Challenges for Biological Interpretation of Environmental Proteomics Data in Non-model Organisms

W. Wesley Dowd

Department of Biology, Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90045, USA

From the symposium “Comparative Proteomics of Environmental and Pollution Stress” presented at the annual meeting of the Society for Integrative and Comparative Biology, January 3–7, 2012 at Charleston, South Carolina.

1Email: wdowd@lmu.edu

Synopsis

Environmental physiology, toxicology, and ecology and evolution stand to benefit substantially from the relatively recent surge of “omics” technologies into these fields. These approaches, and proteomics in particular, promise to elucidate novel and integrative functional responses of organisms to diverse environmental challenges, over a variety of time scales and at different levels of organization. However, application of proteomics to environmental questions suffers from several challenges—some unique to high-throughput technologies and some relevant to many related fields—that may confound downstream biological interpretation of the data. I explore three of these challenges in environmental proteomics, emphasizing the dependence of biological conclusions on (1) the specific experimental context, (2) the choice of statistical analytical methods, and (3) the degree of proteome coverage and protein identification rates, both of which tend to be much less than 100% (i.e., analytical incompleteness). I use both a review of recent publications and data generated from my previous and ongoing proteomics studies of coastal marine animals to examine the causes and consequences of these challenges, in one case analyzing the same multivariate proteomics data set using 29 different combinations of statistical techniques common in the literature. Although some of the identified issues await further critical assessment and debate, when possible I offer suggestions for meeting these three challenges.

Introduction

Recent years have produced a flourishing of large-scale approaches in biology that promise to revolutionize systems-level understanding of cellular and organismal function. These fields, collectively referred to as “omics” (Joyce and Palsson 2006), assess the structure and functioning of organisms at a comprehensive and unprecedented level of detail. Proteomics, the global study of protein abundance and modification patterns, holds particular promise for elucidating the functional responses of organisms to environmental and biological challenges (Cox and Mann 2011; Tomanek 2011). Indeed, many practitioners argue that the proteome best represents the functional molecular phenotype. In the well-known central dogma of molecular biology (depicting information flow from genome to transcriptome to proteome, and now to metabolome), proteins perform most of the molecular work of the cell and constitute a substantial and dynamic component of cellular structures.

The field of proteomics promises two grand, and as yet largely unrealized, outcomes: (1) to assess the abundances, modifications, and interactions of all proteins present in a sample and then to determine how these change through time (e.g., ontogeny, stress, or disease) and (2) in so doing, to unearth novel and integrative functional responses of organisms to diverse environmental challenges, over a variety of time scales and at different levels of organization. The first of these has been hampered primarily by technological factors, such as mass-spectrometer scanning rates (Mulvey et al. 2010) or techniques for separation of native proteins (Monti...
of coastal marine organisms to environmental experience in examining the proteomic responses in the literature and from personal identified by review of recent environmental proteomics studies. Technological advances in protein separation, protein identification by mass spectrometry, and large-scale genomic and transcriptomic sequencing (e.g., next-generation sequencing platforms such as Roche 454 and Illumina) now allow biologists to more easily apply proteomics techniques to non-model organisms in environmentally realistic scenarios (reviewed by Tomanek 2011; Diz et al. 2012). These technologies, which have been reviewed elsewhere and which are succinctly summarized in other contributions to this symposium, have numerous potential applications for integrative biologists interested in physiological, ecological, and evolutionary questions. Such applications lie well outside the typical realm of proteomics, which tends to be biomedically focused. For example, contributions to this symposium apply proteomics methods to examine responses of organisms to environmental and toxicological challenges (Martyniuk and Denslow, this volume); correlations of function with ecology and biogeography (Fields et al., this volume); species’ differences (Tomanek, this volume); the patterns of environmental adaptation (Abbaraju et al., this volume); and limitations to adaptive processes (G. Dilly, unpublished data). Thus, other contributions to this symposium highlight both the scope and depth of what we have learned and can learn by applying study of the proteome to environmental questions.

In contrast, this review considers how proteomics techniques can be more effectively applied to environmental questions to reach biological conclusions. Specifically, I examine the dependence of biological interpretations of proteomics data on three factors:

- The specific experimental context (e.g., treatment conditions and tissue type);
- The choice of statistical analysis methods;
- The degree of proteome coverage and the rate of identification of proteins-of-interest, both of which are typically much less than 100% (i.e., “analytical incompleteness,” Wilkins et al. 2006).

These three challenges to biological inference were identified by review of recent environmental proteomics studies in the literature and from personal experience in examining the proteomic responses of coastal marine organisms to environmental challenges. Although similar concerns have been raised by others (e.g., Broadhurst and Kell 2006; Wilkins et al. 2006; Diz et al. 2011), I argue that the implications of these challenges are especially profound for the field of environmental proteomics, and I demonstrate that these three factors generally have not received sufficient consideration in recent environmental proteomics articles. For each challenge, I first characterize the general status in the field through a review of recent environmental proteomics literature, I then examine causes and consequences of these challenges, using examples from my own work on environmental stress proteomics in sharks and mussels. Finally, when possible, I offer suggestions for addressing these challenges as a field. The first two of these challenges may well be overcome by more robust and more sophisticated application of techniques for experimental design and statistical analysis, whereas the third challenge requires advances in technology and generation of a critical mass of shared sequence information and functional protein analyses in non-model organisms. As we address these challenges explicitly in future studies, the power and contributions of environmental proteomics will only continue to expand.

**Methods: Literature review**

The review of the literature focused on environmental proteomics publications appearing after 2005. Rather than a comprehensive review of all available studies, I focused instead on a representative subsample. The studies were selected from a list generated by searching the Web of Science database (Topic=[(stress or salinity or temperature or hypoxia or environment*) and protein*) AND Year Published=[2011 or 2010 or 2009 or 2008 or 2007 or 2006 or 2005]), and the list included many studies recently reviewed by Tomanek (2011). Although a number of proteomics-focused journals exist, the selected studies were published primarily in journals that cater to integrative and comparative biologists. This focus is consistent with the goals of the present symposium, and the existence of such publications indicates an effort on the part of the environmental proteomics community to make these techniques visible and applicable to a wider audience.

In total, 31 proteomics data sets published in 19 different articles were reviewed; in cases in which more than one species or more than one tissue was assessed in a given article, these were counted as separate data sets (complete list of studies and summary data are provided in Supplemental Table S1). However, the two studies in which more than one
tissue from the same organisms and treatments were examined (Dowd et al. 2010a, 2010b) were treated as single entities when analyzing criteria for experimental design. Consequently, the total number of data sets analyzed was reduced to 29 in some cases. Each experimental data set was initially scored for 27 criteria for experimental design, statistical analysis, and results (Table 1). These criteria covered not only the specific designs, separation methods, and statistical analyses of environmental proteomics studies but also how well those details were explained in the resulting publications. Clearly, this is not an exhaustive summary of the complete body of environmental proteomics literature, but the results are illustrative of the three challenges outlined earlier in the text.

**Summary of the data**

The summary data set included organisms from several life history stages (larvae to adults) and from a number of taxonomic groups (bony and cartilaginous fishes, molluscs, annelids, echinoderms, crustaceans, urochordates, and zooxanthellae symbionts of corals). All the studies used entire organisms for the experimental exposures or treatments, as opposed to cell lines or isolated tissues. Consequently, the results could be interpreted in each case in the physiological and biochemical context of an entire organism. Of the 31 studies, only three fractionated the proteome to look at specific subsets of proteins (in this case, peroxisomal proteins) (Apraiz et al. 2006). The distribution of experimental approaches for separating protein mixtures was heavily skewed toward two-dimensional gel electrophoresis (2D-GE; 28 of 31 data sets) rather than liquid chromatography mass spectrometry (LC-MS). This is likely due to a number of factors that currently hamper application of LC-MS to non-model organisms, most notably a lack of sufficient sequence information to allow reconstruction of predicted protein sequences from numerous digested peptides that separate into different liquid fractions (Cox and Mann 2011). Both the 2D-GE and LC-MS separation approaches have their advantages and disadvantages (Görg et al. 2004; Rabilloud et al. 2010; Cox and Mann 2011); among the most important and well known of these are under-representation of large, membrane-bound proteins in 2D-GE studies and relatively high technical variability in both methods (i.e., large coefficients of variation up to 15–30% for the same peptide/protein between replicate runs of the same sample on the same apparatus) (e.g., Hunt et al. 2005; Bandow et al. 2008).

To overcome high levels of variation, previous authors have advocated increasing levels of biological replication (i.e., analyzing more individuals) before increasing technical replication (e.g., Rocke 2004; Chich et al. 2007), especially when analytical or financial resources are limited. Technical replication alone overestimates the precision of measurements and exaggerates significant differences among treatment groups, increasing the rate of false positives (i.e., incorrect rejection of the null statistical hypothesis of no significant difference) (Chich et al. 2007). Furthermore, these authors caution against the pooling of samples except in cases when the cost of analysis is prohibitively high (Rocke 2004). In the data sets examined, there was a mean of 5 ± 2 (range: 3–11; median: 5; Fig. 1A) biological replicates (individuals or pools). Of these, 38% (11 of 29) used pooled samples, sometimes because of limited availability of tissue, as in studies of larval proteomes (e.g., Silvestre et al. 2010a). Approximately 80% (23 of 29) of the studies did not carry out any technical replication.

**Challenge 1: Experimental contexts**

The challenge of designing rigorous, environmentally relevant, and feasible experiments is not unique to the field of environmental proteomics. It is increasingly clear from many fields of biology that context matters; experimental results depend intimately on the timing, intensity, and duration of the experimental treatment, in addition to other factors such as species’ interactions, behavior, multiple interacting stressors, and chance/stochasticity. The time and expense required to carry out proteomics studies often necessitates a focus on a few experimental conditions or time points, thereby escalating the requirement for very judicious and well-justified experimental design. Furthermore, the large quantity of information generated by a single proteomics comparison (e.g., controls versus one treatment) may lead to researchers (and readers) focusing on the breadth and depth of the results rather than critically examining the relevance of the data to a particular environmental question.

**Literature review**

The review of the literature relevant to this challenge focused on two questions:

1. Were the specific conditions and/or time points chosen for analysis with proteomics justified and relevant to environmental conditions experienced by organisms in the wild? Examples of environmentally relevant experimental contexts included exposures designed to mimic tidal or diel cycles (Dowd et al. 2010b) or...
Table 1 Experimental, statistical, and analytical criteria by which environmental proteomics studies in the literature were evaluated (Supplemental Table S1)

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Codes (or description)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I. Experimental</strong></td>
<td></td>
</tr>
<tr>
<td>Age class</td>
<td>0 = not stated, 1 = larvae or embryo, 2 = juvenile, 3 = adult, 4 = mixed</td>
</tr>
<tr>
<td>Experimental unit exposed to environmental challenge</td>
<td></td>
</tr>
<tr>
<td>Tissues analyzed</td>
<td>0 = whole organism, 1-2 = number if analyzed separately</td>
</tr>
<tr>
<td>Number of biological replicates per treatment</td>
<td>(may be pooled or individuals)</td>
</tr>
<tr>
<td>Nature of biological replicates analyzed</td>
<td>1 = pooled, 2 = individuals</td>
</tr>
<tr>
<td>Technical replicates run</td>
<td>1 = yes, 2 = no, 3 = unknown</td>
</tr>
<tr>
<td>Fractionation of proteome</td>
<td>0 = none, 1 = mitochondrial or nuclear fraction, 2 = phosphoproteome, 3 = other</td>
</tr>
<tr>
<td>Separation technique</td>
<td>0 = LC-MS, 1 = 2D-GE</td>
</tr>
<tr>
<td>pH/pl scale</td>
<td>0 = N/A (LC-MS), 1 = 3–10, 2 = 4–7, 3 = other</td>
</tr>
<tr>
<td>Gel dimensions</td>
<td>0 = N/A (LC-MS), 1 = 11 cm ± 1 cm, 2 = 18 cm ± 2 cm, 3 = other</td>
</tr>
<tr>
<td>Samples collected from laboratory or field exposures</td>
<td>1 = laboratory, 2 = field, 3 = both</td>
</tr>
<tr>
<td>Number of treatments</td>
<td>(number of experimental conditions or time points analyzed with proteomics, including controls)</td>
</tr>
<tr>
<td>Justification of experimental conditions/time points</td>
<td>0 = none, 1 = somewhat, 2 = explicit</td>
</tr>
<tr>
<td>Exposure duration</td>
<td>1 = short (≤ 24 h), 2 = long (&gt; 24 h)</td>
</tr>
<tr>
<td>Conditions/time points discussed separately</td>
<td>0 = no, 1 = somewhat, 2 = separate paragraphs or sections</td>
</tr>
<tr>
<td>Conclusions varied depending on condition/time point</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td><strong>II. Statistical</strong></td>
<td></td>
</tr>
<tr>
<td>Statistical tests outlined</td>
<td>0 = none, 1 = somewhat, 2 = explicit</td>
</tr>
<tr>
<td>Location of description of statistical methods</td>
<td>1 = text, 2 = Supplemental Material</td>
</tr>
<tr>
<td>Number of significant proteins</td>
<td>(according to statistical criteria defined in the text)</td>
</tr>
<tr>
<td>Univariate/multivariate statistical techniques</td>
<td>1 = univariate t-test, U-test, or ANOVA, 2 = multivariate PCA, 3 = multivariate PLS-DA, 4 = multivariate other (e.g., Bayesian, artificial neural networks), 5 = clustering analysis, dendrogram, 6 = Protein Pilot error factor</td>
</tr>
</tbody>
</table>
environmentally realistic levels of exposure in toxicological studies (Martyniuk et al. 2010a, 2010b). Publications were grouped into three broad categories based on the degree of justification of the chosen experimental contexts: none = no justification of levels and durations of exposure, and/or of time points; somewhat = all studies falling between the other two categories; and explicit = clear, thorough justification in the article (Introduction, Materials and Methods, Results, or Discussion) of the specific levels and durations of exposure, and/or of the time points that were chosen for analysis with proteomics, including their relevance to conditions experienced in the wild. An example of explicit justification can be found in Dowd et al. (2010b, 94):

"We focused on the 24 h recovery time point for our proteomics analyses for three reasons: 1) reoxygenation induces as or more severe damage as the period of oxygen limitation; 2) compensatory stress response mechanisms may be more potent during the recovery period than during the actual stress period; and 3) hypoxic episodes are naturally spaced 24 h apart on the reef platform. Thus, our proteomic data represent the phenotype at the time when the next stressful episode would likely be encountered. This 24 h lag also coincides with the onset of the delayed phase of ischemic preconditioning in mammalian models."

Although the latter two categories required a degree of subjectivity, the general pattern is robust. Of the studies reviewed, less than half (12 of 29) included what was deemed to be an explicit justification for the experimental conditions and time points chosen; in four studies, no justification could be found.

(2) How many conditions or time points were compared using proteomics? Of the 29 proteomics data sets examined, approximately half (14) compared only one treatment condition or time point with one control condition or time point, and only two studies compared more than four treatments. The average number of experimental conditions was 3 (range: 1–8; median: 2). Even within the relatively narrow time frame of studies reviewed, there was a positive correlation between the year of publication and the number of experimental conditions or time points analyzed (Kendall’s $\tau_b$ correlation coefficient = 0.415, $P = 0.008$). This encouraging trend suggests that more complex experimental designs are becoming more desirable and/or more tractable in environmental proteomics. Surprisingly, there was no correlation between
the extent of justification of the experimental context and the number of conditions or time points examined (Kendall’s $r_b$ correlation coefficient $= 0.052$, $P = 0.757$).

How does experimental context influence biological conclusions in environmental proteomics?

The clearest evidence for the importance of experimental context comes from analysis of the discussion sections of those studies in which more than two experimental conditions or time points were analyzed. In 13 of these 14 studies (only Silvestre et al. 2010b differed), the biological conclusions drawn from the proteomics analysis varied—often substantially—among the different treatment conditions or time points. These differences in interpretation have very important implications for how proteomics experiments need to be designed to capture the most environmentally relevant patterns. For example, in thermal stress experiments with intertidal organisms, should we sample at the peak temperature or 24 h later when the organisms next would be exposed to high temperature? The choice of tissue type may also influence proteomics results. In the two studies that examined more than one tissue from the same organisms under the same conditions, the conclusions varied dramatically between the different tissues (Dowd et al. 2010a, 2010b). Given that 18 of the 29 studies analyzed only a single tissue, caution must be employed when attempting to extrapolate from tissue-specific effects to consequences for entire organisms. In cases in which critical tissues are carefully chosen and the choice is explicitly justified (e.g., the heart of porcelain crabs) (Stillman and Tagmount 2009), the approach can still be quite powerful. Notably, nine studies analyzed the proteomes of whole organisms; this approach may confound interpretation if tissue-specific responses to a particular treatment are offset by opposing patterns of protein expression in different tissues.

A recent proteomics analysis (W. W. Dowd and G. N. Somero, manuscript in preparation) further demonstrates the importance of defining and justifying experimental contexts for environmental proteomics studies. This study examined context-specific proteome responses to the same body temperature in intertidal species of mussels of the genus *Mytilus*, with the goal of examining underlying protein expression changes that might explain differential organismal consequences of elevated body temperature in two blue mussel species (Braby and Somero 2006b; Lockwood and Somero 2011a; W. W. Dowd and G. N. Somero, submitted for publication). The relatively warm-adapted invasive species *Mytilus galloprovincialis* and the relatively cool-adapted native species *Mytilus trossulus* can compete for space in regions of the Pacific coast of North America where their distributions overlap (Shinen and Morgan 2009). The competitive outcomes in the field seem to be dependent on the specific
environmental context, including whether a specific site is located in a warm intertidal location versus a relatively cool subtidal location (Schneider and Helmuth 2007; Schneider 2008). In addition, both species exhibit lower survival following thermal challenges in seawater compared with thermal challenges in air (W. W. Dowd and G. N. Somero, manuscript in preparation). Although elevated body temperatures are more likely to be encountered at low tide when the organisms are emersed (Denny et al. 2011), transcriptomic and proteomic studies of mussels’ gill tissues (principal site of both respiratory gas exchange and filter feeding) conducted to date have only compared the responses of these two species to elevated body temperatures (24–32°C) when the organisms were immersed in seawater throughout the experiment (Lockwood et al. 2010; Tomanek and Zuizow 2010). Furthermore, episodes of elevated body temperature tend to be clustered into multi-day windows, during which body temperature rises in the middle of 2 or more consecutive days (Denny et al. 2011). Consequently, we examined the proteomic responses of individuals of these two species over 3 consecutive days during which temperatures gradually ramped up to 33°C and back down to 13°C over a 4-h period each day, a thermal regimen that approximates field patterns. The mussels were either immersed or emersed during these temperature ramps. Using 2D-GE, we then compared protein expression patterns with individuals held in 13°C seawater throughout the same period (W. W. Dowd and G. N. Somero, manuscript in preparation).

The proteomic response to the same elevated body temperature varied significantly, depending on the mussels’ emersion/immersion status (Table 2). *Mytilus galloprovincialis* mounted a greater proteomic response to elevated body temperature in seawater, whereas *M. trossulus* mounted a greater response to the same temperature in air. Qualitative differences in protein expression were also noted among the treatments within the same species. Of the 1340 spots on the two-dimensional gels for *M. galloprovincialis*, only 7 of the 51 spots that were found to be significantly differentially expressed among the three treatment groups changed in the same direction under both emersion and immersion. Of the 1374 spots on the gels for *M. trossulus*, only 3 of the 34 significantly differentially expressed spots changed in the same direction under both conditions. Mass spectrometry protein identification and subsequent functional interpretation of these data are pending, but it is reasonable to expect that different biochemical processes are represented in the different lists of proteins responsive to thermal stress in air versus in water. These results have potentially critical implications for understanding the context-specific outcomes of interactions between these two species and for design of future comparative studies.

### Meeting the challenges of experimental context

Overall, these findings highlight the importance of designing experiments that maximize the environmental relevance before conducting proteomics analyses. Explicit justification of the specific experimental conditions and/or time points should be clear from the Introduction or Materials and Methods sections of publications. Results of the literature review also demonstrate the importance of not extrapolating too liberally from two-treatment or single-tissue experiments. The technology and expertise now exist to design more complex, integrative experiments covering a wider variety of experimental contexts and types of tissues; trends in the reviewed data sets suggest that the environmental proteomics field is moving in that direction. Of course, even the recent laboratory experiments described earlier for *Mytilus* mussels still fail to capture the contextual richness of field exposures. For example, other recent work has documented species-specific biochemical responses in these mussels to changes in salinity (Evans and Somero 2010; Lockwood and Somero 2011b). This is in agreement with field observations showing that a complex interplay between thermal and osmotic stress may be driving mussels’ distributions in estuarine habitats (Braby and Somero 2006a). Notably, only one of the reviewed studies analyzed samples that were collected directly from the field (Silvestre et al. 2010b). Field studies are perhaps the best way to capture the complex environmental contexts that many are ultimately interested in understanding, and efforts to extend proteomics into the field should be encouraged.

### Table 2 Summary of the numbers of statistically significantly differentially expressed proteins in two congeneric mussel species exposed to the same elevated body temperature under two different contexts: emersion (air) and immersion (seawater)

<table>
<thead>
<tr>
<th>Species</th>
<th>Seawater</th>
<th>Air</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Mytilus galloprovincialis</em> (1340 protein spots)</td>
<td>25 ↑; 11 ↓</td>
<td>12 ↑; 4 ↓</td>
</tr>
<tr>
<td><em>Mytilus trossulus</em> (1374 protein spots)</td>
<td>3 ↑; 8 ↓</td>
<td>13 ↑; 7 ↓</td>
</tr>
</tbody>
</table>

Up and down arrows indicate proteins that increased or decreased in abundance relative to controls, respectively. Data were arcsine-square root transformed and analyzed with a univariate analysis of variance and a significance threshold of \( P < 0.01 \).
Challenge 2: Statistical choices

In all quantitative biological studies, robust and appropriate statistical analyses are paramount for reaching sound biological inferences (Sokal and Rohlf 1995). This requirement assumes even greater importance in environmental proteomics (and in “omics” studies generally), given the massive and complex data sets that are produced (Domon and Aebersold 2006; Chich et al. 2007). These data do not always conform to the assumptions of normality and homoscedasticity of parametric, univariate (i.e., feature-by-feature, where each protein is a feature) statistical tests (Rocke 2004; Wilkins et al. 2006). Consequently, a wide array of data transformations and modified univariate statistical approaches (e.g., generating an empirical distribution by resampling the data) (Dudoit et al. 2003) have been proposed and tested with “omics” data sets, each designed to overcome one or more shortcomings of the standard, parametric Student’s t-test or analysis of variance (Table 3).

Proteomics data sets also suffer from an extreme degree of undersampling (Smit et al. 2007), such that the number of features analyzed for each replicate (average of 592 proteins in the reviewed studies; see later) far exceeds the number of replicates (average of five replicates; see earlier). In undersampled scenarios, the risk of spurious, false positives in univariate, feature-by-feature statistical tests (i.e., the type I error rate \( \alpha \), set by the author as the critical \( P \) value) needs to be taken into account. For example, if 592 univariate tests are conducted on a proteomics data set with an \( \alpha \) of 0.05, then \( 592 \times 0.05 = 29 \) proteins may be called significant by chance alone. Consequently, several methods that control the proportion of false positives (i.e., the false discovery rate [FDR]) have been proposed: the classical Bonferroni correction that controls the family-wise error rate with 0% FDR (Chich et al. 2007); the Benjamini and Hochberg (1995) linear step-up method; and the Bayesian \( q \) value (Storey 2002; Storey and Tibshirani 2003). These FDR methods vary considerably in their stringency, giving researchers a choice in the tradeoff between avoiding false discoveries (e.g., use of the conservative Bonferroni correction \( P = \alpha/n \) is advocated for biomarker development) (Broadhurst and Kell 2006; Chich et al. 2007) and making no discovery (Devlin et al. 2003). Of these methods, the Bayesian \( q \) value approach is the least conservative. This method determines the FDR when a given level of \( \alpha \) sets the statistical significance level; researchers are free to choose the proportion of “false leads” (typically 10% or less) (Karp et al. 2007) they are willing to risk. At the opposite extreme, some would argue that there is no need to correct for multiple comparisons (e.g., Rothman 1990), because a “universal null” hypothesis does not align with conceptual models of natural processes.

Ultimately, no univariate statistical test is capable of detecting correlated changes in protein abundance or of assessing and describing global changes in protein expression patterns. Importantly, such patterns may offer greater insight into pathway-level or systems-level responses to environmental challenges. In recent years, multivariate statistical approaches capable of addressing these systems-level issues have gained greater prominence in the proteomics community. The most commonly used are principal components analysis (PCA) (e.g., Marengo et al. 2006) and partial least-squares discriminant analysis (PLS-DA). PCA determines the combinations of individual protein expression levels (so-called latent variables) that best explain the global spread in the data, regardless of the treatment group to which each sample belongs (i.e., the test is unsupervised). The more powerful, supervised PLS-DA technique is a regression extension of PCA; PLS-DA takes into account treatment-group membership to maximize between-groups variation explained by the latent variables (Pérez-Enciso and Tenenhaus 2003; Karp et al. 2005). The PLS-DA technique is robust to correlated expression changes (in statistical terms, multicollinearity), it outperforms PCA when between-group variability does not dominate data sets (Dai et al. 2006), and it allows the proteins to be ranked by their influence on treatment-group separation using the magnitudes of variable importance in the projection (VIP) scores (Wold et al. 2001; Palermo et al. 2009). PLS-DA has proven very effective in proteomics analyses, especially when used in combination with univariate approaches (e.g., Lee et al. 2003; Karp et al. 2005; Pedreschi et al. 2009; Dowd et al. 2010b). Other multivariate methods are rapidly becoming available, including artificial neural networks (Smit et al. 2008), genetic algorithms (Li et al. 2004), and Bayesian networks (Werhli and Husmeier 2007). Although these new methods hold great promise for elucidating global patterns in “omics” data sets, their utility for environmental proteomics remains to be demonstrated.

Literature review

Given the plethora of possible techniques available and the statistical considerations just described, literature review relevant to this challenge focused on the following two questions:

1. Which statistical techniques of the many available univariate and/or multivariate approaches were
used to analyze proteomics data sets, and were the assumptions of the tests explicitly mentioned and tested? Each of the reviewed studies used a univariate, feature-by-feature statistical approach; 26 of 29 studies used a standard \( t \)-test/analysis of variance (ANOVA) or Mann–Whitney \( U \) test. As far as could be determined from the publications, few, if any, of these studies explicitly tested the assumptions of the chosen univariate tests. In five cases, a univariate resampling approach was used because the authors stated that the assumptions of a standard parametric test likely could not be met. Only seven of 29 studies combined the univariate analyses with a multivariate approach (five used PCA and two used PLS-DA and VIP for feature selection).

(2) Was any correction or adjustment made for testing multiple hypotheses to control the FDR? Although each of the studies incorporated some form of univariate, feature-by-feature statistical analysis, only 6 of 29 explicitly mentioned multiple-hypothesis testing and the false-discovery problem. However, none of the reviewed studies used the FDR-controlling procedures described earlier. Instead, only six studies used a more stringent critical \( P \) value (e.g., \( P < 0.01 \) or \( P < 0.02 \)), lowering the type I error rate but perhaps doing so at the expense of missed information.

### Challenges for environmental proteomics

Differences in statistical analysis choices confound biological interpretation across proteomics studies. In the data sets that were reviewed, the percentage of proteins in the data set reported to be significantly differentially expressed ranged from 0.04% to 54.6% (mean 14.8\% ± 13.4\% standard deviation [SD]; median: 11.8\%; Fig. 1B). There was no relationship between the number of proteins that changed significantly and the number of proteins that were analyzed (linear regression, \( P = 0.893 \)). Similarly, there was no detectable effect of analyzing individual tissues versus analyzing entire organisms (Mann-Whitney \( U \) test, \( P = 0.572 \)). Notably, the fraction of the proteome found to change significantly in abundance increased with the number of biological

### Table 3

Partial list of univariate statistical methods that may be employed to discern differentially expressed features among treatments in “omics” data sets; not all methods are commonly used in proteomics data analysis

<table>
<thead>
<tr>
<th>Statistical test or transformation</th>
<th>Transformation formula</th>
<th>Comments or assumptions</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data</td>
<td>( Y = X )</td>
<td>Assumptions of normality and homoscedasticity</td>
<td></td>
</tr>
<tr>
<td>Parametric ( t )-test/ANOVA</td>
<td>( Y = X )</td>
<td>Less power than parametric tests</td>
<td></td>
</tr>
<tr>
<td>Non-parametric Mann–Whitney ( U ) test</td>
<td>( Y = X )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance-stabilizing transformations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logarithmic</td>
<td>( Y = \log_{10} X )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power (Box-Cox)</td>
<td>( Y = X^b )</td>
<td>( b ) determined by regression</td>
<td></td>
</tr>
<tr>
<td>Arcsine-square root</td>
<td>( Y = \sin^{-1}(X^{0.5}) )</td>
<td>Useful for % volumes in 2D-GE studies</td>
<td></td>
</tr>
<tr>
<td>Transformations that specifically protect against high variance at low expression level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inverse hyperbolic sine</td>
<td>( Y = \sinh(X) )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generalized logarithm</td>
<td>( Y = \ln[(X - \alpha + ((X - \alpha)^2 + c)^{0.5})] )</td>
<td>Requires technical replication for empirical calibration of ( c ), ( \alpha ) = mean background expression level</td>
<td>Durbin et al. (2002)</td>
</tr>
<tr>
<td>Started logarithm</td>
<td>( Y = \ln(X + c) )</td>
<td></td>
<td>Rocke and Durbin (2003)</td>
</tr>
<tr>
<td>Log-linear hybrid</td>
<td>( Y = \frac{X}{k} + \ln(k) - 1, X \leq k )</td>
<td></td>
<td>Rocke and Durbin (2003)</td>
</tr>
<tr>
<td>Other alternatives</td>
<td></td>
<td>May not be appropriate with small sample sizes (see text)</td>
<td>Dudoit et al. (2003)</td>
</tr>
<tr>
<td>Resampling (bootstrap, permutation)</td>
<td></td>
<td>Not well-developed</td>
<td>Vuong et al. (2011)</td>
</tr>
</tbody>
</table>
| Outlier-based methods             |                        |                         | |}

Many transformations are intended to stabilize variance across all features in a data set. For each transformation of data, the raw data \( X \) are transformed according to the formula presented, and statistical analyses are conducted using the resulting data \( Y \).
replicates analyzed (logistic regression, $P<0.001$). However, because of the variety of statistical techniques, underlying assumptions, and critical $P$ values used, it is nearly impossible to directly compare these percentages—and, more importantly, their corresponding biological implications—across studies. Any underlying relationships would also be confounded by differences among studies in other factors such as the species under study, the technique employed for separating proteins, the tissue analyzed (see earlier in the text), and, perhaps most importantly, the different experimental conditions that represent different degrees of challenge to the organisms.

To more concretely demonstrate the influence of the choice of statistical analysis on subsequent biological conclusions, I analyzed a single published 2D-GE proteomics data set (Dowd et al. 2010b) using 29 different combinations of statistical methods. The list of univariate analyses included both non-parametric (Mann–Whitney $U$ test) and parametric methods using several data transformations from Table 3 ($t$-test on raw data, $t$-test on logarithm base 10 transformed data, $t$-test on arcsine-square root transformed data, and $t$-test on inverse hyperbolic sine transformed data). The percentages of proteins in the untransformed data set that met the assumptions of normality (Lilliefors’ test) and homoscedasticity (Levene’s test) for a $t$-test were 82.0% and 91.8%, respectively. The only data transformation to achieve any increase in both values was an arcsine-square root transformation (slight improvements to 88.0% and 92.2%, respectively). The arcsine-square root transformation is recommended by statistical texts for percentage data, and 2D-GE data are often presented for each protein as a percentage of the total proteome. For each of the parametric methods, I conducted the $t$-test using both the standard $t$-distribution and a resampled $t$-distribution generated by 2000 bootstrap iterations of the data for that protein. Furthermore, the results of each statistical test were adjusted using two of the three different FDR-controlling methods described earlier (the Bonferroni technique is far too conservative). The analyses were conducted in Matlab v7.12 (The Mathworks, Inc., Natick, MA, USA) using commands (matstat, mafdr) available in the Bioinformatics Toolbox. To examine systems-level patterns in the data, I also used two multivariate approaches: PCA in Matlab and PLS-DA in SIMCA-P+ software (Umetrix, Umeå, Sweden).

The data set, including abundances for 768 protein spots on 2D gels, was previously used to assess the responses in the rectal gland of epaulette sharks (*Hemiscyllium ocellatum*) to episodic hypoxia and anoxia (Dowd et al. 2010b). Animals were exposed to one of four treatments: normoxic handling controls ($n=5$), one episode of hypoxia ($n=6$), one episode of anoxia ($n=4$), or two episodes of hypoxia separated by 24 h ($n=6$). In the original analysis, we were most interested in comparing expression patterns within each treatment group with the normoxic controls (Dowd et al. 2010b). For comparison with the published results and for simplicity, the present univariate statistical analyses were limited to comparing protein expression levels in each of the low oxygen groups with those in the control group (i.e., a two-treatment comparison); analogous trends emerged when the data were analyzed across all treatments with a more appropriate one-way ANOVA (data not shown). The multivariate analyses were conducted on data from all four treatments.

The consequences of choosing different statistical tests included both quantitative and qualitative effects. The number of proteins found significant with a univariate, parametric $t$-test varied by 27–41% of the maximum number depending on the transformation (Table 4). In each pair of standard $t$-distribution versus bootstrap $t$-distribution analyses, the bootstrap generated a greater number of significant differences (Table 4). This pattern may arise from the sample sizes being too small for application of the bootstrap (Molina et al. 2005; Chich et al. 2007). Qualitatively, the lists of proteins found to be significant also varied widely among the different statistical methods. For example, only 13 of the 37 different proteins that returned a statistically significant result for the anoxia treatment in one or more technique were found to be significant by all the techniques (Fig. 2). Clearly, such differences will have profound consequences for subsequent biological interpretation.

The FDR-controlling methods also yielded substantially different results. Although the Bonferroni correction may be dismissed as far too strict for environmental proteomics data, even the less-stringent Benjamini and Hochberg (1995) method yielded few, if any, significant differences in the pairwise comparisons (Table 4). Similarly, the Bayesian $q$ value FDR-controlling technique (at $<10\%$ FDR) also revealed no statistically significant results when comparing rectal gland proteomes after exposure to anoxia relative to control conditions (Table 4). In comparison, using an arbitrarily “strict” critical $P$ value of 0.01 without controlling the FDR produced four significant proteins for the anoxia treatment. Treatment differences were substantial; when comparing a single episode of hypoxia with controls, the same Bayesian $q$ value approach yielded...
reasonable lists of 31–42 significant proteins for the standard parametric tests (37–38 using a critical $P$ value of 0.01). It is unclear why, in some cases, the Bayesian $q$ value approach inflated the number of significant proteins so dramatically when performed on statistics derived from bootstrap distributions (Table 4; hypoxia 1 data).

The multivariate techniques varied in their ability to discriminate the four treatment groups. A PCA analysis on the complete log-transformed and mean-centered data set (to normalize across spots with highly variable volumes on 2D gels) explained 41% of the variation in the data set with the first three latent variables (a.k.a. principal components), but the analysis failed to clearly distinguish the four treatment groups (Supplemental Figure S1A). It should be noted that although a PCA or other multivariate analysis conducted using only those proteins found to be significantly differentially expressed with a univariate statistical test (as in several of the reviewed studies) will certainly distinguish the treatment groups, such an analysis generates no further insight than do the univariate tests alone. Rather, multivariate analyses should first be conducted on

### Table 4

The number of epaulette shark rectal gland proteins meeting statistical significance criteria in pairwise comparisons of low oxygen treatments with normoxic controls

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Statistical test and transformation</th>
<th>Mann–Whitney $U$</th>
<th>t-test</th>
<th>t-test log10</th>
<th>t-test asin-sq rt</th>
<th>t-test asinh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard t-distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anoxia</td>
<td>20 / 0 / 0</td>
<td>28 / 0 / 0</td>
<td>22 / 0 / 0</td>
<td>30 / 0 / 0</td>
<td>27 / 0 / 0</td>
<td></td>
</tr>
<tr>
<td>Hypoxia 1</td>
<td>87 / 0 / 41</td>
<td>126 / 1 / 34</td>
<td>126 / 0 / 42</td>
<td>133 / 0 / 35</td>
<td>127 / 1 / 34</td>
<td></td>
</tr>
<tr>
<td>Hypoxia 2</td>
<td>37 / 0 / 0</td>
<td>63 / 0 / 0</td>
<td>58 / 0 / 0</td>
<td>62 / 0 / 0</td>
<td>63 / 0 / 0</td>
<td></td>
</tr>
<tr>
<td>Bootstrap t-distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anoxia</td>
<td>35 / 0 / 0</td>
<td>30 / 0 / 0</td>
<td>35 / 0 / 0</td>
<td>34 / 0 / 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypoxia 1</td>
<td>151 / 0 / 220</td>
<td>147 / 0 / 211</td>
<td>147 / 0 / 233</td>
<td>140 / 0 / 243</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypoxia 2</td>
<td>72 / 0 / 0</td>
<td>69 / 0 / 0</td>
<td>74 / 0 / 2</td>
<td>79 / 0 / 3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The same data set was analyzed using each of the tests and data transformations listed (described in Table 3). For each of the nine statistical tests, results are presented for three degrees of control of the false discovery rate (FDR): no correction (i.e., $P<0.05$) / correction using the Benjamini and Hochberg (1995) linear step-up method / correction using the Bayesian $q$ value approach with FDR less than 10%. Treatments are as described in text.
the complete data set, after which a subset of proteins with significant influence on the separation can be determined using additional criteria (Chong and Jun 2005; Niijima and Okuno 2009; Palermo et al. 2009). In contrast to the PCA, the PLS-DA analysis of the complete data set clearly separated the four treatment groups (data not shown), and the first three latent variables described 35% of the variation in the data set. When we (arbitrarily) selected the top 5% of proteins (38 proteins) based on their VIP scores (VIP scores have a mean of 1; all selected scores were >1), the new reduced PLS-DA model explained 49% of the variation in the data (Supplemental Figure S1B).

Meeting the challenges of statistical choices

It is clear from the examples earlier that the choice of statistical test may have profound effects on the biological conclusions drawn from a single proteomics data set. Despite the availability of a wide array of statistical techniques in the literature, only a few were employed in the reviewed studies. In the most general sense, the level of statistical sophistication in the field of proteomics has lagged behind that of the transcriptomics community and has also lagged behind the rapidly increasing sophistication of methods for separation and identification of proteins. This is perhaps a consequence of vendor-specific analysis packages included in proteomics software; as far as could be determined from the publications, 20 of the 31 studies that were reviewed conducted statistical analyses solely in proprietary proteomics software rather than in statistical software packages. In contrast, shared analysis workflows such as those hosted on Bioconductor (Gentleman et al. 2004) for the R statistical programming environment have essentially standardized data processing for microarrays, including use of FDR-controlling procedures.

Consistent, robust use of statistical tools and easy interpretation of experimental designs and statistical methods should be focal points for the environmental proteomics community. At the very least, there is a need for FDR-controlling techniques in univariate analyses, for two reasons: (1) not only to prevent false discoveries but also (2) in some cases, to maximize the yield of information from costly and time-consuming proteomics experiments rather than using an arbitrarily strict $P$ value cutoff. For multivariate, systems-level analyses, PCA is useful as an exploratory technique, whereas PLS-DA is more powerful. Although not the most statistically rigorous technique, Karp et al. (2005) recommended a combination of univariate analysis and multivariate PLS-DA, highlighting overlap in the lists of proteins generated by the univariate tests and the VIP method. In the analysis of epaulette sharks, only 14 of the 38 top VIP scores were found to be statistically differentially expressed in any treatment relative to the control using the univariate approach. Thus, PLS-DA corroborated 40–70% of the univariate results (depending on the univariate test chosen), and it generated a reasonable list of additional proteins-of-interest for downstream interpretation and subsequent validation studies. However, several outstanding questions remain regarding the PLS-DA approach (e.g., how many latent variables to include in the model and what VIP-score criteria to use to select features of interest).

Further collaboration between practitioners of environmental proteomics and biostatisticians is crucial for developing greater ability to interpret the systems-level biological consequences of the large data sets produced in proteomics studies. Any new techniques will need to account for the facts that subtle numerical changes in protein abundance may be biologically meaningful yet difficult to detect with statistical rigor (e.g., for transcription factors or other low-abundance proteins), that individual proteins can have multiple functions in numerous biochemical pathways (e.g., glucokinase functions both in glycolysis and apoptosis) (Kim and Dang 2005), that individuals may vary in their response to a given environmental challenge (Rocke et al. 2005), and that protein interaction networks are complex and non-linear (Urfer et al. 2006).

Challenge 3: Analytical incompleteness

The third challenge, termed analytical incompleteness (Wilkins et al. 2006), concerns variation among studies in the degree of proteome coverage and the proportion of proteins-of-interest (i.e., those that are judged significant by some statistical criteria) that were identified by mass spectrometry. Low levels of both measures create two problems: (1) difficulty in comparing biological conclusions from studies with different levels of proteome coverage and (2) difficulty in drawing systems-level mechanistic conclusions when only a subset of proteins that met the criteria of statistical significance are identified.

Literature review

I compiled data from recently published environmental proteomics studies to examine two questions:

(1) How many proteins were quantified in environmental proteomics studies? On average, recent studies assessed $592 \pm 526$ SD proteins.
This average number represents only approximately 6% of the proteome, if we very conservatively assume a proteome size of 10,000 proteins per tissue (Harrison et al. 2002). In general, LC-MS studies analyzed fewer features, due to the bioinformatic complications of reconstructing proteins from individual peptides in non-model organisms (see earlier). Theoretically, there may be 5000 to 20,000 or perhaps many more proteins depending on the organism, cell type, and developmental stage. Thus, biological conclusions are generally limited to a very small subset of the most abundant and most soluble proteins present in a particular sample. The accompanying risk of substantial bias is one that deserves attention in environmental proteomics (as in mammalian studies) (Petrak et al. 2008). Variable proteome coverage of the same tissue under different experimental conditions may also confound biological interpretations and comparisons across studies. In the only example from the literature review in which the same tissue was analyzed twice in the same laboratory (two studies of mussels’ gills; Tomanek and Zuzow 2010; Fields et al. 2012), there was a 17% (554 versus 458 proteins in *M. galloprovincialis*) to 29% (465 versus 331 proteins in *M. trossulus*) difference in the number of proteins on 2D gels in the separate studies. More dramatically, there was a more than 100% difference in proteome coverage between these two studies and the more recent analysis of *Mytilus* mussels described in Challenge 1 earlier (1340 and 1374 spots) (W. W. Dowd and G. N. Somero, manuscript in preparation). Even if identical experimental designs and statistical methods were used in these different studies, the biological inferences would undoubtedly differ, based on these differences in proteome coverage.

(2) Of the proteins that met the criteria for statistical significance chosen in each study, what proportion was unambiguously identified using mass spectrometry? In the studies reviewed, protein-of-interest identification rates ranged from 0 to 100% with a mean of 44.1% (median: 41.8%). Although explanations for these low rates were generally not given, they are likely a consequence of lack of sequence data for non-model organisms, variations in stringency of search criteria for bioinformatics (Kültz et al. 2007), technical limitations (e.g., for identifying post-translationally modified forms of proteins), limited abundance of proteins for tryptic digestion and detection, or limited access to instrument time. Regardless, only approximately two of every five proteins-of-interest are identified and interpreted biologically in recent environmental proteomics studies.

**How does analytical incompleteness influence biological conclusions in environmental proteomics?**

On an average, only a small fraction of the proteome is currently assessed in any given environmental proteomics study, and more than half of the significantly differentially expressed proteins in a given study remain unidentified. These findings represent unquantifiable, but certainly profound, limitations for biological interpretation of environmental proteomics data sets. For example, although a low rate of protein identification may not completely impede discovery (e.g., by validation of the roles of individual proteins), it will certainly hamper systems-level insight. Consider methods of downstream interpretation such as Ingenuity Pathway Analysis (Ingenuity® Systems, www.ingenuity.com). In some cases, failure to identify a single protein-of-interest could disrupt the algorithmic recognition of a biochemical pathway crucial to the organisms’ response to a given environmental challenge. A systematic analysis of such effects is beyond the scope of this review, but the implications warrant consideration.

**Meeting the challenges of analytical incompleteness**

Increasing proteome coverage (including increasing protein identification rates) is one of the major goals and one of the primary technological challenges in proteomics (Mann and Kelleher 2008), especially when attempting to apply these techniques to non-model organisms (Kültz et al. 2007). It should be emphasized that substantial progress has already been made and that current levels of proteome coverage far surpass anything previously possible. Very promising hypotheses have been generated using existing technologies and databases (Cravatt et al. 2007; see also contributions to this symposium). The possibility also exists of mining existing data sets as new information on sequences becomes available. However, the relatively low protein identification rates and their consequences (outlined earlier) constitute a strong argument for greater investment in generation and annotation of genomic sequence information and greater access to mass spectrometers among integrative biologists.
Concluding thoughts

Although the results of the literature review presented in this study are somewhat sobering, they in fact highlight the exceptional potential of environmental proteomics approaches. The simple recommendations outlined earlier in some cases may improve the design, analysis, and interpretation of environmental proteomics data sets. Numerous opportunities exist to further address these challenges and to make significant contributions to our understanding of organisms’ interactions with their environment, especially when interdisciplinary collaborations include comparative physiologists, biochemists, ecologists, evolutionary biologists, and biostatisticians.

One might reasonably ask whether the true challenge in environmental proteomics is philosophical, rather than experimental, statistical, or analytical. Rather than focusing on the biological and logistical tradeoffs in experimental designs, the sophisticated and standardized analysis of complex data sets, or the under-representation or lack of identification of particular groups of proteins in data sets, perhaps practitioners of proteomics should solely embrace the discovery-driven power of the approach. They might then focus attention and resources on following the most compelling of the many leads generated by any given proteomics experiment (although only seven of the reviewed studies included any orthogonal validation). Although this stance holds substantial practical merit (Kültz et al. 2007), particularly given current analytical limitations and pressures to generate biological stories for publication, I argue that widespread adherence to such a philosophical approach would hinder the development of true systems-level understanding. Sophisticated, global analyses that capitalize on the systems-level patterns emergent in complex proteomics data sets are likely to hold the greatest potential for generating novel biological insights. Only by meeting and overcoming the three challenges presented herein (and others) can practitioners of environmental proteomics make concrete steps toward realizing this vast potential.

Acknowledgments

The author thanks Lars Tomanek for the invitation and for the opportunity to open the symposium, which was sponsored by the SICB Division of Comparative Endocrinology, the Division of Comparative Physiology and Biochemistry, and the Division of Neurobiology. Kellen Flanigan provided invaluable assistance with the literature review. The ideas presented here were refined through discussions with D. Kültz and participants in the symposium, especially L. Tomanek, P. Fields, and B. Rees.

Funding

The author thanks the National Science Foundation [OISE-0713887 to W.D. and IOS-0542755 to D. Kültz] and the Partnership for Interdisciplinary Studies of Coastal Oceans [grant to G. Somero] for funding the proteomics studies of sharks and mussel A. This is PISCO contribution 419. The National Science Foundation and Loyola Marymount University provided travel support to participate in the symposium.

Supplementary material

Supplementary material is available at ICB online.

References


Challenges for environmental proteomics


Martyniuk CJ, Kroll KJ, Doperalski NJ, Barber DS, Denslow ND. 2010b. Genomic and proteomic responses