene profiling studies. One of the most prominent genes within the ER-related luminal gene cluster is the transcription factor GATA3. The favorable prognostic impact of GATA3 expression was confirmed at the protein level in 552 (104) and 139 tumors (117). Genes included in the basal cluster were also tested, and their negative prognostic role was confirmed in several hundreds of tumors: crystallin alpha B (118), cytokeratins 5 and 17 (119), and annexin A8 (120). Using DNA microarrays, we identified a gene expression signature associated with ERBB2 status (121). We then used TMAs (~250 tumors) to validate at the protein level the positive correlation of GATA4 and Ki67 with ERBB2. Similarly Rouzier et al. (9) confirmed on 122 independent tumors the correlation between low expression of the microtubule-associated protein TAU and sensitivity to neoadjuvant pacitaxel identified by gene profiling. Identification of a multiprotein signature of prognosis by supervised analysis, as reported at the RNA level with DNA microarrays, is more challenging with IHC data in terms of statistics. It has been reported so far in two studies. We developed a supervised method to search for the best combination within 26 tested proteins that would improve the prognostic classification of 552 tumors (104). We identified a 21-protein set (Fig. 1A) that optimally classified patients in two classes (“good prognosis” and “poor prognosis”) with different 5-year MFS. Initially identified in a learning set of 368 patients, this prognostic signature was validated in an independent set of 184 patients, showing its robustness. The classification based on the 21-protein predictor was associated with a difference in 5-year MFS (90 versus 61%; Fig. 1B). This difference remained significant when molecular grouping was done according to lymph node invasion, ER status, or type of adjuvant systemic therapy. Our predictor performed well in patients irrespective of ER status, suggesting it provides more accurate clinical information than ER status alone, possibly reflecting functional differences in the ER pathway or interacting pathways. In multivariate analysis, which included other prognostic factors and each protein separately, the 21-protein set was the strongest independent predictor of clinical outcome. This study was the first to apply both unsupervised and supervised analyses to quantitative IHC data. More recently, supervised analysis was applied to nine proteins and 324 stage I to III breast cancers treated with adjuvant tamoxifen (122). Univariate analysis showed the importance of conditional interpretation of certain markers on others due to their interdependency. In combination with the application of sophisticated informatics, it enabled the development of a predictive model that performed better than the current standards.

SELDI-TOF Studies

In breast cancer, SELDI-TOF MS was used to investigate serum/plasma (123–126) but also NAF (127–130) and tumor tissues (131) as a potential source for diagnostic biomarkers and recently as a potential tool to predict outcome (131, 132) and/or to monitor cancer treatment (133, 134).

SELDI and Diagnosis—Serum/plasma is the most easily available sample type for clinical proteomic studies. It may be a reliable surrogate tissue for a physiologic or pathologic process such as cancer. Some studies have addressed the potential of SELDI to provide serum biomarkers that differentiate breast cancer from benign disease and/or healthy subjects (123, 124, 126). Enrolling between 133 and 169 patients, these studies have identified diagnostic protein profiles with sensitivities and specificities of 76–93 and 90–93%, respectively. Protein peaks that distinguished healthy women from those with cancer were found at m/z 4,300 and 8,900 in two studies. However, no protein identification was provided. Another serum-based study with diagnostic purpose (125) analyzed 30 serum samples from BRCA1 mutation patients who either developed cancer (n = 15) or remained cancer-free (n = 15). It revealed 23 specific markers in BRCA1 cancer specimens that classified the two groups with 87% sensitivity and specificity. This study also included a comparison of serum profiles of BRCA1 mutation cancer developers and patients with sporadic breast cancers. Various proteins were up-regulated in the BRCA1 specimens and classified the two groups with 94% sensitivity and 100% specificity. Again no protein identity was reported, and the limited sample size precluded independent validation.

Other SELDI-based diagnostic studies used materials more
directly related to the tumor such as NAF or ductal lavage (128–130, 135–138). These studies, most of them with limited sample size, compared NAF or ductal lavage from healthy subjects (including benign breast disease for some) with unrelated patients with cancer or examined matched pair samples from women with unilateral primary invasive breast cancer using the contralateral (unaffected) breast as the “normal” sample. Considerable interpatient heterogeneity was reported, whereas similarity between matched NAF from cancerous and unaffected breast was high (128, 136). However, protein peaks with differential expression between cancerous versus healthy NAF were observed, some of them correlating
with axillary lymph node metastases and extent of the disease (135). One of the largest studies compared NAF from 27 cancer patients with 87 control women and differentiated invasive cancer not only from healthy subjects but also from ductal carcinoma in situ and benign breast disease (138).

**SELDI and Prognosis**—Whereas a large number of SELDI-based diagnostic studies were rapidly reported, little is known about the potential of this technique for prognostic evaluation. Using SELDI-TOF MS, we have reported the first evidence that serum protein profiling can provide prognosis information in EBC. We analyzed postoperative early serum from a retrospective series of 81 EBC patients receiving adjuvant anthracycline-based chemotherapy (132). Sera collected after surgery and before any specific adjuvant treatment were fractionated by combining chromatographic anion exchange beads and retention on chromatographic protein chips and analyzed by TOF-based mass spectrometry. Proteins differentially expressed (Fig. 2A) according to the metastatic outcome were selected and subjected to statistical analysis combining supervised partial least squares projection and a logistic regression model. A 40-protein index was generated that correctly predicted the clinical outcome in 83% of patients and was validated using leave-one-out cross-validation. This protein index identified two classes of patients with a difference in 5-year MFS and overall survival (Fig. 2B). It was the strongest independent prognostic factor in multivariate analysis, which included the clinicopathological factors with prognostic impact in univariate analysis, such as lymph node invasion, pathological tumor size, grade, and hormonal receptivity. Some of the components of the protein index were identified: haptoglobin α-1 and transferrin levels were up-regulated, whereas C3a complement fraction was down-regulated in serum from patients with high probability of metastatic relapse. Other identified proteins included apolipoproteins A-I and C-I. All these proteins are non-tumor-specific and are most certainly derived from host response. However, some of them are involved in biological processes such as angiogenesis and immune response, which may influence metastasis outcome. Although still preliminary and requiring validation on an independent dataset, these results provide incentive to further explore SELDI-based serum proteomics as a prognostic and/or predictive tool.

Another study has recently investigated the ability of SELDI technology to address prognostic issues (131). This study profiled tumor cytosolic extracts from 60 patients with node-negative EBC (most of them without adjuvant chemotherapy), 30 of them displaying a metastatic relapse and 30 of them being disease-free after the follow-up period. Differentially expressed proteins according to outcome were examined, and two protein peaks, proved by univariate Cox regression to have a good cumulative prognostic-informative power, improved the conventional Saint-Gallen prognostic classifica-
in tumor tissues by Western blotting analyses and IHC. This up-regulation was confirmed using ELISA and ferritin light chain; their up-regulation was validated, refined, and extended. Several obstacles have to be addressed to allow proteomics to reach its optimal yield and to translate into tangible benefits for patients. The accessibility, relatively limited to some academic facilities or companies during the past years, is increasing rapidly, notably for TMAs, and commercial TMAs are now available. The availability, complexity, and cost of other proteomic tools will likely improve with the next generation of tools. Because of the complexity and high throughput of experiments, exquisite quality controls are required, including controlled experimental conditions, signal-to-noise ratio, dynamic and linear range, sensitivity, and reproducibility. With regard to SELDI, identification of the discriminator protein(s) may be problematic. A major issue is the handling of data for multiprotein experiments. Collaborations being established between researchers across disciplines (physicians, biologists, mathematicians, statisticians, and computer scientists) to develop adequate tools required for producing, storing, analyzing, visualizing, and interpreting data. Previous works on gene expression profiling data constitute precious and useful bases for progresses, notably for statistical analyses.

Another field of interest for SELDI-based profiling of human breast cancer samples may be the prediction of therapeutic response to anticancer agents when applied to more advanced tumors. Compared with early diagnosis-oriented studies, this approach (which can be named “theragnostics”) may be more easily successful. Large tumor burden is frequently present in this setting, increasing the theoretical concentration of putative tumor markers that could regulate drug activity and therefore the probability of their detection in body fluids. In our institution, we examine SELDI-generated serum and tumor profiles from metastatic breast cancer patients received specific anticancer treatment (trastuzumab-based in ERBB2-positive patients and docetaxel-based in ERBB2-negative patients). Discriminator proteins between responder and non-responder patients might be identified.

**PERSPECTIVES**

Altogether these preliminary studies have revealed the potential of proteomics in clinical issues of breast cancer both to validate results from high throughput molecular studies and to define new potential diagnostic, prognostic, and predictive protein signatures. Results obtained on relatively small number of samples for non-TMA-based studies must now be validated, refined, and extended. Several obstacles have to be addressed to allow proteomics to reach its optimal yield and to translate into tangible benefits for patients. The accessibility, relatively limited to some academic facilities or companies during the past years, is increasing rapidly, notably for TMAs, and commercial TMAs are now available. The availability, complexity, and cost of other proteomic tools will likely improve with the next generation of tools. Because of the complexity and high throughput of experiments, exquisite quality controls are required, including controlled experimental conditions, signal-to-noise ratio, dynamic and linear range, sensitivity, and reproducibility. With regard to SELDI, identification of the discriminator protein(s) may be problematic. A major issue is the handling of data for multiprotein experiments. Collaborations being established between researchers across disciplines (physicians, biologists, mathematicians, statisticians, and computer scientists) to develop adequate tools required for producing, storing, analyzing, visualizing, and interpreting data. Previous works on gene expression profiling data constitute precious and useful bases for progresses, notably for statistical analyses.
Proteomics of Breast Cancer

the most attractive material that could be analyzed by proteomics in the near future would be cancer stem cells. Breast cancer stem cells have been identified (43). They are thought to fuel the tumor mass and should be the cells to target by specific therapy.

Finally once a potential marker has been identified by proteomic analysis and before any clinical application, it needs to be validated in a large series of samples. In this context, IHC on TMA is the most interesting approach provided that the appropriate antibody is available or can be made. It is too early to define the format of proteomic tool that will be applied in clinical practice. One example includes protein arrays that could detect, in the tumor and among multiple kinases, the activation of kinases to be targeted by specific therapies.

CONCLUSION

Proteomics of breast cancer is still in its infancy, but there is no doubt that it has a bright future for several reasons. First, it allows identification of targets for specific therapy. Proteins are excellent targets when mutated, truncated, or expressed as specific isoforms or in large amount (e.g. after gene amplification). Cancer isoforms are currently found after genomic studies rather than protein studies (e.g. DNA sequencing and molecular cytogenetics), but this could change with the advent of new techniques to study tumor samples and fluids. Second, proteins are good markers. Tumors and patient fluids may express specific protein isoforms that may serve as beacons for early detection and/or follow-up. Third, proteomics may be easily coupled with functional tests. Antibody arrays or cell arrays are good examples of this. Fourth, in contrast to other levels of information, proteomics may be applied to components of the cell, such as nucleus, membrane, and organelles, which may deserve specific attention or be more convenient to study in some cases. Even if, as mentioned above, a number of issues remain to be addressed before the proteomics can enter clinical practice, it is anticipated that in the future such an approach will affect the current generation of physicians in charge of breast cancer patients, hopefully transforming disease management into a more structured and logical science and a successful medicine. The characterization of discriminator proteins will certainly provide new markers useful for screening, diagnosis, prognosis, and monitoring and will help in developing new molecularly targeted anticancer drugs.

* This work was supported by Institut Paoli-Calmettes, INSERM, Université de la Méditerranée, and grants from the Ministries of Health and Research (Cancéropôle AC2I2004 and AC2I2005) and Ligue Nationale contre le Cancer (label). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

To whom correspondence should be addressed: Dépt. d’Oncologie Moléculaire, Inst. Paoli-Calmettes, 323 Bd. Sainte-Marguerite, 13009 Marseille, France. Tel.: 33-4-91-22-35-37; Fax: 33-4-91-22-36-70; E-mail: bertuccif@marseille.fncicc.fr.

REFERENCES


Proteomics of Breast Cancer

mase using automated quantitative analysis (AQUA): a novel paradigm for automated and continuous biomarker measurements. Am. J. Pathol. 164, 831–840


Proteomic analysis reveals that 14-3-3-σ is down-regulated in human breast cancer cells. Cancer Res. 61, 76–80


Proteomics of Breast Cancer


positive breast cancers are infrequent and lack KIT gene mutations.


*Int. J. Cancer* **107**, 854–862


