Proteomic strategies for the discovery of novel diagnostic and therapeutic targets for infectious diseases

Moushimi Amaya1, Alan Baer1, Kelsey Voss1, Catherine Campbell1, Claudius Mueller2, Charles Bailey1, Kylene Kehn-Hall1, Emanuel Petricoin III2 & Aarthi Narayanan1

1 National Center for Biodefense and Infectious Diseases, George Mason University, Manassas, VA, USA
2 Center for Applied Proteomics and Personalized Medicine, George Mason University, Manassas, VA, USA

Proteomic approaches can be a useful means of identifying targets for identifying and treating a variety of infectious agents. This is especially true for agents responsible for emerging infections or those important to biodefense.

Keywords
reverse-phase protein microarray; mass spectrometry; protein–protein interactions.

Abstract
Viruses have developed numerous and elegant strategies to manipulate the host cell machinery to establish a productive infectious cycle. The interaction of viral proteins with host proteins plays an important role in infection and pathogenesis, often bypassing traditional host defenses such as the interferon response and apoptosis. Host-viral protein interactions can be studied using a variety of proteomic approaches ranging from genetic and biochemical to large-scale high-throughput technologies. Protein interactions between host and viral proteins are greatly influenced by host signal transduction pathways. In this review, we will focus on comparing proteomic information obtained through differing technologies and how their integration can be used to determine the functional aspect of the host response to infection. We will briefly review and evaluate techniques employed to elucidate viral–host interactions with a primary focus on Protein Microarrays (PMA) and Mass Spectrometry (MS) as potential tools in the discovery of novel therapeutic targets. As many potential molecular markers and targets are proteins, proteomic profiling is expected to yield both clearer and more direct answers to functional and pharmacologic questions.

Introduction
Viruses have evolved an arsenal of mechanisms for infecting their host and to establish a productive infectious cycle. Hosts in turn have developed innate cellular defenses to block infection and replication of these viruses. It has been well established that the introduction of a virus into a host cell alters the host's proteome resulting in novel protein interactions. The novel interactions may not only occur between integral host components, but may also involve viral proteins. A comprehensive understanding of the molecular mechanisms underlying these protein–protein interactions (PPIs) would therefore enable the utilization of these mechanisms for the discovery of antiviral drugs and the identification of virally altered host proteins (Zheng et al., 2011; Kshirsagar et al., 2013; Mancone et al., 2013).

“Omics” technologies have advanced tremendously over the last few years, with genomics and transcriptomics pushing the gene expression field ahead (Braun and Gingras 2012; Mancone et al., 2013; Noisakran et al., 2008; Keating et al., 2013). However, gene expression profile datasets are limited in that they do not necessarily correlate with steady state protein products and are often not truly representative of the interactome (Graves & Haystead, 2002; Noisakran et al., 2008; Mancone et al., 2013). For the purposes of this review, we consider host proteins that interact with one another and with viral proteins forming distinct complexes as an interactome. Furthermore, genomics and transcriptomics studies alone cannot elucidate PPIs on a structural and functional level (Mancone et al., 2013).

PPIs can be studied using a variety of proteomic approaches ranging from biochemical and genetic to large-scale high-throughput technologies (Zheng et al., 2011; Kshirsagar et al., 2013). Here, we will review and evaluate techniques employed to elucidate PPIs with a
primary focus on Protein Microarrays (PMA) and MS as potential tools in the discovery of novel therapeutic strategies.

**Traditional approaches**

There are multiple traditional molecular biological and biochemical approaches that have been used for many decades to successfully detect PPIs. However, discussion of all of the methods is out of the scope of this review. In this review, we will include contextual information on the most popular classical methods employed for analyzing PPIs which include yeast two-hybrid systems (Y2H) and co-immunoprecipitations (co-IP) (Drewes & Bouwmeester, 2003; Miernyk & Thelen, 2008; White & Howley, 2013). Together, these two methods were responsible for deciphering over 70% of interactions in the InAct database as of May 2008 (Aranda et al., 2010; Chen et al., 2010).

The Y2H system is popular for the detection of binary physical interactions (Drewes & Bouwmeester, 2003; Hsu & Spindler, 2012; White & Howley, 2013). This method relies on the interaction between a ‘bait’ fusion protein and a ‘prey’ fusion protein to activate reporter genes or selectable markers. The advantages to this system are that it is economical and can be used to process large sample sizes (Drewes & Bouwmeester, 2003; White & Howley, 2013). For example, a pooled library screen approach can be used where a library of known ‘prey’ clones are combined and tested as pools against ‘bait’ strains. An example of determining PPI by the Y2H system was demonstrated by the influenza virus NS1 protein and human Stau proteins in vivo (Falcón et al., 1999; Cho et al., 2013), where coupling genome wide expression profiling with the Y2H system suggested a large number of host-viral protein interactions (Shapira et al., 2009; Tafforeau et al., 2011). Additionally, this system was used to identify: nine putative host cell proteins interacting with NSm of Rift Valley Fever Virus (RVFV); Vaccinia virus interactions with a variety of proteins (McCrailth et al., 2000); Severe Acute Respiratory Syndrome-CoronaVirus (SARS-CoV) proteins self-interacting and/or interacting with other viral proteins to form multimeric complexes (Von Brunn et al., 2007); and numerous cellular proteins binding to Dengue virus (DENV) proteins (Khadka et al., 2011). Unfortunately this procedure is also limited in that interactions between more than two proteins cannot be detected (Drewes & Bouwmeester, 2003; White & Howley, 2013) and the PPIs are observed outside of their natural context and will always require independent validation in relevant model systems.

The analysis of PPIs traditionally involves binding systems, where the protein of interest is tagged for recovery in methods such as co-immunoprecipitations (co-IP) (Miernyk & Thelen, 2008; Noisakran et al., 2008). Co-IPs rely on the idea that a fairly strong interaction between two proteins is stable under stringent elution conditions (such as high salt). When one partner of such an interaction is precipitated, the interacting protein is also precipitated and can be detected by antibody-based methods. Co-IP usually uses a sepharose-protein A column. Briefly, an antibody specific to the protein of interest is incubated with prepared protein samples to form an antibody-protein complex, which is then analyzed by SDS-PAGE and Western blotting, or further processed for more modern approaches including Matrix Assisted Laser Desorption/Ionization – Time of Flight Mass Spectrometry (MALDI-TOF MS) (Drewes & Bouwmeester, 2003; Miernyk & Thelen, 2008; Zheng et al., 2011). As examples, the use of co-IP has allowed for the identification of host–viral PPIs such as: DENV nonstructural protein 1 and human heterogeneous nuclear ribonucleoprotein C1/C2 (Noisakran et al., 2008); the methyltransferase domain of West Nile Virus NS5 protein and mammalian cGMP-dependant protein kinase, protein kinase G (Keating et al., 2013); and Vaccinia virus p37 with the host Rab9 and tail interacting protein of 47 kDa (TIP47) (Chen et al., 2009). Additional examples of co-IP studies include the Venezuelan Equine Encephalitis Virus, nonstructural protein 2 with the host’s major ribosomal phosphoprotein Pp66 (Montgomery et al., 2006) and Ebola virus VP35 protein with the host inhibitor of xB kinase ε (IKKε) and TANK-binding kinase 1 (TBK-1) (Prins et al., 2009). While valuable information has been determined through co-IP studies, the major drawbacks to this technique are that it may not be an accurate representation of an in vivo scenario, and weak or transient PPIs may be overlooked (Hsu & Spindler, 2012).

Additional methods are emerging to analyze PPIs that utilize a new generation of aptamers that contain chemically modified nucleotides. Aptamers are short single-stranded oligonucleotides that bind with high affinity and specificity to proteins, peptides, and small molecules (Gold, 1995; Brody & Gold, 2000). One method uses Systematic Evolution of Ligands by EXponential enrichment (SELEX) to select aptamers from libraries of randomized sequences (Ellington & Szostak, 1990; Gold, 1995). Although this technology has high potential for high-throughput biomarker identification, there have been some difficulties creating high affinity aptamers for some protein targets (Gold, 1995). In addressing this problem, a new form of aptamers has been developed, called Slow Off-rate Modified Adapters (SOMA-mers). The basis for SOMA-mers is that the addition of functional groups can give aptamers protein-like properties that enable a wider variety of high affinity aptamers (Gold et al., 2010). The aptamer-based technology wave has advantages, but requires further characterization, validation and standardization with standards such as affinity capture-MS (Gold et al., 2010).

Due to the limitations of traditional methods, improved approaches for assessing PPIs are in demand, which can not only increase our capacity to analyze broad ranges of interactions in a systems biology approach, but also function synergistically with traditional methods. PPIs cannot be observed in isolation as binary interactions, which strengthen the need for new approaches that would encompass both the ability to identify PPIs but also relevant biological pathway information. More recently, large-scale ‘precision proteomics’ based on MS and microarrays have enabled the system-wide characterization of host-based events at the levels of post-translational modifications, PPIs and changes in protein expression. This advancement
delivers accurate and unbiased quantitative information regarding protein modifications in response to any perturbation. As large-scale proteomics-based signaling research continues to develop and integrate existing and novel technologies alongside improved databases, it is likely that our understanding of signaling networks will undergo significant change in the coming years.

Protein microarrays

PMA have been used to profile PPIs leading to the discovery of possible drug targets at the molecular level (Wulfkuhle et al., 2004; Gulmann et al., 2006; Wilson et al., 2010; Mancone et al., 2013). The basis of a protein microarray is an array of immobilized protein spots, arranged on a slide so that each spot contains either a homogenous or heterogeneous set of bait molecules (Liotta et al., 2003). Currently there are two popular types of PMAs: traditional or forward phase microarrays (FPMA) and reverse phase protein microarrays (RPMA; Liotta et al., 2003; Sheehan et al., 2005; Wilson et al., 2010; Zhu et al., 2012; Sutandy et al., 2013). In FPMA, bait molecules, usually antibodies are immobilized on a glass slide and a cell lysate sample (the antigens) is placed on the array. A layer of signal-generating antibodies is then added to allow for binding. A signal is emitted only at positive spots on the array (Fig. 1a). In contrast, RPMA methodology immobilizes the test sample analytes (the antigens) on the slide, and then antibodies are applied as the second mobile phase. The RPMA format immobilizes a different test sample in each spot, meaning that one array can be comprised of hundreds to thousands of different samples. The array is incubated with a signal-generating antibody to obtain positive signals (Fig. 1b). Each of these arrays will be described further below.

Forward phase protein microarrays

FPMA is a protein microarray that is designed by fixing and immobilizing multiple bait molecules on a glass slide surface, where the baits are usually antibodies (Fig. 1a), but can also be nucleic acids, peptides and phage lysates (Liotta et al., 2003; Wulfkuhle et al., 2004; Sheehan et al., 2005; Gulmann et al., 2006; Wilson et al., 2010; Hu et al., 2011). FPMA allows the simultaneous measurement of multiple proteins from a single sample (Liotta et al., 2003; Sheehan et al., 2005; Gulmann et al., 2006; Wilson et al., 2010; Hu et al., 2011). The analytes from either serum or tissue samples are prelabeled, often with fluorophores such as Cy3 or Cy5, is placed over the array. A layer of signal-generating antibodies is added to the analytes to allow binding such that a signal will be emitted only at positive spots on the array (Liotta et al., 2003; Gulmann et al., 2006; Wilson et al., 2010; Hu et al., 2011).

Studies using FPMA have been used in the study of a wide range of topics such as, determination of proteomic profile of endogenous proteins in LoVo colon carcinoma cells exposed to ionizing radiation to identify regulatory sites for radiation-induced apoptosis signaling (Sreekumar et al., 2001). Additional applications of FPMA involved profiling protein tyrosine phosphorylation and characterizing changes in post-translational modifications, such as acetylation and ubiquitination in mammalian cells (Ivanov et al., 2004). Other studies have developed a proteomic profile of cancer progression in oral cavity cancer tissue samples where a combination of FPMA was employed with Laser Capture Microdissection (LCM) (Knezevic et al., 2001). LCM enables researchers to isolate cells of interest from heterogeneous tissue (with multiple kinds of cells) and arrive at a homogeneous cell population without contamination from surrounding nonrelevant cells (Espina et al., 2006). For
additional technical information on LCM, the reader is
directed to Espina et al. (2006). The data obtained from
these studies illustrate just a few examples of applying
FPMA as a high-throughput tool to profile a variety of
post-translational modifications in cells under different
treatments.

The two major limitations for this procedure are as follows:
fluorophore labels can potentially disrupt binding capabilities
of the analyte, and labeling all of the proteins in the sample
can result in a high background due to non-specific binding.
This is because analytes in a complex mixture may be
present in low abundance, alongside an excess of other
proteins. Performing serial dilutions of the samples is often
necessary to dilute protein concentrations to alleviate high
background issues (Gulmann et al., 2006). An additional
limitation is that FPMA requires specific antibodies to
targeted proteins, with differences in antibody design further
introducing potential variability when comparing indepen-
dent experiments (Gulmann et al., 2006).

Reverse phase protein microarrays

RPMA is a protein microarray that immobilizes the test
sample analytes on the slide, and then antibodies are
applied on top of those analytes (Fig. 1b, compared to
FPMA in Fig. 1a). RPMA is a newer iteration of traditional
FPMA technology developed in 2001 for the cancer field by
Drs. Lance Liotta and Emanuel Petricoin III, where cell
populations from samples taken from cancer tissue were
subjected to LCM and used to analyze the state of
pro-survival checkpoints and growth regulation proteins
(Wulfkuhle et al., 2004; Paweletz et al. 2001). Since this
publication, RPMA technology has been adopted by many
research groups and applied not only to LCM tissue, but
also to heterogenous tissue samples, cell culture lysates
(Nishizuka et al., 2003), and serum/plasma samples (Janzi
et al., 2005; Mueller et al., 2010), ovarian effusions (Dav-
sdon et al., 2006), vitreous (Davuluri et al., 2009), fine needle
aspirates (Rapkiewicz et al., 2007), and peptides (MacBe-
ath & Schreiber, 2000). Moreover, the use of RPMA has not
been limited to pre-clinical research studies, but has been a
critical component of several clinical trials [for review see
Mueller et al. (2010)], cancers (VanMeter et al., 2008;
Wilson et al., 2010; Einspahr et al., 2012), cellular pathway
characterizations, bacterial infection mechanisms (Popova
et al., 2009), immunological disorders and viral-host inter-
actions (Wilson et al., 2010; Einspahr et al., 2012; Agu-
ilar-Mahecha et al., 2009; Narayan et al., 2012; Baer
et al., 2012; Austin et al., 2012; Popova et al., 2010). RPMA
is capable of monitoring protein dynamics as a function of
time, in diseased vs. nondiseased states before, during, and
after treatments (Sheehan et al., 2005; Spurrier et al., 2008)
and to quantitatively monitor protein expression levels of
many samples simultaneously (Liotta et al., 2003; Wulfkuhle
et al., 2004; Sheehan et al., 2005; Gulmann et al., 2006;
Hultschig et al., 2006; Spurrier et al., 2008; Wilson et al.,
2010; Hu et al., 2011; Sutandy et al., 2013).

RPMA technology provides the necessary analytical
sensitivity to measure even ultra-low abundant proteins.

For this reason, the combination of LCM technology with
RPMA creates a powerful analytical tool based on its
increased sensitivity, reduced sample volume/concentration
input requirements and combination of hundreds to thou-
sands of samples on a single array. Moreover, because the
RPMA is not a two-site sandwich-type antibody array and
requires only one well-performing primary antibody for
detection, the number of analytes that can be measured is
much greater than what could be performed on a for-
ward-phase array. Typically, with the RPMA nanoliters (nl)
of each sample are printed in duplicate or triplicate onto
glass-backed nitrocellulose slides using either contact
(Sheehan et al., 2005; Gulmann et al., 2006; Lübke et al.,
2008) or non-contact (Schena, 2000) arrays. The source of
these analytes could be in vitro samples generated from
cultured cells or in vivo samples generated from animal
tissue or clinical samples. With a spot size of only tens to a
few hundred microns in diameter, the RPMA platform allows
for thousands of samples to fit on one slide and hundreds of
slides can be analyzed in every array run giving this
approach a high throughput flavor. Each array is then
probed with a single primary antibody, in principle similar to
other immunoassays.

The biggest challenge for RPMA is the same as for any
immunoassay: the need and availability of high quality,
specific antibodies. Prior to application to RPMA, antibodies
have to be validated using Western blotting to demonstrate
high target specificity. Currently, the development of vali-
dated antibody libraries is an individual effort by each lab. In
George Mason University, we currently possess a large
repertoire of more than 400 validated antibodies relevant to
phosphorylated and unphosphorylated proteins which map
to diverse nodes in many phospho-signaling cascades. It will
be beneficial for the protein microarray field to combine such
efforts in the future and assemble and maintain a central
repository of validated and “RPMA-certified” antibodies.

Diagnostics and current applications of RPMA

to describe the workflow for RPMA (Fig. 2), we will follow
an example of a RPMA study conducted by Popova et al. (2010).
In this study, Popova et al. used RPMA to identify
phospho-signaling protein pathways modulated in cultured
human cells at very early (of the order of a few minutes) to
late time points after RVFV infection. Human cells were
infected with the wildtype strain of RVFV, ZH-501 and cells
were lysed at various time points. For the RPMA procedure,
c. 30 nL of each sample (equivalent to the amount of
material from 40 lysed cells) was arrayed on nitrocellulose
slides. An advantage of the array layout is that each sample is
printed in a 4–8 point dilution series which ensures that the
quantification of positive signals can be performed in the
linear range. Each slide was then probed with one of 60
different antibodies specific against phosphorylated or total
forms of signaling proteins. The antibodies were selected to
monitor the molecular networks involved in host responses
most likely affected by virus exposure such as apoptotic and
cell survival pathways. To validate the RPMA data, the
levels of multiple selected proteins were analyzed by
Western blots using antibodies against phosphorylated and total forms of proteins; both procedures were in agreement with each other. This demonstrated that while RPMA has enormous technical utility independently, it can also function in a synergistic capacity with traditional methods.

Following imaging of arrays, the software technology used to capture and quantify the analyte spots is similar to software used for DNA microarray analysis [i.e. IMAGEQUANT (GE Healthcare Life Sciences, Pittsburgh, PA) or MICROVIGENE™ (VigeneTech, North Billerica, MA)]. Some software packages, such as MICROVIGENE, include several features to facilitate quantification of array data: automated spot-boundary detection with image contrast enhancement, dust/scratch removal, outlier flagging, regional background correction and total protein, negative controls and internal reference standard normalization options (Gulmann et al., 2006). Data normalization algorithms range from calibration curve normalization (Sevecka & MacBeath, 2006), spiked-in internal standard normalization (Korf et al., 2008), Robust-Linear-Model normalization (Sboner et al., 2009), to normalization of respective sample total protein. In the example of Popova et al. (2010), the average total level of cellular protein at every time point was determined by staining with Sypro Ruby Protein Blot Stain of four randomly selected slides throughout the print run (Popova et al., 2010). Spot finding software programs convert pixel density for each spot into numerical values (Mueller et al., 2010). The MICROVIGENE software offers automated curve-fitting approaches to quantify the analyte across a five-point dilution curve, including replicates (Gulmann et al., 2006). For additional technical information about the software the reader is directed to Gulmann et al. (2006).

The downstream biostatistical analysis of the raw, normalized array data depends highly on the type of samples printed on the array and the scientific questions that are being addressed. In general, many of the statistical methods employed in RPMA data analysis are traditional methods used to evaluate confidence intervals. Methods such as unsupervised hierarchical clustering, k-means clustering, self-organizing maps and principal component analysis are...
tools used to determine a relationship between the data points generated from the microarray (Gulmann et al., 2006). Clustering of closely related data points can be determined with a cut-off point that makes the data biologically sensible; however, statistically significant differences between data points cannot be provided with hierarchical clustering methods (Gulmann et al., 2006). Other common statistical methods such as, ANOVA, t-tests and Mann-Whitney U-test can be employed to determine differences between data sets provided prior knowledge of the classification of the samples is known (Gulmann et al., 2006).

Several technologies exist to graphically map protein signaling pathway activation data from RPMA. One approach is to overlay the RPMA data over a static image of PPIs that are generally accepted by the larger scientific community, such as the well-known “Pathways in Human Cancer” cell signaling map by Dr. Robert Weinberg. Cscape uses the Google Maps API to overlay this cell signaling map with relative protein/phosphoprotein abundances, enabling the visual identification of “hot spots” within the cell signaling architecture or whole activated/nonactivated signaling pathways (Fig. 3). Another approach is to visualize protein abundance and protein–protein correlations (i.e., as calculated by Spearman’s Rho analysis) in a Bayesian network like computational analysis using general network analysis systems such as Gephi (Bastian et al., 2009). The graphs generated with this application graphically illustrate influences of pathway components on each other in the form of nodes and arcs, where nodes represent the variables and the arcs statistically significant relationships between the variables.

One of the strongest advantages of RPMA is its ability to generate multiple hypotheses for subsequent testing and development of the field. The outcome of each RPMA run is a virtual treasure trove of information on multiple signaling events analyzed as a single snapshot in time. Each one of those altered signal events will form a basis of an individual hypothesis. If this involved specific viral strains or individual viral proteins, for example, we can go to great lengths in deciphering functional consequences of PPIs. This was exemplified by a RPMA study conducted by Popova et al. (2012) that revealed phosphorylation and activation of the NF-kB signaling cascade. Narayanan et al. (2012) demonstrated in a follow up study that phosphorylation of p65 (serine 536) involved phosphorylation of IxB through the classical NF-kB cascade and that RVFV utilized the host NF-kB signaling cascade to establish a robust infection. In addition, this group demonstrated that inhibition of the NF-kB cascade was able to inhibit viral replication (Narayanan et al., 2012). The more important outcome of that study was that the viral protein NSs may have an influence on the activation status of the cascade. The same RPMA study also revealed that the tumor suppressor protein, p53 was phosphorylated in the event of ZH501 infection (Austin et al., 2012). A study by Austin et al. (2012) expanded on this data to show that in RVFV infections, p53 was activated, phosphorylated and localized in the nucleus following transient interaction with the RVFV virulence factor NSs. Further characterization of p53 phosphorylation prompted investigation into the DNA damage signaling pathway during RVFV infection. Baer et al. (2012) demonstrated that RVFV infection induced an NSs-dependent DNA damage signaling response and concurrent S-phase arrest, providing evidence for a novel function of NSs that was shown to directly impact viral replication. Collectively, modifications in phospho-protein signaling in RVFV examined by a single RPMA study resulted in the elucidation of numerous cellular mechanisms that played important roles during viral infection and that were subsequently exploited in targeting viral replication. The results drawn from these studies illustrate how the high throughput identification of PPIs between the host and the virus can lead to the identification of potential therapeutic targets during viral infection (Baer et al., 2012).

In 2009 Popova et al., utilized RPMA to characterize and measure the innate cell signaling responses of lung epithelial cells to Bacillus anthracis infection (Popova et al., 2009). The authors demonstrated RPMA that B. anthracis infection

![Fig. 3 Development of Protein Interaction Networks](https://academic.oup.com/femspd/article-abstract/71/2/177/2911565/17291855)
RPMA technology in a clinical environment to detect biomarkers. This example demonstrates the feasibility of using RPMA technology to create a rapid, accurate and sensitive diagnostic tool to detect biomarkers for SARS-CoV. To address this problem, Zhu et al. (2006) developed the first coronavirus protein RPMA. The entire proteome of the human SARS-CoV, HCoV-229E virus and partial proteomes of other coronaviruses were imprinted on the microarray and probed with symptomatic patient sera (Zhu et al., 2006). The presence of viral-specific antibodies was detected using Cy3-labeled goat anti-human IgG antibodies (Zhu et al., 2006). Analysis of the microarray revealed increased sensitivity, reactivity and accuracy for biomarker detection in SARS infected individuals (Zhu et al., 2006). This example demonstrates the feasibility of using RPMA technology in a clinical environment to detect biomarkers.

**Mass spectrometry**

While the advent of novel and high throughput techniques such as RPMA have greatly expanded on the ability to generate informative datasets, complementary approaches are necessary to account for the dependence on antibodies. Additionally, antibodies have to be selected by the researcher and are inherently directed to a small number of known signaling events and modifications, giving rise for the need to complement antibody validation using a technique that is independent of these limitations. To get around the inherent limitations and bias of antibody-based research, modern quantitative MS is increasingly being used to balance and direct antibody-based approaches (De Chassey et al., 2012; Zhou et al., 2011; Patwa et al., 2010; Patwa et al., 2009). MS-based proteomics is not limited to specific sites or proteins of interest which makes it suitable for use in an unbiased (hypothesis free) and systems-wide manner, representing a fundamentally different approach to studying cell signaling (Go et al., 2006; Zheng et al., 2011). One commonly used MS approach depends on peptide separation normally performed by liquid chromatography (LC) tandem MS or LC-MS-MS (Gonzalez-Galarza et al., 2012).

When looking at host cell signaling responses, complex protein samples may be derived from whole cells or from biological fluids, and while all of the proteins in an entire sample may be processed as is, it is often preferable to perform separation techniques to better characterize information on proteins of particular biological interest (Brewis & Brennan, 2010). Protein separation is initially performed prior to this step, by methods such as one dimensional (1D) or two dimensional (2D) gel electrophoresis, column chromatography, or affinity purification, followed by enzymatic digestion (often using trypsin). Sample preparation is critical to subsequent data interpretation, particularly when dealing with large datasets such as those generated by LC-MS. Techniques such as subcellular fractionation can provide cleaner samples while also giving valuable protein localization information compared with that from whole cell lysates alone (Brewis & Brennan, 2010; Gonzalez-Galarza et al., 2012). With biological fluids, it is also possible to remove the most abundant proteins by immunodepletion or enrichment.

An alternative or complimentary method for global protein separation is to resolve proteins by 1D or 2D electrophoresis before subjecting individual protein bands to digestion and LC-MS (Brewis & Brennan, 2010). In both methods, many gel segment/protein band cuts can be processed to identify numerous separated proteins. In a typical workflow, an individually separated protein is physically isolated and removed, trypsin digested and the resulting peptides are separated on the basis of relative hydrophobicity through LC before tandem MS (MS/MS Brewis & Brennan, 2010). The tandem MS data are then used to search existing protein databases to achieve a matching protein spectra based on amino acid sequences typically derived from MS/MS spectra. Low abundance, very large, or very small proteins have, however, proved difficult to resolve using 2D gels, and for global analysis it is now much more commonplace to trypsin digest the entire solubilized protein mixture to produce a peptide “soup” of all the proteins in the sample (gel-free LC-MS proteomics) (Petricoin et al., 2002). Peptides can then be separated by LC on the basis of relative hydrophobicity and charge as a multidimensional separation (Petricoin et al., 2002). Extensive MS/MS and database searches can then be performed to identify many of the proteins in the original sample.

One of the advantages of this protein interactome mapping workflow is that it is possible to achieve quantitative data at the same time by introducing a peptide labeling step, such as isobaric Tags for Relative and Absolute Quantification (iTRAQ) method, using labeled peptides as internal standards, or to perform protein labeling, such as stable isotope labeling by amino acids in cell culture (SILAC), without the need for additional LC-MS (Brewis & Brennan, 2010). iTRAQ represents a method for differentiating between multiple conditions (up to eight samples) simultaneously (Wiese et al., 2007; Brewis & Gadella, 2010). This approach lends itself to the labeling of *in vivo* samples and/or primary cells as such labels are applied postsample preparation. As an example, iTRAQ has been used to investigate influenza virus infection in primary human macrophages (Lietzén et al., 2011). Typically, two sets of samples are differentially labeled with a stable isotope, one sample with the light form, the other with the heavy form of the label. Depending on the type of the chemical labeling reagent used, samples are labeled before, during, or after enzymatic digestion, combined then subjected to a separation/enrichment technique followed by analysis with MS for a quantitative comparison. Changes in protein expression levels are then quantified by determining the ratio of the peak intensities of the light and heavy forms of the generated peptides (Brewis & Brennan, 2010). The pairwise comparison of the peak intensities of peptides
labeled with the heavy and light form of the label serves as the basis for quantitative protein analysis (Brewis & Brennan, 2010). It should be noted that the accuracy of the quantitative measurement depends on when the label was incorporated, with the highest level of accuracy being achieved when the label is incorporated early on in the process due to sample loss and processing variability.

SILAC has also been applied to study viral infections in cell culture to provide quantitative information regarding pathogen–host cell interactions (Brewis & Gadella, 2010; Zhou et al., 2011; Toss et al., 2013). During SILAC, cells and viruses are differentiated by growing each cell population (or the virus within a cell population) in media containing unique stable isotope labeled amino acids that become incorporated into newly synthesized proteins, eventually supplanting their equivalent nonlabeled homologs (Brewis & Gadella, 2010). This labeling technique greatly reduces sample complexity and increases labeling efficiency prior to sample preparation, while allowing the relative quantification of proteins in samples by MS (Brewis & Gadella, 2010). When analyzed on a mass spectrometer, pairs of chemically identical peptides of different isotope composition can then be differentiated, as the labeled amino acids induce a shift in the mass/charge ratio (m/z) in comparison to the unlabeled peptides (Brewis & Brennan, 2010; Brewis & Gadella, 2010). By comparing the intensities of the labeled and unlabeled m/z peaks, it is possible to obtain accurate quantitative data on the relative abundance of labeled and unlabeled peptides present in the sample (Brewis & Brennan, 2010; Brewis & Gadella, 2010). In this way, proteins that are increased or decreased in abundance in virus-infected, compared with mock or drug-treated cells, can be simultaneously identified and quantified (Brewis & Brennan, 2010; Brewis & Gadella, 2010).

The relative quantification of peptides using iTRAQ or SILAC labeling, coupled to LC-MS/MS and bioinformatic analysis, is one of the most popular and powerful options for global proteomic quantification and has provided an excellent resource for studying host cell proteomes and is readily being applied in the study of host–pathogen infections (Wiese et al., 2007; Brewis & Brennan, 2010; Brewis & Gadella, 2010; Munday et al., 2012).

Biomarker identification

In the proteome biomarker discovery pathway, the aim is to combine multidimensional fractionation and labeling methods along with MS/MS analysis to identify proteins that are unique or highly abundant in complex samples obtained from specific disease states and comparing those profiles to healthy matched controls. Through the use of peptide labeling such as the previously mentioned iTRAQ or SILAC, samples from both disease-affected and healthy controls can be run side by side through an MS workflow with the data obtained then processed using bioinformatic algorithms and programs such as iTRACKER for iTRAQ that search for differences in peak intensities between the sample sets or SILACAnalyzer for SILAC which is an open source tool for the fully automated analysis of quantitative proteomics data (Petricoin et al., 2002; Wiese et al., 2007; Munday et al., 2012). SILACAnalyzer identifies pairs of isotopic envelopes with fixed m/z separation and requires no prior sequence identification of the peptides. The discriminating pathogenic pattern formed by the key subset of proteins or peptides buried among the entire repertoire of thousands of proteins represented in the sample spectrum, can then be compared to its control group and potentially identified.

A recent study of the Human respiratory syncytial virus (HRSV) using SILAC in conjunction with LC MS/MS allowed the direct and simultaneous identification and quantification of both cellular and viral proteins (Munday et al., 2010). To reduce sample complexity and increase data return on potential protein localization, cells were further fractionated into nuclear and cytoplasmic extracts (Munday et al., 2010). Novel HRSV-host cell interactions, including those associated with the antiviral response and alterations in subnuclear structures such as the nucleolus and ND10 (promyelocytic leukemia bodies) were identified (Munday et al., 2010). In addition, novel changes in mitochondrial proteins and functions, cell cycle regulatory molecules, nuclear pore complex proteins and nucleocytoplasmic trafficking proteins were observed in infected A549 cells (Munday et al., 2010). Commonly available bioinformatics programs such as Ingenuity Pathway Analysis were used in the organization, expansion and interrogation of the derived data sets from virus-infected cells which were then validated using tradition orthogonal assays (Munday et al., 2010; Lietzén et al., 2011). In this study, the use of SILAC in conjunction with LC MS/MS resulted in the identification and potential interaction of 1140 cellular proteins along with six viral proteins (Munday et al., 2010). In another study using Flock House Virus (FHV)-infected Drosophila cells, profile changes in the protein expression levels were found based on the direct analysis of intensities using labeling in combination with LC-MS/MS (Go et al., 2006). Overall a total of 1500 host proteins were identified and quantified, of which 150 were up-regulated while 66 were down-regulated in response to viral infection (Go et al., 2006).

While stable isotope labeling has traditionally been used in comparative host–viral proteomics, a label-free approach is becoming a viable and attractive alternative, through the use of chromatographic separation and high mass accuracy measurements along with data normalization methods. Moreover, informatics algorithms are employed to facilitate the data analysis due to the large datasets that can be obtained from LC-MS experiments. Overall, this is highly effective in comparative proteomics due to its comprehensiveness and high throughput nature. To date, label-free quantitative proteomics approaches have been successfully applied in the analysis of numerous samples and organisms such as: human serum, yeast, and Shewanella oneidensis (Fang et al., 2006; Brewis & Brennan, 2010). Despite these significant advances and efforts, challenges remain with current label-free methods such as those associated with the dynamic range of measurements and the extent of proteome coverage, confidence in peptide/protein identifications, quantization accuracy, analysis throughput, and the
robustness of present instrumentation are all issues that still need to be addressed and improved on before these technologies can be reliably used in a clinical setting (Qian et al., 2006; Findeisen & Neumaier, 2009).

There are many options available for quantitative, unbiased proteome studies, only some of which have been discussed in this review. New techniques utilizing traditional and novel next-generation targeted proteomics have started to shift the traditional paradigm from using discovery-based identification to targeted quantification, such as in the utilization of triple quadrupole mass spectrometers (QQQ; Boja & Rodriguez, 2012; Shi et al., 2012; Fung et al., 2013). With a discovery-based strategy, the goal is usually to identify as many proteins as possible, while the goal of a targeted proteomics is to monitor a select few proteins of interest with high sensitivity, reproducibility and quantitative accuracy. In the traditional discovery-based approach peptide ions are automatically selected in the mass spectrometer for fragmentation on the basis of their signal intensities, generating rich but complex tandem mass spectra for each peptide sequence, requiring careful analysis and complex bioinformatics. In contrast when using a targeted workflow, the QQQ mass spectrometer can be programmed to detect specific peptide ions derived from proteins of interest, and can select specific ‘precursor’ ions (on the basis of their m/z ratio) for fragmentation. In the second mass filter, target ions are selected and guided to the detector for quantification, resulting in a trace of signal intensity vs. retention time for each precursor ion–product ion pair. This process is called selected reaction monitoring (SRM) or multiple reaction monitoring (MRM; Lange et al., 2008). While this up-front approach is much more labor intensive to develop than a traditional discovery-based pipeline, once a reliable assay is generated for a specific protein, analysis of the MS data is relatively straightforward and uncomplicated in comparison. While SRM is the most mature MS-based technology for targeted proteomics analysis, new methodologies that obviate the need for laborious SRM assay optimization are currently being developed. One example of several novel methodologies is SWATH; complex mass spectra generated by data-independent acquisition (in which peptides are selected for fragmentation without regard to signal intensity) are queried for the presence of specific peptides using libraries of qualified peptide fragment spectra (Gillet et al., 2012).

For MS, novel approaches and developments continue to improve and push the field rapidly forward towards targeted approaches and provide many viable options when looking at proteome interactomes from a host–pathogen perspective. While every toolset discussed in this review has its strengths and weaknesses, combining MS along with other high throughput methods of protein detection such as RPMA, in conjunction with the use of traditional tools and means of validation such as Y2H screens or immunoblot analysis, have demonstrated the ability to greatly strengthen the inherent deficiencies of any individual approach. The use of these orthogonal methods in conjunction is thus able to arrive at a much more well-rounded and comprehensive picture of the proteome and its mechanisms when used in combination (Patwa et al., 2009; Mancone et al., 2013).

Informatics approaches for integrating proteomics data

As previously discussed, cellular function is driven by a complex network of interacting proteins that are controlled by various signaling cascades among other mechanisms. The proteomics experiments described here aim to paint a molecular portrait of the essential proteins activated in specific cellular states, such as comparing infected vs. uninfected states (Fig. 3). Following statistical analysis to identify proteins that are significantly changed between desired analytical states (described above), protein lists are generated that must be transformed into informative networks that are representative of the functional changes observed in the disease state.

There are several key analytical steps that must be undertaken to develop these comprehensive networks, including mapping protein IDs to function, cross-comparing protein function, and identifying, visualizing and analyzing protein interaction networks (Sanz-Pamplona et al., 2012). Additionally, protein function, structure and interaction databases are sparse and represent only a small fraction of known human proteins, with host–pathogen interaction databases even more sparse (Kshirsagar et al., 2013; Kshirsagar, Carbonell, and Klein-Seetharaman 2012). Trying to find commonalities among functions is therefore essential for generating the most complete networks possible. Tools such as Ingenuity Pathway Analysis (IPA) (Goulet et al. 2013), National Institute of Allergy and Infectious Diseases’ Database for Annotation, Visualization and Integrated Discovery (DAVID; Huang, Sherman, and Lempicki 2009), and Cytoscape (Sanz-Pamplona et al. 2012) among others can be used to identify common and statistically overrepresented pathways and ontological terms (functions) from protein lists which can then be used to interrogate broader protein interaction networks. These tools draw from large public databases that describe protein functions—such as the Gene Ontology (GO), pathways—such as KEGG, Reactome and Biocarta, and protein interaction databases—such as DIP, BIND, and IntAct among others. These tools can also be used to identify important pathways and interactions between host and pathogen proteins (Kshirsagar et al., 2012; Goulet et al. 2013).

Although MS experiments are comprehensive in their measurements of cellular proteins, issues described above such as variability in protein abundance, can confound experiments and result in results that cannot reproducibly identify individual proteins. This can confound the types of basic network analyses described in the previous paragraph. Several existing tools can help with assessing the similarities of identified gene functions and merging similar proteins into like functional groups, including SORA and simGIC (Teng et al. 2013). These tools can be used to merge protein expression information by protein function across different methods such as RPMA and tandem MS to help cluster proteomics data into functional bins which can then be used to identify pathways and interaction networks. SORA and simGIC, among others, measure the functional similarity among proteins in terms of information content in
the context of GO terms, which allows for more precise binning of proteins with similar functions and roles (Teng et al., 2013).

Networks generated from proteomics experiments are often highly complex and benefit greatly from further analysis with interactive visualization tools that can help to tease out key subnetworks for further analysis or laboratory validation experiments. Tools such as IPA (Goulet et al., 2013), Cytoscape (Kacprowski, Doncheva, and Albrecht 2013; Sanz-Pamplona et al. 2012) and CScape (Einspahr et al., 2012) allow for insightful visualizations and analysis of these complex cellular networks analysis that allow scientists to directly interact with their data and help to turn information into knowledge of key cellular functions affected by disease. For example, nodes and edges within these networks may be weighted based on common functionality, and/or directionality or magnitude of expression to identify key connections among experimentally identified proteins.

Conclusion

Proteins are currently the major drug targets of choice and play a critical role in the process of modern drug design. Host proteins are particularly attractive targets as pathogens are less likely to develop an obstructive mutation to the therapeutic and develop resistance. Therapeutic development typically involves numerous steps: the construction of drug compounds based on the structure of a specific drug target, validation for therapeutic efficacy of the drug compound, evaluation of drug toxicity, and finally clinical trial. RPMA and MS along with other functional proteomic approaches can be employed at all steps in the investigative process. The use of MS in combination with the use of high throughput antibody platforms will allow a much more global, comprehensive, and directed approach in the study of viral infection or any pathogenic state for that matter, revealing the complexity of the events within an infected cell while compensating for the limitations of a technique in isolation.

Post-translational modifications—including phosphorylation, acetylation, and ubiquitination—are specific modifications that can alter the activity of an individual protein target. The cumulative effect of these small modifications is the regulation of large signaling pathways and networks within cells. A comprehensive understanding of the molecular mechanisms underlying viral infection remains a major challenge in the discovery of new antiviral drugs and host susceptibility factors. New advances in the field are expected to arise from systems-level modeling and the integration of proteomic and genomic disciplines, with current wet lab techniques. Here, we attempted to briefly explore the importance and benefits of using two powerful proteomic techniques in a combinatorial capacity, to better understand the molecular relationships between viruses and their host, chiefly looking at cellular signaling pathways from a global perspective. Past studies have already demonstrated that viral proteomes target a wide range of functional and inter-connected modules of proteins within the human interactome and that using a global and systems-level approach can provide relevant, accurate and valuable information (Go et al., 2006; Brewis & Gadella, 2010; Munday et al., 2010; Lietzn et al., 2011; Zhou et al., 2011; De Chassey et al., 2012).

To summarize, protein interaction studies have come a long way in the past decade in terms of technology/platform development and protein chemistry. Together with the use of microarrays and bioinformatics tool sets, along with other “Omics”: the identification of the molecular signatures of diseases based on protein pathways and signaling cascades, holds great promise and utility for disease diagnosis and therapeutic development.

Acknowledgements

The authors would like to thank members of the Narayanan and Kehn-Hall laboratories and Center for Applied Proteomics and Molecular Medicine at GMU for help in manuscript preparation and revision. The authors declare that there are no conflicts of interest relevant to this manuscript.

Authors’ contribution

M.A. and A.B. contributed equally to this work.

References


Proteomic strategies for target discovery


Popoveva T, Espina V, Bailey C, Liotta L, Petricoin E & Popov S
Munday DC, Surtees R, Emmott E, Dove BK, Digard P, Barr JN,
Munday DC, Emmott E, Surtees R, Lardeau CH, Wu W, Duprex
et al. Popova TG, Turell MJ, Espina V
Narayanan A, Kehn-Hall K, Senina S, Lundberg L, Van Duyne R,
Noisakran S, Sengsai S, Thongboonkerd V,
Proteomic strategies for target discovery M. Amaya
Nishizuka S, Charboneau L, Young L,
188 Pathogens and Disease (2014), 71, 177–189, © 2014 Federation of European Microbiological Societies. Published by John Wiley & Sons Ltd. All rights reserved