OXFORD

Briefings in Bioinformatics, 2017, 1–9

doi: 10.1093/bib/bbx128
Paper

Advanced bioinformatics methods for practical applications in proteomics

Wilson Wen Bin Goh and Limsoon Wong

Corresponding authors: Wilson Wen Bin Goh, School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551. Tel.: +65-6904-7149; E-mail: wilsongoh@ntu.edu.sg; Limsoon Wong, Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417. Tel.: +65-6516-2902; E-mail: wongls@comp.nus.edu.sg

Abstract

Mass spectrometry (MS)-based proteomics has undergone rapid advancements in recent years, creating challenging problems for bioinformatics. We focus on four aspects where bioinformatics plays a crucial role (and proteomics is needed for clinical application): peptide-spectra matching (PSM) based on the new data-independent acquisition (DIA) paradigm, resolving missing proteins (MPs), dealing with biological and technical heterogeneity in data and statistical feature selection (SFS). DIA is a brute-force strategy that provides greater width and depth but, because it indiscriminately captures spectra such that signal from multiple peptides is mixed, getting good PSMs is difficult. We consider two strategies: simplification of DIA spectra to pseudo-data-dependent acquisition spectra or, alternatively, brute-force search of each DIA spectra against known reference libraries. The MP problem arises when proteins are never (or inconsistently) detected by MS. When observed in at least one sample, imputation methods can be used to guess the approximate protein expression level. If never observed at all, network/protein complex-based contextualization provides an independent prediction platform. Data heterogeneity is a difficult problem with two dimensions: technical (batch effects), which should be removed, and biological (including demography and disease subpopulations), which should be retained. Simple normalization is seldom sufficient, while batch effect-correction algorithms may create errors. Batch effect-resistant normalization methods are a viable alternative. Finally, SFS is vital for practical applications. While many methods exist, there is no best method, and both upstream (e.g. normalization) and downstream processing (e.g. multiple-testing correction) are performance confounders. We also discuss signal detection when class effects are weak.

Key words: proteomics; networks; biostatistics; bioinformatics; biotechnology

Introduction

Proteomics, as the high-throughput study of proteins, is undergoing vast technological advances resulting in more efficient protein extraction, higher-resolution spectra acquisition and improved scalability. These have helped proteomics mature into an independent discovery platform. Notable examples include determination of the first draft human proteomes via high-resolution mass spectrometry (MS) [1, 2], demonstrating that MS-based technologies can independently identify a significant proportion of the translated products (proteins) from known genes (~80%; 17 294 for Kim *et al.* [1] and 15 721 for Wilhelm *et al.* [2], of ~20 000 genes) across a gamut of human tissues (including isoforms, with open accessibility to raw spectra). Such large-scale endeavours pave the way for cross-validating new data and investigating tissue-specific biology from a proteome-first perspective. Another example is the rise of big (proteomics) data because of the emergence of data-independent acquisition (DIA) [3], which leverages on sophisticated separation and high-resolution instruments to capture all detectable spectra within each analytical window. Although this resolves the semi-random preselection problem present in older proteomics paradigms (data-dependent acquisition; DDA),

Wilson Wen Bin Goh is a lecturer in the School of Biological Sciences, Nanyang Technological University.
Limsoon Wong is Kwan-Im-Thong-Hood-Cho-Temple Chair Professor of Computer Science at the National University of Singapore. He currently works mostly on knowledge discovery technologies and their application to biomedicine. He is a Fellow of the ACM, inducted for his contributions to database theory and computational biology. Some of his other awards include the 2003 FEER Asian Innovation Gold Award for his work on treatment optimization of childhood leukemias, and the ICDT 2014 Test of Time Award for his work on naturally embedded query languages.

Submitted: 14 July 2017; **Received (in revised form):** 11 September 2017

it creates another. Specifically, DIA spectra profiles do not have a direct one-to-one correspondence between precursor and fragmentation peptide ions, thereby complicating the process of obtaining good quality Peptide-Spectra Matches (PSMs). Even so, coupled with efficient protein extraction and shorter running times, DIA has rapidly gained dominance, and the first truly large proteomics data sets are emerging [4]. Note, there are variations of DIA, e.g. SWATH [5] and MS(E) [6].

These advances generate greater data volume, but quality can suffer, therefore creating new computational challenges. Simultaneously, traditional problems regarding coverage (i.e. inability to survey the entire proteome simultaneously) and consistency (i.e. different proteins are identified across different runs of the same sample, and different proteins are observed between different samples from the same experiment) persist.

Given these developments, it is timely to consider how bioinformatics evolve to meet these new challenges. Some notable achievements include a common data standard for proteomics is now widely adopted (mzML [7]), while mega open-access software (e.g. OpenMS [8]) provides unprecedented cross-hardware comparability and analytical flexibility. It is impossible to cover all new bioinformatics developments. So, we focus on four practical issues: peptide/PSM, missing-protein (MP) prediction, data heterogeneity and statistical feature selection (SFS).

Peptide/PSM: Genomics technologies provide direct sequence information per read. In contrast, MS data are merely an obscure series of peak intensities and mass-to-charge (mz) ratios, which must be mapped to peptides first. The mapping process is error-prone (e.g. incomplete fragmentation, mixed signals from multiple peptides and large numbers of potential matches per spectra, all lead to increased uncertainty). When confronted with several options, the best PSM may be wrong [9]. Notice we refer to peptides, not proteins, as proteins are predigested to facilitate ionization and detection. Therefore, identified peptides must be mapped to the parent protein: if unambiguously mappable, the PSM is retained as evidence of the parent protein's existence. Unfortunately, most PSMs do not map unambiguously, and are therefore discarded. Moreover, this procedure ignores splice variants (as only canonical full parent sequences are typically considered) [10, 11].

MP prediction: The human proteome project estimates that the protein products of ~20% genes are non-detected by MS [12, 13], while significant proportions are inconsistently observed on a routine basis because of difficulties in protein isolation and solubilization, sequence ambiguity, varied analysis algorithms and non-standard statistical thresholds. This results in irregular and irreproducible data. Given that many proteins are unreliably characterized via MS, orthogonal approaches are often required including antibody-based identification (for proteins lacking trypsin cleavage sites), subcellular/organelle enrichment and targeted-MS (e.g. selective or multiple reaction monitoring) [12]. Bioinformatics also has a role, e.g. missing-value imputation (MVI) provides estimates of values in 'data-holes', while network/pathway/protein complex-based analysis predicts the presence of completely undetected proteins.

Data heterogeneity—analysis of real human data is confounded by biological heterogeneity, e.g. disease subpopulations, demographics (age, race, gender, etc.) and technical heterogeneity, e.g. batch effects where samples are strongly correlated with non-phenotypic factors. Batch effect is an important confounder but seldom investigated in proteomics [14, 15].

Finally, identifying biomarkers (prognostics and classification) from proteomics data is accomplished via SFS where a quantitative metric (e.g. a test statistic or a P-value) is used to determine relevance (and therefore predictive power). Unfortunately, many SFS methods exhibit poor reproducibility: When a test is applied independently on two data sets of the same disease, the two lists of significant proteins lack agreement, partly because of misunderstanding and misinterpretation of the P-values [16]. Moreover, the erroneous assumption of independence amongst individual proteins also means that multiple-testing corrections (MTCs) overcorrect (lowering sensitivity), while including mutually correlated proteins in a signature (from same complex/pathway) is redundant and prevent other proteins (adding novel information) from inclusion.

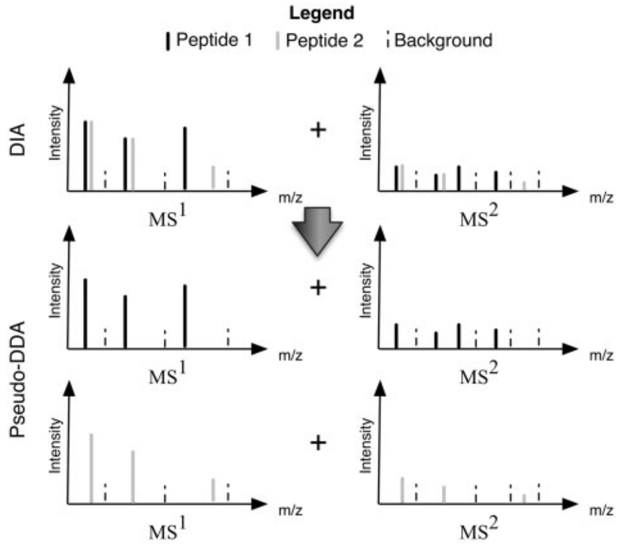

Figure 1. In DIA, an analytical window may comprise mixed signals from multiple peptides such that there is no fixed MS^1-MS^2 correspondence. This hinders sequence identification. A get-around is to decompose DIA spectra into pseudo-DDA spectra such that there is a fixed MS^1-MS^2 correspondence, amendable for use with well-established library search algorithms.

Peptide/PSM

In DDA, proteomics comprises a tandem set-up between two MSs. The first determines peptide (precursor or MS^1) masses within a unit of analytical time. If several peptides co-elute concurrently, then only one is preselected for subsequent fragmentation in a collision chamber followed by analysis in the second MS (MS^2). This set-up enforces a fixed MS^1-MS^2 correspondence, simplifying peptide identification. But preselection is semi-random such that across different runs (even across technical replicates for the same sample), peptides whose peaks co-elute simultaneously are reported inconsistently, resulting in different identification proteins each round.

DIA is a new class of brute-force spectra-acquisition strategies that eschew preselection. However, the PSM problem is harder, as DIA indiscriminately captures all precursor and fragment information within specific mz and retention time (rt) windows. An mz/rt window comprises peaks from multiple peptides, making disambiguation/disentanglement difficult (Figure 1).

A few strategies have been devised to resolve this. Group-DIA uses global information across samples (or runs), combining the elution profiles of precursor ions and their respective fragment ions, to determine exact precursor-fragment ion pairs [17]. In doing so, the paired data become pseudo-DDA, amendable to DDA library search algorithms. This approach requires individual runs to be made comparable from the onset, achieved by aligning retention times, and maximizing the correlation coefficient of the extracted ion chromatograms of the product ions. Pairing is achieved by selected fragment ions with high profile similarity to the precursor. False discovery rate (FDR) is calculated by random selection from unselected fragment ions. While combining runs increases peak assignment confidence, it also identifies false signals, as these exhibits limited inter-run reproducibility. The concept is sound, and takes advantage of higher scalability possible with DIA, thereby boosting sensitivity. Combining inter-run data has another benefit: by searching for consistent yet low-intensity signal, one may identify low-abundance proteins confidently.

Group-DIA is reported to outperform DIA-Umpire [18] (using the SWATH-MS Gold Standard data [19]). The authors reported that the more DIA data files used, the better Group-DIA became, even with different search engines and quantitative thresholds [17]. Although comparable with Open-SWATH [19], Group-DIA produced more consistent quantitation [17]. However, as Group-DIA relies on run alignment, it is vulnerable to real data noise and/or heterogeneity, making it difficult to align individual samples, skewed by extreme samples or loss of power with small sample size [17].

Alternatively, one may simply compare individual spectra (from known peptides) iteratively against DIA-spectra. In MSPLIT-DIA, annotated library spectra are compared against DIA-spectra, generating a list of potential Spectra-Spectra Matches (SSMs) [20]. Redundancy amongst SSMs is eliminated via pairwise comparisons, while statistical evaluation is based on decoys generated from randomly selected matches. MSPLIT-DIA's main advantage is sensitivity, and it can detect up to 10 peptides per DIA-spectra. When benchmarked on the SWATHAtlas spectral library [21], MSPLIT-DIA identified 66–89% times more peptides than DIA-Umpire per run [20]. However, as library spectra are compared iteratively against each DIA-spectra, it may be difficult to compute efficiently (although it appears amendable to parallel processing). Also, although FDR is typically fixed at 1% based on decoy estimations, actual numbers of false positives are ostensibly higher.

DIA-Umpire v1/2 is amongst the first DIA-search algorithms, and a standard against which newer ones are compared [18]. Although superseded, it does offer comprehensive workflows for various applications (signal extraction, untargeted identification, targeted extraction, etc.) Similar to Group-DIA, DIA-Umpire can be used for SSM, but it does not use information across individual runs to improve confidence. To match fragments to precursor, the Pearson correlation coefficient is calculated based on the chromatographic peak profiles provided they co-elute. Precursor-fragment pairs are modelled as a bipartite graph and filtered by a combination of thresholds, generating pseudo-DDA spectra compatible with DDA library-search methods. This is similar to GROUP-DIA's pseudo-DDA spectra generation method. An upgrade for signal extraction implements an improved feature-detection algorithm with two additional filters using isotope pattern and fractional peptide mass analysis. Targeted re-extraction is now implemented with a new scoring function and more robust, semi-parametric mixture modelling of the resulting scores for computing posterior probabilities of correct peptide identification.

Besides the aforementioned, there are also other emerging technologies, e.g. DIA with variable width windows such that each window captures roughly an equal number of precursor ions [22]. It is noteworthy that the technological landscape in proteomics changes rapidly.

MP prediction

MPs are proteins that are present in a sample but fail to be detected by the proteomic screen for various reasons (lack of unique PSMs, low abundance, etc.) Conventionally, an MP is one that has never been observed before in MS-based proteomics, but it is generalizable to include inconsistently detectable proteins (Figure 2A). MPs impede analytical efforts in comparative/clinical studies, and must be addressed.

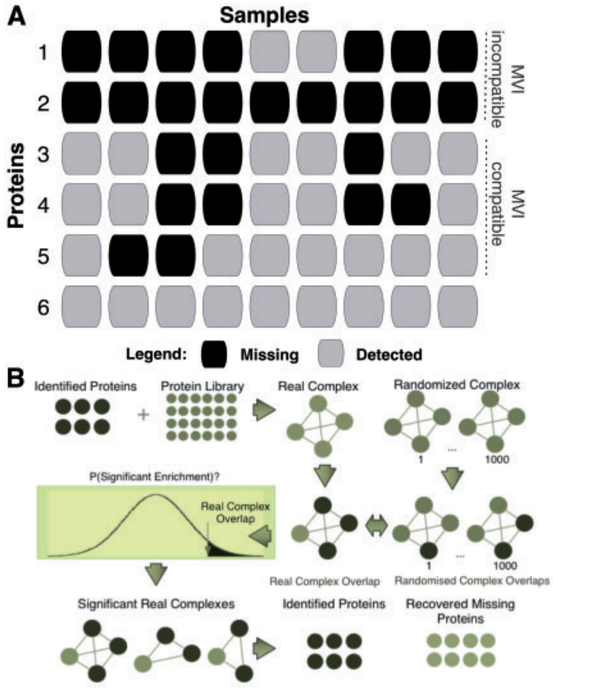

Figure 2. Missing proteins. (**A**) MVI is the process of predicting the value of a missing entry but requires that the protein is detected in at least one sample. However, prediction accuracy is ostensibly more unreliable if fewer samples are available as reference. (**B**) The complete absence of a protein in a proteomics screen does not mean it is not there, and it may fall beyond the limits of detections. We may use networks as a means of predicting presence via 'guilt-by-association'. The procedure shown above are the steps of the FCS method (see text for full description).

MPs may be resolved via experimental and technical procedures. Bioinformatics can also play important roles, providing two solution types, viz. MVI and network/pathway-based recovery cum deep-spectra mining.

MVIs are inferential methods and can be used if the MP is observed at least once in the data (Figure 2A). MVIs range from simple (where a missing value is replaced by a constant, or a randomly generated number), to local (where missing values are estimated based on protein expression profiles of other proteins with correlated intensity profiles) and to global [where missing values are estimated based on high-level data structures, e.g. principal components (PCs)]. In proteomics, MVIs are reportedly ineffective: using 3 MS-data sets (a cleaned but controlled dilution experiment, human clinical data with high heterogeneity within and between groups and mouse data with high homogeneity within experimental group) and 10 MVIs, Webb-Robertson *et al.* [23] concluded that local MVIs are better at accuracy, but no MVI consistently outperforms the others. As the actual missing values are known, it is found that MVIs have poor accuracy (the root-mean-square deviations are high). Thus, MVIs can mislead. Even so, it is statistically unsound to simply ignore missing values in general. Doing so can result in an overestimated mean protein abundance value calculated across biological replicates if the missing values were the result of some replicates having protein abundances below the MS sensitivity threshold. Also, imputation by simply using a constant value can lead to underestimation of the SD and type I errors.

MVIs are pure quantitative approaches, and do not leverage on biological context. Moreover, the MPs must have been observed at least once. What if the MPs are never observed? In such cases, we may look to networks/pathways/protein complexes. Network-based methods use biological information and can predict completely unobserved proteins. As proteins work together in functional units (as a complex or a module), MPs correlated with observed proteins falling within common complexes are more likely present [24, 25]. If more protein components from a complex are detected, the more likely the complex is formed, and therefore, the more likely the constituent MPs are present. Several methods leverage on this reasoning, e.g. functional class scoring (FCS) [26, 27]. In FCS, an overlap is calculated between the observed proteins and each complex. A random selection of proteins equals to the size of the complex is taken repeatedly (from a pool of proteins belonging to at least one complex), and a randomized overlap determined. As proteins in a complex are correlated, a true enrichment would be one that is significantly higher than randomly generated complexes where proteins are non-correlated. The empirical *P*-value is, therefore, the proportion of randomized samples having an overlap greater than the observed (Figure 2B). FCS is considerably more powerful than other network-based approaches such as Maxlink [28] and Proteomics Expansion Pipeline (PEP) [29], particularly in recall. FCS exemplifies the notion that biological reasoning/context can lead to powerful quantitative approaches (in this case, for MP prediction).

However, FCS has some limitations: the FCS *P*-value alone does not provide a means of ranking individual predicted proteins (based on relevance) nor is (1−*P*-value) the exact probability the MP is present. This is a generic issue because of the *P*-value, and not FCS *per se* [16, 30]. Determining the exact likelihood, an MP is present is an open problem, but it is possible to leverage on the joint probabilities of confidence from detected members from the same complex.

FCS predicts MPs independently of the spectra and therefore does not yield protein expression information. Therefore, it must be paired with spectra-mining to determine expression level [29]. If a peptide is present, signal from the molecular ion is almost always present in MS^1 but may be obfuscated by low signal-to-noise ratios, peak misalignments or PSM ambiguity (unsure which protein the observed peptide belongs to). While the spectra may be searched manually, automation is required for scalability [29]. It is possible to use targeted search approaches like DeMix-Q, which propagates information from runs with positive identification to runs where the peptides are reported absent [31]. Although DeMix-Q can be used standalone, it potentially returns large amounts of false positives if there is no prioritization/predetermination of search targets. As network/protein complex-based analysis directs the search towards better quality targets, the two may be integrated.

Data heterogeneity

Heterogeneity refers to variations uncorrelated with the factor of interest, e.g. a disease. High heterogeneity inevitably leads to bias, which makes findings irrelevant and irreproducible. We need to distinguish two forms of heterogeneity: technical (batch) and biological (class) (Figure 3A). Technical heterogeneity stems from use of specific technologies, or running conditions (batch) [32], whereas biological heterogeneity (class) arises from cohort demographics and aetiologies. Although both are confounding factors, the latter should be conserved (but is often removed by accident).

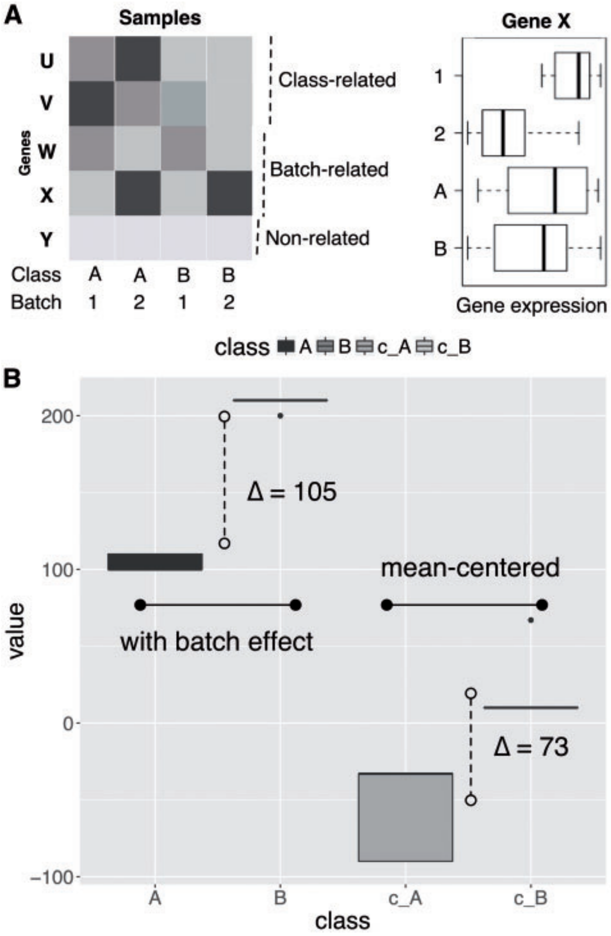

Figure 3. Heterogeneity in biological data. (**A**) Suppose, we have a data set with two classes, A and B, with two batches (technical replicates) 1 and 2, we may find that the variation of some genes correlates with class, while others correlates with batch or some other factor. In the case of gene X, it is clear that it is highly correlated with batch, and to a lower extent, with class effects. By chance, gene X can be wrongly selected during SFS, leading to analytical error. (**B**) Batch effects can cause erroneous estimation of effect size (true Δ = 100). In this example, presented as annotated boxplots, batch-representation imbalance in Classes A and B (because of poor experimental design) can create problems. Removing batch effects via mean-centering (c_A and c_B are batch-corrected Classes A and B, respectively) results in drastic underestimation of true effect size (see text for full description).

To some degree, heterogeneity can be minimized via normalization, i.e. the standardization of data across multiple samples. This is critical, as the choice of normalization method directly impacts SFS and downstream functional analysis (see next section). Unfortunately, most normalization methods are borrowed directly from genomics without considering proteomic idiosyncrasies. But this consideration is necessary: within a laboratory, the top two biases from proteomics data stem from rt and charge state; whereas between laboratories, the top biases stem from retention time, precursor m/z and peptide length [33]. These factors are proteomics-specific. Based on mock biomarker data, Rudnick *et al.* [33] described how these proteomics-specific factors can be used for developing a stepwise normalization procedure with highly beneficial effects [31].

There are also recent evaluations on what normalization procedure works well on proteomics data. Valikangas *et al.* [34]

benchmarked 11 normalization methods using three spike-in data sets and one experimental data set, based on the ability to (i) reduce variation between technical replicates, (ii) effect on differential protein expression analysis and (iii) effect on estimation of log-fold change. These are useful evaluation metrics, but not necessarily uncorrelated: reducing inter-technical replicate variation is a global metric; while it may mean irrelevant variation is removed, it can also mean that a large proportion of variation (useful and non-useful) are lost and data integrity is affected. A practical measure of functional outcome, especially for SFS (see next section), is to check the precision and recall. However, it is insufficient to simply know the set of differential features, as the magnitude of the effect size (i.e. expression levels) also matters. It is possible that the log-fold relationship changes unstably because of normalization, yet retain statistical significance.

Valikangas *et al.* suggested that the scarcely used variance stabilizing normalizationis best suited for proteomics data but tends to underestimate log-fold changes (effect size). An extremely important point raised was on the nature of the phenotypes being compared: most normalization methods assume that only a small portion of proteins are differentially expressed, and force the total intensity levels between samples (from different phenotypes) to be the same, e.g. z-scaling normalizes the gene expression/protein abundance distribution of each sample to the normal distribution. Unfortunately, this assumption usually does not hold in real samples. It is known that gene expressions in a disease sample are dissimilar to normal samples [35] and if violated, normalization creates false effects: especially in cancer, quantile normalization can reduce or remove true up-regulation relationships, and more severely, reversal [36].

Suppose gene expressions are similarly affected by heterogeneity (and inter-phenotype distributions are similar), simple normalization techniques (e.g. mean/median-centring, z-scaling and quantile) should work. However, in batch effect (when proteome measurements correlate with technical variables, e.g. time of experiment, technician/sample handler, reagent vendor and instrument), individual genes might be affected dissimilarly and, therefore, unresolvable via simple normalization. Oftentimes, on seemingly normalized data, individual genes susceptible to batch effect retain batch correlation and samples still cluster by processing date [15]. In proteomics, this problem should also exist. Moreover, if batch effect is suspected, detecting and addressing it is important and data analysis often benefits even from simple batch-mean centring, which assumes batch-effect uniformity [37].

Batch effects are commonly visualized via principal components analysis (PCA), but this multivariate approach can also be used for removing batch effect. For example, the top *n* PCs significantly correlated with batch are simply removed. The remaining PCs are then used as variables for feature selection and clustering [38]. However, this method can remove useful information, if batch and class effects are strongly confounded (i.e. in the same PC). Extending this PC-removal principle, the individual PCs can be scanned for variance correlated to batch, followed by removal, given a user-defined threshold (to control the amount of biological signal lost) [39]. The cleaned PCs are then recombined, and transformed back into the original data set. This approach, embodied in Harman [36], reportedly removes more batch noise and preserves more signal at the same time. This method makes intuitive sense, but an evaluation against established batch effect-correction algorithms (BECAs) has not yet been performed.

Established BECAs include ComBat [40] and surrogate variable analysis (SVA) [41]. ComBat is based on empirical Bayesian inference, and requires pre-specification of batch variable (which is not always known) but not formal indication of class variable (the sample phenotypes). Conversely, SVA requires class variable but not batch, which it estimates by first isolating class-associated variation and projecting the remaining variation into discrete PCs (termed surrogate variables, which are estimated batch variables). BECAs are error-prone: in ComBat, *P*-values are generally lower post-correction, with concomitantly higher false-positive rates, suggesting data integrity is compromised [38]. SVA recognizes that direct removal of variation from the data matrix reduces the actual degrees of freedom (making it more likely to generate false positives), and so, it does not directly return a batch effect-corrected data set [41]. Instead, the surrogate variables are saved separately as covariates, which are incorporated in downstream linear models for follow-up analysis, e.g. feature selection.

There are other tricky issues. For SVA, the sole preservation of class effects—at the expense of all else—loses valuable information: while we may pick out class-differential proteins post-SVA-correction, if their corresponding gene expression variability is further stratifiable based on secondary factors (e.g. age, gender and demographics), this information is lost [14]. Conversely, if class effect is false or many errors are made during class assignment (e.g. misdiagnosis), then SVA may amplify false effects. BECAs should be used carefully.

Less known is that BECAs should not be used on data with batch-design imbalances (i.e. the classes are unevenly distributed across batches) as the inter-batch class proportion differences can induce pseudo-batch effect. Depending on the BECA, the inter-batch class proportion differences deflate true class effect, or inflate false effect [42]. Suppose we have two batches where we have 5 subjects in Class A and 20 subjects in Class B in Batch 1, and 10 subjects in Class A and 5 subjects in Class B in Batch 2 (Figure 3B). Suppose everyone from Class A has a true value of 100 and everyone from Class B has a true value of 200. Then, the true class difference is 100. Suppose there is a batch effect such that everyone in Batch 1 gets a value 10 added to his true value. Then, the observed class difference considering both batches is $|(5 * 110 + 10 * 100)/15 - (20 * 210 + 5 * 200)/25| = 105$. Without normalization, the observed class difference is thus slightly magnified, causing false positives when the two batches are naively pooled to, e.g., increase sample size. Conversely, suppose as a normalization, we mean-centre each batch. Everyone from Class A in Batch 1 now gets the value $100-(5 * 110 + 20 * 210)/25 = -90$, and everyone from Class B in Batch 1 now gets the value $200-(5 * 110 + 20 * 210)/25 = 10$. However, everyone from Class A in Batch 2 now gets the value $100-(10 * 100 + 5 * 200)/15 = -33$, and everyone from Class B in Batch 2 now gets the value $200-(10 * 100 + 5 * 200)/15 = 67$. Then, the class difference observed post-normalization considering both batches is $|(5*-90 + 10*-33)/15 - (20 * 10 + 5 * 67)/25| = 73$. Thus, the observed class effect is diminished post-normalization, potentially causing false negatives when the two batches are pooled.

BECAs can be complex, and so one may consider alternatives such as batch effect-resistant normalization (BERN) [32]. These use ranks rather than absolute values (therefore, not making assumption on identical expression distribution) and fuzzification, which reduces fluctuations from minor rank differences and discards noise from rank variation in low-expression genes/proteins [16]. One BERN, Gene Fuzzy Scoring (GFS), is an unsupervised normalization approach that first sorts genes per

Table 1. Traits of an ideal statistical feature-selection (SFS) method

Trait	Definition
Reproducible/stable	The SFS should always give you the same set of features, given any data set comparing the same phenotype classes
Generalizable	The differential feature set must have high predictive power in any independent data set comparing the same phenotype classes and not outperformed by randomly selected features
Relevance	The feature set should be directly relatable to the phenotype; preferably as upstream as possible, and plays a real role in the phenotype as opposed to merely being correlated
Resistant to technical and other irrelevant sources of variability (noise)	This is a difficult one: the SFS should be resistant to unwanted non-class-specific effects, but at the same time, it should also preserve subpopulation effects (which look like non-class-specific effects). The difference being that the former are technical variations and the latter because of biological subtypes

sample based on their expression rank, assigns a value of 1 to a gene if it falls above an upper rank threshold, between 0 to 1 if it falls between the upper and low rank threshold and 0 if it falls below the lower rank threshold [35]. GFS exhibits strong reproducibility and selection of relevant biological features in genomics data. Combined with protein complexes, it exhibits high batch-effect resistance compared with other SFS methods [38]. GFS transformation is a crucial factor [43]: following GFS, even the typically poor-performing hypergeometric enrichment test improves dramatically in reproducibility across batches relative to non-GFS transformed data in SFS [44]. Moreover, individual checks on top-ranked protein complexes confirm specific association with phenotype class (not batch), and therefore, their constituent proteins are more likely clinically relevant.

Statistical feature selection

SFS is a large class of statistical methods with varied prerequisites (distribution, sample size, etc.) An ideal SFS should be reproducible, generalizable, selects relevant features and noise resistant (Table 1). While it is fashionable to evaluate and rank SFSs, in practice, there is no dominant SFS that works well across all circumstances. Confounding factors (e.g. mixed data distributions, between-class/within-class variability ratios, normalization and MTC) unpredictably affects SFS performance. Moreover, despite many options, the classical *t*-test works well across many scenarios [16].

We discuss two recent SFS evaluations by Langley and Mayr [45] and Christin *et al.* [46] on proteomics data. While they cannot be compared directly (different data sets, SFSs and evaluation metrics), they do introduce interesting and powerful methodologies. Christin *et al.*'s [46] approach relies on a single well-designed real data set built by spiking known concentrations of proteins to generate known true positives. By varying sample size, spiked protein concentrations and sample background, they can control intra- and inter-class variability, thereby generate a combination of test scenarios, with low inter-class and high intra–class variability and small sample size being the most challenging scenario. Using the f- and g-scores as scoring metrics, they concluded that when sample sizes are small, the univariate *t*-test and the Mann–Whitney U-test with MTCs perform badly, and when sample size increases beyond 12, provided inter-class variability is high, these classical methods outperform most methods. However, they are also highly sensitive to alterations in both inter- and intra-class variability. Multivariate methods—e.g. PC discriminant analysis and partial least squares discriminant analysis—leverage on higher-order data transformations and are less sensitive to these alterations but suffer from lower precision. Overall, they concluded that NSC (nearest shrunken centroid) offers the best compromise between recall and precision. The strength of this study lies in the reference data design, which provides a powerful means of simulating various test scenarios. However, the evaluations are potentially limited as the conclusions come from only one possible means of generating reference data (i.e. we do not know if the results will change given a second independent spiking experiment).

In contrast, Langley and Mayr [45] used *in silico* simulations on real data sets. The procedure involves taking proteomics data from a single class, random splitting into a pseudo-reference and test class and inserting effect sizes into randomly selected features in the latter. Across 2000 simulated data sets (1000 simulations from 2 data sets), their conclusions are more generalizable than Christin *et al.*'s [46]. They pointed out that all SFSs are essentially compromises (high precision but low recall; low precision but high recall), and none of the methods tested (including the *t*-test) could fully capture the differential landscape, even when inserted effect sizes were maximal (at 200% increment). However, they only evaluated univariate SFS methods. Data normalization/preprocessing [16], choice of MTC, choice of classifier and manner of *P*-value calculation (nominal or based on bootstrap) are additional confounding factors not examined by these works.

Moreover, these evaluations are based on the nominal null-hypothesis testing framework (where the null is a conservative statement denoting no differences between classes, and the alternative suggesting there is). The goal is to reject the null hypothesis at a predefined statistical threshold (usually 0.05 or 0.01) based on a theoretical (nominal) distribution. However, rejecting the null does not imply the alternative is true. For example, Venet *et al.* [47] suggested that signatures (a set of differential features) selected in this manner reveal little regarding phenotype association. Indeed, most random signatures are as good at predicting phenotype. Hence, it is imperative that selected features be checked for specific association with phenotype [47].

But failure does not lie solely in feature-selection approaches or statistical test paradigms. Proteomics-based quantitation is noisy, and idiosyncratic noise-eliminating procedures can improve performance. For example, Goeminne *et al.* [48] introduced an extension over traditional peptide-based linear regression models for estimating the true values of each protein.

First, let us express protein quantitation based on a linear regression model (Daly *et al.* [39] Clough *et al.* [22] and Karpievitch *et al.* [40]):

$$y_{ijklmn} = \beta_{ij}^{treat} + \beta_{ik}^{pep} + \beta_{il}^{biorep} + \beta_{im}^{techrep} + \epsilon_{ijklmn},$$

where y_{ijklmn} is the *n*th log-normalized signal intensity for the *i*th protein under the *j*th condition (treat), the *k*th peptide sequence (pep), the *l*th biological repeat (biorep) and the *m*th technical repeat (techrep) and ε_{ijklmn} a normally distributed error term with a mean of zero and variance $\sigma_i{}^2$. Each β denotes effect size for treat, pep, biorep and techrep for the *i*th protein, respectively.

Given the *i*th protein, the ordinary least squares estimate is defined as the parameter estimate that minimizes the loss function:

$$\sum_{jklmn} \epsilon_{ijklmn}^2 = \sum_{jklmn} \left(y_{ijklmn} - \beta_{ij}^{treat} - \beta_{ik}^{pep} - \beta_{il}^{biorep} - \beta_{im}^{techrep} \right)^2.$$

Goeminne *et al.*'s extension based on ridge regression shrinks regression parameters via penalization weights, and the ridge regression estimator is obtained by minimizing a penalized least squares loss function:

$$\min\left(\sum_{jklmn} \epsilon_{ijklmn}^2 + \lambda_i^{treat} \sum_j \left(\beta_{ij}^{treat}\right)^2 + \lambda_i^{pep} \sum_k \left(\beta_{ik}^{pep}\right)^2 + \lambda_i^{biorep} \sum_l \left(\beta_{il}^{biorep}\right)^2 + \lambda_i^{techrep} \sum_m \left(\beta_{im}^{techrep}\right)^2 \right),$$

where each λ is a ridge penalty for each estimated parameter β. If λs are generally positive, then estimators for β will decrease, thus reducing its variability (higher stability and accuracy). If evidence for β is sparse (e.g. many missing values), then it will also be corrected towards 0. Conversely, if evidence for β is strong (many observations), then λ encapsulates the sum of squared errors over these observations, suggesting more accurate estimation of β. The authors also reported that variability because of peptide effects is stronger than any of the other estimated parameters. This is consistent with what we know as well [10]. Evaluated on a CPTAC (Clinical Proteomic Tumor Analysis Consortium) data set, Goeminne *et al.* [48] suggested that, while computationally more complex, ridge regression stabilizes protein estimations with higher precision, but it is noteworthy there are also other methods that can be deployed as extensions including empirical Bayes, which stabilizes variance estimators and M-Huber weights, which reduces the impact of outlying peptide intensities.

Improving protein-level estimations improves feature selection but does not resolve collinearity issues (same-complex or same-pathway proteins are highly correlated and do not provide additional predictive information). Complex-based/network-based feature selection in proteomics is a new paradigm [49, 50], providing strong reproducibility and high phenotype relevance [24, 25]. However, this is also a new area. A key shortcoming is that the feature set is limited to known protein complexes.

An alternative is doing away with protein-based SFS: while it is intuitive to think in terms of proteins and their expression, in proteomics, this information is derived indirectly. Rather, it is the PSMs that are being analysed. The issue is that protein summarization relies on incomplete information and ignores splice variation. The incomplete component arises because only unique PSMs are retained and the remainder discarded. Splice variants are prevalent in real biological systems, and the consequent protein expression is a mixed function of its constituent splice-variants [10]. Consequently, protein-based summarization can be misleading, and may contribute towards poor SFS-reproducibility issues. We may circumvent this problem by performing SFS on MS^1 peaks or peptides, followed by functional analysis (mapping to specific splice forms, including differential but potentially ambiguous peptides) instead [10].

The examples discussed thus far assume that class-differentiating signal is easily detectable. SFS in itself is of limited utility if class-differentiating signal is weak (most variation is uncorrelated with class effects). Multivariate methods—e.g. PCA—can help. For example, SFS is applied on each PC such that even those signals from lower PCs (accounting for smaller proportion of total variation) can be isolated. A more radical approach involves injecting independent noise into the data set, such that those meaningful PCs that initially carry a small amount of variance now carry significantly more variance (because of the noise injection) [51]. In contrast, non-useful PCs are expected to continue carrying small amount of variance, uncorrelated with the injected noise. Gene set enrichment analysis (direct statistical testing of predefined gene sets for differential expression analysis) is yet another strategy [52], but is still inferior to most other network-based methods [43, 49].

Summary

Technological advancements in proteomics call for innovative solutions to new and old problems.

For PSM on new DIA data, two strategies have emerged: the first is transforming DIA spectra to pseudo-DDA spectra. The second involves brute-force searching each DIA spectra against known reference libraries iteratively.

MPs cannot be resolved satisfactorily via MVI, which is devoid of context. A better strategy is to incorporate biological information, e.g. using protein complexes for predicting MPs, followed by spectra-mining.

Data heterogeneity in proteomics is a difficult emerging problem. Standard normalization has limited utility in removing bias and, depending on assumptions, can introduce false effects. Technical variation (including batch effect) is traditionally countered through BECAs. But BECAs can be difficult to use, and may compromise data integrity. Alternatively, BERNs and complex-based methods may be used.

SFS is integral towards functional analysis. While many SFS methods exist, there is no best method. Evaluative frameworks usually fail to consider the confounding effects of upstream (normalization) and downstream (MTCs) data processing, which consequently affects SFS performance. In proteomics, thinking in terms of protein expression, as opposed to spectra peak or peptide intensities may not be the best option (as it is indirect information). Additionally, if class effects are small, creative multivariate approaches (based on PCs) are necessary.

Key Points

- New proteomics technologies create new data challenges that are solvable with bioinformatics.

- The MP problem is better resolved via contextualization based on protein networks/complexes.
- Resolving technical bias (batch effects) is a difficult emerging problem.
- SFS is confounded by both upstream and downstream data processing.

Funding

A Singapore Ministry of Education tier-2 grant (MOE2012-T2-1-061) to L. W.

References

1. Kim MS, Pinto SM, Getnet D, *et al*. A draft map of the human proteome. *Nature* 2014;**509**:575–81.
2. Wilhelm M, Schlegl J, Hahne H, *et al*. Mass-spectrometry-based draft of the human proteome. *Nature* 2014;**509**:582–7.
3. Egertson JD, Kuehn A, Merrihew GE, *et al*. Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods* 2013;**10**:744–6.
4. Guo T, Kouvonen P, Koh CC, *et al*. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat Med* 2015;**21**:407–13.
5. Gillet LC, Navarro P, Tate S, *et al*. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* 2012;**11**:O111 016717.
6. Plumb RS, Johnson KA, Rainville P, *et al*. UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun Mass Spectrom* 2006;**20**:1989–94.
7. Deutsch EW. Mass spectrometer output file format mzML. *Methods Mol Biol* 2010;**604**:319–31.
8. Bertsch A, Gropl C, Reinert K, *et al*. OpenMS and TOPP: open source software for LC-MS data analysis. *Methods Mol Biol* 2011;**696**:353–67.
9. Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol* 2010;**604**:55–71.
10. Goh WWB, Wong L. Spectra-first feature analysis in clinical proteomics—a case study in renal cancer. *J Bioinform Comput Biol* 2016;**14**:1644004.
11. Tavares R, Scherer NM, Ferreira CG, *et al*. Splice variants in the proteome: a promising and challenging field to targeted drug discovery. *Drug Discov Today* 2015;**20**:353–60.
12. Baker MS, Ahn SB, Mohamedali A, *et al*. Accelerating the search for the missing proteins in the human proteome. *Nat Commun* 2017;**8**:14271.
13. Paik YK, Jeong SK, Omenn GS, *et al*. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat Biotechnol* 2012;**30**:221–3.
14. Jaffe AE, Hyde T, Kleinman J, *et al*. Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis. *BMC Bioinformatics* 2015;**16**:372.
15. Leek JT, Scharpf RB, Bravo HC, *et al*. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**:733–9.
16. Wang W, Sue AC, Goh WW. Feature selection in clinical proteomics: with great power comes great reproducibility. *Drug Discov Today* 2017;**22**:912–18.
17. Li Y, Zhong CQ, Xu X, *et al*. Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nat Methods* 2015;**12**:1105–6.
18. Tsou CC, Avtonomov D, Larsen B, *et al*. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* 2015;**12**:258–64, 267, following 264.
19. Rost HL, Rosenberger G, Navarro P, *et al*. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* 2014;**32**:219–23.
20. Wang J, Tucholska M, Knight JD, *et al*. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nat Methods* 2015;**12**:1106–8.
21. Rosenberger G, Koh CC, Guo T, *et al*. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci Data* 2014;**1**:140031.
22. Zhang Y, Bilbao A, Bruderer T, *et al*. The use of variable Q1 isolation windows improves selectivity in LC-SWATH-MS acquisition. *J Proteome Res* 2015;**14**:4359–71.
23. Webb-Robertson B-JM, Wiberg HK, Matzke MM, *et al*. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res* 2015;**14**:1993–2001.
24. Goh WWB, Wong L. Integrating networks and proteomics: moving forward. *Trends Biotechnol* 2016;**34**:951–9.
25. Goh WWB, Wong L. Design principles for clinical network-based proteomics. *Drug Discov Today* 2016;**21**:1130–8.
26. Goh WWB, Sergot MJ, Sng JC, *et al*. Comparative network-based recovery analysis and proteomic profiling of neurological changes in valproic acid-treated mice. *J Proteome Res* 2013;**12**:2116–27.
27. Pavlidis P, Lewis DP, Noble WS. Exploring gene expression data with class scores. *Pac Symp Biocomput* 2002;474–85.
28. Goh WWB, Lee YH, Ramdzan ZM, *et al*. A network-based maximum link approach towards MS identifies potentially important roles for undetected ARRB1/2 and ACTB in liver cancer progression. *Int J Bioinform Res Appl* 2012;**8**:155–70.
29. Goh WWB, Lee YH, Zubaidah RM, *et al*. Network-based pipeline for analyzing MS data: an application toward liver cancer. *J Proteome Res* 2011;**10**:2261–72.
30. Goodman SN. A comment on replication, p-values and evidence. *Stat Med* 1992;**11**:875–9.
31. Zhang B, Kall L, Zubarev RA. DeMix-Q: quantification-centered data processing workflow. *Mol Cell Proteomics* 2016;**15**:1467–78.
32. Goh WW, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 2017;**35**(6):498–507.
33. Rudnick PA, Wang X, Yan X, *et al*. Improved normalization of systematic biases affecting ion current measurements in label-free proteomics data. *Mol Cell Proteomics* 2014;**13**:1341–51.
34. Valikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform* 2016, in press.
35. Belorkar A, Wong L. GFS: Fuzzy preprocessing for effective gene expression analysis. *BMC Bioinformatics* 2016;**17**(Suppl 17):540.
36. Wu D, Kang J, Huang Y, *et al*. Deciphering global signal features of high-throughput array data from cancers. *Mol Biosyst* 2014;**10**:1549–56.
37. Gregori J, Villarreal L, Mendez O, *et al*. Batch effects correction improves the sensitivity of significance tests in

spectral counting-based comparative discovery proteomics. *J Proteomics* 2012;**75**:3938–51.

38. Goh WWB, Wong L. Protein complex-based analysis is resistant to the obfuscating consequences of batch effects—a case study in clinical proteomics. *BMC Genomics* 2016;**18**(Suppl 2): 142.
39. Oytam Y, Sobhanmanesh F, Duesing K, *et al*. Risk-conscious correction of batch effects: maximising information extraction from high-throughput genomic datasets. *BMC Bioinformatics* 2016;**17**:332.
40. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**:118–27.
41. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 2007;**3**: 1724–35.
42. Nygaard V, Rodland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* 2016;**17**:29–39.
43. Goh WWB, Wong L. NetProt: complex-based feature selection. *J Proteome Res* 2017;**16**:3102–12.
44. Goh WWB. Fuzzy-FishNET: A highly reproducible protein complex-based approach for feature selection in comparative proteomics. *BMC Med Genomics* 2016;**9**(Suppl 3):67.
45. Langley SR, Mayr M. Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. *J Proteomics* 2015;**129**:83–92.
46. Christin C, Hoefsloot HC, Smilde AK, *et al*. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol Cell Proteomics* 2013;**12**:263–76.
47. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 2011;**7**:e1002240.
48. Goeminne LJ, Gevaert K, Clement L. Peptide-level robust ridge regression improves estimation, sensitivity, and specificity in data-dependent quantitative label-free shotgun proteomics. *Mol Cell Proteomics* 2016;**15**:657–68.
49. Goh WWB, Wong L. Advancing clinical proteomics via analysis based on biological complexes: a tale of five paradigms. *J Proteome Res* 2016;**15**:3167–79.
50. Goh WWB, Wong L. Evaluating feature-selection stability in next-generation proteomics. *J Bioinform Comput Biol* 2016;**14**: 1650029.
51. Giuliani A, Colosimo A, Benigni R, *et al*. On the constructive role of no in spatial systems. *Phys Lett A* 1998;**247**:47–52.
52. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; **102**:15545–50.