Journal of proteome research

Technical Note

pubs.acs.org/jpr

NetProt: Complex-based Feature Selection

Wilson Wen Bin Goh*,†,‡,§ and Limsoon Wong*,§,∥

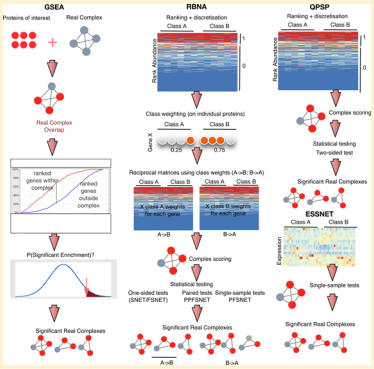
†School of Pharmaceutical Science and Technology, Tianjin University, 92 Weijin Road, Tianjin 300072, China

‡School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551

§Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417

∥Department of Pathology, National University of Singapore, 5 Lower Kent Ridge Road, Singapore 119074

ABSTRACT: Protein complex-based feature selection (PCBFS) provides unparalleled reproducibility with high phenotypic relevance on proteomics data. Currently, there are five PCBFS paradigms, but not all representative methods have been implemented or made readily available. To allow general users to take advantage of these methods, we developed the R-package NetProt, which provides implementations of representative feature-selection methods. NetProt also provides methods for generating simulated differential data and generating pseudocomplexes for complex-based performance benchmarking. The NetProt open source R package is available for download from https://github.com/gohwils/NetProt/releases/, and online documentation is available at http://rpubs.com/gohwils/204259.



KEYWORDS: *proteomics, bioinformatics, networks, feature-selection*

INTRODUCTION

Proteomics, as the high-throughput study of proteins in living systems, has advanced greatly. However, while it has achieved unprecedented proteome coverage in recent years, obtaining sufficient intersample consistency and quantitation accuracy for the purpose of selecting relevant proteins out of all observed proteins (feature selection) for the purpose of biomarker development and drug targeting remains challenging.[1] Continued efforts to improve scalability and quantitation accuracy are undoubtedly useful, but this only constitutes half of the problem. The other half lies in the inadequacies of contemporary statistics.[2] In general, conventional statistical methods falter given relatively small sample sizes, population heterogeneity, technical bias, and missing proteins (not to mention if their fundamental assumptions, e.g., independent and identically distributed (i.i.d.), are also violated). Therefore, statistically significant proteins in one study are nonreproducible in another and also lack predictive power (non-generalizable).[2]

Better designed analytical methods combine biological context with protein-based expression data (proteomics data).[3] Context encompasses prior knowledge on biological function, including gene groups,[4] biological complexes,[5,6] and biological networks.[7,8] Here gene groups refer to ensembles of genes or proteins where members are expressionally correlated and share common functions; physical interactions are implied but nonessential. Biological complexes (or protein complexes) refer to physical assemblies of proteins; physical interactions are expected, but exact binding configuration needs not be known, and biological networks refer to protein−protein interaction, regulatory, signaling, and metabolic[9−11] networks. As it turns out, real protein complexes (as opposed to predicted ones from a biological network) are very useful:[12] They are quite stable and are easily obtainable from centralized repositories, for example, CORUM[5,6] and MIPS.[13,14] They are also highly enriched for biology information. In particular, given a multitude of predictors (expression correlation, genetic interactions, etc.) in yeast, protein complexes are the best class predictors.[15] Even merging multiple predictors cannot improve upon protein complex-based predictions, suggesting that these other predictors do not add more information. These findings were independently confirmed in a large-scale analysis by Michaut et al., who used genetic knockout data from yeast spanning 191 890 genetic interactions between 4415 genes.[16] Our own experience[17] suggests likewise where missing-protein recovery based on real protein complexes are highly sensitive (∼95%) with decent precision (∼50%). In contrast, predicted complexes fall short, with the best outcome having a sensitivity of 45% with precision of 35%.[17] Protein-complex prediction algorithms, while diverse and sophisticated, do not meet practical requirements and cannot supersede real complexes.[18−20]

Aside from high precision and sensitivity, protein complex-based feature selection (PCBFS) is surprisingly robust against technical bias, particularly batch effects.[21−23] Batch effects are nonbiological variation and arise from confounding sources such as different experiment times, handlers, reagents, and instruments.[24] It is an extremely important problem for contemporary data analysis and is commonly resolved by

Received: June 1, 2017
Published: June 30, 2017

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112

batch effect-correction algorithms (BECAs) such as ComBat[25] and SVA.[26] However, BECAs can affect data integrity and introduce false-positives.[23] They can also be highly sophisticated but unwieldy in the hands of an amateur analyst.[26] In the case of SVA, class effects (i.e., phenotype) can be supplied to the algorithm, thereby explicitly protecting variation correlated to class. SVA identifies and removes surrogate variables uncorrelated with class. The downside of doing this is that overall variation is decreased concomitantly (alongside the actual degrees-of-freedom) such that any standard statistical test is more likely to produce false-positives.[26] Leek et al. have designed statistical tests to account for this and have offered an adjustment based on the F-test in their R package.[24,27,28] Nonetheless, people manipulating the batch-corrected data matrix may not know how to deal with this problem appropriately. (This issue applies to other BECAs as well.) Another problem is that BECAs may remove other interesting sources of variation (factors) if not explicitly spelled out to the algorithm. For example, if a class-differentiating protein, which has gender-dependent expression levels, is identified postbatch correction, we are likely only able to observe its class-differentiating effect but not gender effect.[26] Hence, it is unsurprisingly that when Jaffe et al. wrote about the limitations of SVA, they advised users to be very careful, understand SVA's limitations, and run iterative analyses, combining class with other potentially interesting factors (age, gender, race, etc.) as input. Unfortunately, this can be complicated, and, as with any sophisticated BECA, this issue is not limited to SVA alone. In contrast, as recently demonstrated, PCBFS is inherently batch-effect-resistant; that is, without using BECAs (and potentially affecting data integrity) and when tested on data where both batch and class effects exist, top protein complex-based features were strongly associated with class but not batch effects, while individual-protein features selected by parametric statistical approaches were strongly correlated with batch effects.[23]

A good diagnostic signature must be relevant to the phenotype and therefore ought not be outperformed by randomly generated signatures (i.e., random signature superiority). In gene-expression assays, this is a known problem,[29] but the same problem also exists in protein-expression assays. Protein complex-based scores, however, are quite robust against random signature superiority.[30]

Given these many useful properties, it may be beneficial for scientific investigators to consider PCBFS. Hence, we develop the R package, NetProt, which provides representative PCBFS methods from each of five current paradigms[31] and also some basic functionalities for benchmarking. We provide guidelines on when to use what, extensions to other useful R modules (for visualization and functional annotation), and some in-depth discussions on limitations and future development. Nonetheless, we note that PCBFS is not a panacea and has disadvantages as well. The most obvious limitation is that not all protein complexes are known, resulting in some functionalities being poorly represented. Fortunately, our collective knowledge on protein complexes (a.k.a. the complexome) is rapidly expanding given the emergence of new databases (e.g., Bioplex[32]), purification technologies,[33] and computational methods.[34] The second limitation arises from redundancy: Overlapping complexes within and across different databases remain open problems with no universally accepted solution. (A simple fix is to simply merge when two complexes share at least n common proteins,[35] but this may generate irrelevant *faux* complexes.[17,36]) Third, not all complexes are expected to be present in some tissue: Just as tissues have tissue-specific expressions, resulting in idiosyncratic cellular networks,[37] different tissues should also have unique tissue-specific complexomes, allowing us to match the appropriate protein-complex background against the data being studied.[38] Regardless, depending on the goal, users are free to supply real complexes, predicted complexes, and gene-annotation groups, of their choosing, to NetProt's methods. It is not compulsory to use only a fixed set of real complexes.

MATERIALS AND METHODS

Gene Fuzzy Scoring Normalization

NetProt provides Belorkar and Wong's Gene Fuzzy Scoring (GFS) as an independent function (gfs) for data processing on omics data without networks/complexes.[39] GFS is an example of a Signal Boosting Transformation (SBT), which are data normalization techniques with the following traits: boosting of high confident signals, penalization of low-confident signals and discretization of wide but nonuseful range of measured values.[2]

In GFS, an expression matrix is transformed by weighting individual variables (viz. genes or proteins) per sample based on expression ranks. GFS uses two thresholds, θ_1 (e.g., top 10%) and θ_2 (e.g., between top 10−20%). Variables ranked above θ_1 are assigned a weight of 1. Variables ranked between θ_1 and θ_2 are chopped into n equal-width intervals, and those in the same intervals are assigned the same interpolated weight between 0 and 1 (to account for measurements with inherent high variations). Variables with ranks below θ_2 are assigned 0 (and effectively ignored). For further details, refer to Belorkar and Wong.[39]

By itself, GFS transformation improves reproducibility of feature selection, with demonstrable robustness against batch effect.[39] It is also noteworthy that several NetProt methods (FSNET/PFSNET/PPFSNET[30,40] and QPSP[31]) improve upon GFS, providing additional power by taking advantage of autocorrelations among proteins within the protein complex,[30] with the added benefit of indirectly recovering low-abundance variables lost by GFS, as a significant complex may comprise both high- and low-abundance proteins[30,40] (see next paragraph).

Complex-based Feature-Selection Methods

We provide a quick introduction on the defining characteristics of each paradigm; cf. Figure 1. (For detailed descriptions, refer to Goh and Wong, 2016.[31]) Over-Representation Analysis (ORA) is a two-stage procedure: univariate feature selection on proteins followed by enrichment test. ORA can be highly unstable as it is very sensitive to the test type and stringency

Paradigm	Representative methods			
Over-Representation Analysis (ORA)	Hypergeometric Enrichment (HE)			
Direct Group (DG)	Gene Set Enrichment Analysis (GSEA)			
Hit-Rate (HR)	Quantitative Proteomics Signature Profiling (QPSP)			
Network Paired (NP)	Extremely Small Subnets (ESSNET)			
Rank-Based Network Analysis (RBNA)	SNET	FSNET	PFSNET	PPFSNET

Figure 1. Representative feature-selection methods in NetProt from each of five paradigms.

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112

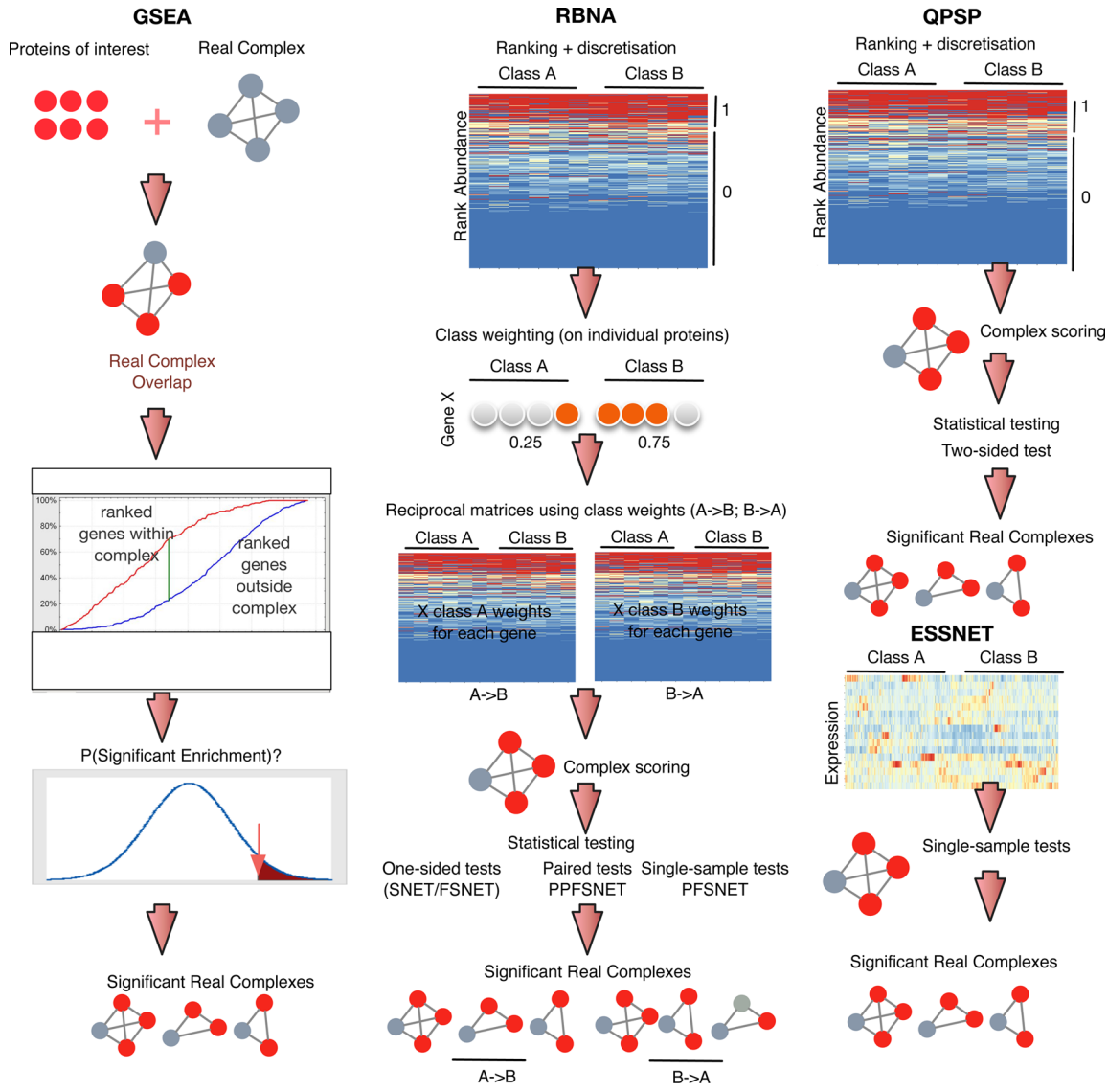


Figure 2. Overview on four methods included in NetProt (HE is not shown).

conditions in the univariate feature-selection step[30,41] as well as the appropriateness of the null hypothesis of the enrichment test.[42] As mitigation, Direct Group (DG) analysis does away with the univariate feature-selection step and directly determine whether a complex is differential by comparing the distribution of constituent protein expression between phenotype classes against that of proteins outside the complex. Hit-Rate (HR) analysis is initially conceived to address inconsistency issues in Data-Dependent Acquisition (DDA)-proteomics. Its defining characteristic involves calculation of overlaps (or hit-rates) among detected proteins against a vector of complexes to generate a hit-rate vector, which is used for class discrimination and feature selection.[36] Rank-Based Network Analysis (RBNA) incorporates rank weights based on the expression level of identified proteins and class weights based on the proportion of supporting information among samples,[30,40] while Network Paired (NP) analysis does not use any rank weights and work by testing the distribution of paired class differences on a given set of subnets or complexes.[42]

ORA is probably the most well known paradigm, and the Hypergeometric Enrichment (HE) pipeline is its most common realization. HE involves two steps: differential-protein selection via the two-sample t test, followed by enrichment analysis via the hypergeometric test (equivalent to a one-sided Fisher's exact test). While intuitive and simple, its key limitations are its hypersensitivity toward the upstream univariate protein-selection step[43] and its enrichment test's error-prone null hypothesis, which assumes proteins have mutually independent expression levels.[42]

DG avoids these shortcomings via global non-threshold-dependent evaluation. Given all observed proteins, it tests whether a protein complex is differential as a whole versus some constructed background. NetProt provides a vanilla implementation of Gene Set Enrichment Analysis (GSEA),[4] which is probably the best known DG method. GSEA first ranks all observed proteins based on effect size, for example, the t-statistic (if the comparison is between two classes). This is followed by the Kolmogorov–Smirnov (KS) test to determine if protein ranks in the complex and the ranks of proteins outside the reference complex arise from the same distribution. The significance of the KS test statistic is evaluated against a null distribution obtained by permuting sample class labels. For further details, refer to Subramaniam et al.[4]

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112

Table 1. Rough When-To-Use-What Guide

paradigm	when to use	when not to use
Over-Representation Analysis (ORA)	*General exploratory purposes	*Sample size is too small and therefore, the univariate test statistic is unreliable
	*There is some flexibility in the choice of the upstream feature-selection approach	*If reproducibility is a priority
	*When only a list of differential proteins—but not the full data matrix available	
Direct Group (DG)	*If we want to avoid general instability issues stemming from the upstream feature-selection approach	*If data is noisy
		*If individual sample scores per complex are required
Hit-Rate (HR)	*If the data matrix has high proportion of data holes, use PSP	*If data matrix is complete, then NP and RBNA are better options
	*If the data matrix is complete, then GFS can be used to produce differential top proteins per samples (QPSP)	
Network-Paired (NP)	*If an exhaustive list of complex-based features is desired (including low-abundance complexes)	*If individual sample scores per complex are required
Rank-Based Network Analysis (RBNA)	*If low-abundance complexes can be excluded	*If low-abundance proteins/complexes are of priority
	*An appropriate RBNA can be selected given specific experimental designs—e.g., if samples are pairable, use PPFSNET.	

Table 2. Advantages and Disadvantages of Each Complex-Based Feature-Selection Paradigm

paradigm	advantages	disadvantages
Over-Representation Analysis (ORA)	*Intuitive and simple two-step process involving a two-class feature-selection test first, followed by an enrichment test.	*Highly sensitive to the statistical stringency in the feature-selection test
		*Null hypothesis of the enrichment test is sometimes inappropriate
Direct Group (DG)	*Avoids instability issues by forgoing the upstream feature-selection test	*Highly sensitive to noise
Hit-Rate (HR)	*Can be used on data with large proportion of data holes	*Not as powerful as NP or RBNA
	*Resistant to noise	
Network-Paired (NP)	*Highly sensitive and stable	*May suffer from hyper-sensitivity issues if degrees-of-freedom or minimum overlap with complex is not determined appropriately
	*Works on small data sets	
	*Can be used to detect low-abundance complexes	
Rank-Based Network Analysis (RBNA)	*Highly sensitive and stable	*Low-abundance information is potentially lost
	*Works on small data sets	
	*Several flavors exist for different experimental designs	

There are two variants of HR: Proteomics Signature Profiling (PSP)[44] and its newer counterpart, quantitative PSP (QPSP).[36] In the latter, differential protein lists are derived by applying GFS such that each sample is defined by its top-ranked proteins. For PSP, instead of GFS, the data matrix is simply binarized (1 for present, and 0 for absent) before hit-rate vectorization (Figure 2).

Like QPSP, RBNAs also deploy GFS as the initial data-processing step. However, it goes one step further and defines a class-representation proportion (class weighting) per gene, given the rationale that a top-ranking gene frequently observed in the top *n*% of samples in its respective class is more likely a true-positive (Figure 2). Indeed, RBNAs are extremely powerful, greatly improving signal-to-noise ratios over a series of clinical blood cancer genomics data sets[45] and also display very high utility on proteomics data.[17,41] There are currently four RBNAs: SubNET (SNET),[45] Fuzzy SNET (FSNET), and paired FSNET (PFSNET)[40,45] and class-paired PFSNET (PPFSNET).[21] SNET and FSNET use the same one-sided reciprocal statistical test (Figure 2) but differ in the upstream data transformation: SNET uses a simple binarization procedure, while FSNET uses GFS. PFSNET uses GFS upstream but differs from FSNET by swapping the one-sided test in favor of a single-sample test. PFSNET uses unpaired tests and has reduced power if samples are pairable (e.g., the normal and disease tissues are derived from the same individual), PPFSNET addresses this shortfall by replacing the unpaired tests in PFSNET with the paired version. For detailed descriptions, please refer to Goh and Wong.[30,31]

NP is the newcomer among PCBFS paradigms, and there is only one representative method so far, Extremely Small SubNET (ESSNET),[42] which tests the distribution of paired class differences across the constituent proteins within a complex[42] (Figure 2). ESSNET is the only approach that explicitly considers low-abundance proteins.[42]

Choosing a Complex-Based Feature-Selection Method

We provide some rough guidelines on when to use certain methods (Table 1) and their key advantages/disadvantages (Table 2). NetProt provides vanilla implementations of ORA (HE) and DG (GSEA). Given recent benchmarking evaluations, we generally do not recommend use of ORA and DG, except for exploratory and comparative purposes, as they are unstable, even between technical replicates.[30,31,41] There are optimal conditions, however: For HE, performance improves if sample size is large;[30] for GSEA, performance improves if noise is minimal.[31]

Given a high proportion of data holes (as is typical in Data-Dependent Acquisition), PSP is ideal as it explicitly deals with data holes: Instead of trying to predict the values of the missing values via missing value imputation (MVI), PSP embraces data holes as informative. MVI is typically carried out using some statistical approach for estimating missing values based on observed values but is generally inaccurate and also inflates the degrees-of-freedom (DOF), potentially increasing both type I and II errors during feature selection.[2,46]

If the data matrix is mostly complete, then RBNAs and ESSNET are good options (QPSP is outperformed by these).[30,31] Among RBNAs, SNET and FSNET are superseded

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112

by PFSNET but are included nonetheless for benchmarking and evaluative purposes. ESSNET explicitly considers low-abundance proteins, with slight loss in stability.[31] Between PFSNET and ESSNET, the choice boils down to whether sample clustering is desired or inclusion of low-abundance proteins takes priority: ESSNET explicitly considers low-abundance proteins and provides a statistical evaluation of which complexes are significant given two classes but not individual sample scores. PFSNET provides the latter; therefore, the resultant data matrices can be used for clustering (but low-abundance proteins are generally ignored due to GFS).

Differential Data Simulation and Pseudocomplex Generation

Complex-based feature selection in proteomics is a relatively new area;[3,12] new paradigms and methods are expected to emerge, and appropriate approaches are needed for evaluation and benchmarking. NetProt provides an implementation (generate_proteomics_sim) of the univariate data simulation approach described by Langley and Mayr[47] (Figure 3A). On natural proteomics data with only one true class, the user chooses the proportion of randomly selected proteins to insert an effect size (p, which itself is also randomly sampled from a user supplied vector), which samples to assign into the pseudoclasses, and the number of simulated data sets to generate. The effect size is increased in only one of the pseudoclasses and is expressed as

$$SC_{i,j}{}' = SC_{i,j}{}^*(1 + p)$$

where $SC_{i,j}$ and $SC_{i,j}{}'$ are, respectively, the original and simulated spectral counts from the jth sample of protein i.

The Langley–Mayr approach, however, only simulates differential proteins but not complexes. Hence, NetProt also provides a method (generate_pseudo_cplx) for generating pseudocomplexes for protein complex-based feature selection from simulated differential data (Figure 3B). Differential proteins are reordered such that those with similar expression pattern are adjacent to each other. This reordered list is then split at regular intervals to generate differential pseudocomplexes (Users may also adjust the proportion of differential proteins, referred to as purity, in differential pseudocomplexes.) An equal number of nondifferential proteins are randomly selected, reordered on expression similarity, and split to generate nondifferential pseudocomplexes. To make it harder to detect a differential complex, constituent significant proteins are randomly selected and the effect size is removed. The proportion of significant proteins to remove is determined by the purity parameter (where 100% purity means the significant pseudocomplex is comprised solely of differential proteins). The differential and nondifferential pseudocomplexes can be used for precision-recall-based performance benchmarking on new data or on new feature-selection methods.

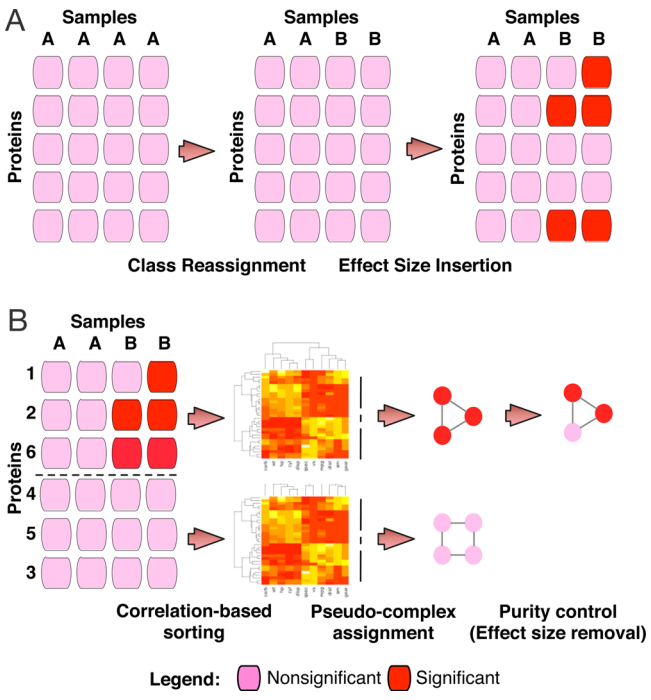


Figure 3. NetProt's differential data and pseudocomplex simulation methods. (A) Differential protein assignment. Samples initially are from one true class and randomly assigned into one of two pseudoclasses. Proteins are randomly selected in one pseudoclass, and an effect size is added. This process can be repeated many times to generate many simulated data sets. (B) Pseudocomplex generation: Following simulated differential features (cf. panel A), differential proteins are grouped based on expression correlation and further split into equal-sized pseudocomplexes. An equal number of nonsignificant proteins are randomly selected and assembled into pseudocomplexes similarly. To increase difficulty of detection of significant pseudocomplexes, significant proteins are randomly selected and the effect size is removed. The proportion of significant proteins to remove is determined by the purity parameter.

RESULTS

Using NetProt

NetProt adopts a modular approach, so that users have freedom to modify the methods according to their needs. (For details, refer to the NetProt documentation; http://rpubs.com/gohwils/204259.) For example, if users do not want to use GFS, they may replace GFS in favor of some other data-processing approach (e.g., quantile normalization) of their design or choice. If users do not want to use protein complexes, then they can simply use GFS as standalone.

The modular design also clarifies the logic behind some of the newer paradigms, for example, RBNAs. Broadly, NetProt's functions can be generally classified along the following levels: univariate-data transformation, class-proportion weighting, complex-based scoring, statistical test, and meta-methods. These are explained in Table 3.

NetProt functions (Figure 4, Table 3) are assembled linearly to produce each PCBFS method. NetProt's modular approach facilitates mix and match with functions from other R packages or linking to new methods, and because we place no restriction on the output, users are free to select how to further analyze the outputs using R's diverse clustering, visualization, annotation, and machine-learning packages (see next section).

As for input, NetProt takes in a standard protein-by-sample expression matrix and compares it against a list of complexes (based on CORUM's format). Code examples are provided in the online RPubs documentation (http://rpubs.com/gohwils/204259). NetProt does not provide automapping functions for identifiers, as it is often a clumsy and error-prone process. The identifiers used in the complex vector (whether real or predicted complexes) must be the same as those in the protein expression matrix. If identifier mapping is required, then the Biomart R package in Bioconductor (https://www.

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112

Table 3. Non-Exhaustive List of NetProt Functions Arranged by General Classification

function	explanation	used in
Univariate-Data Transformation		
gfs	*Applies Gene Fuzzy Scoring (GFS) onto a data vector. *Use in conjunction with R's apply function to apply it onto a data matrix	FSNET/PFSNET/ PPFSNET/QPSP
gnfs	*Discretizes data such that the top 20% takes on a value of 1, and the bottom 80%, 0. *Simplification of GFS. No fuzzification.	SNET/PSP
binarize_mat	*Binarizes a data sample such that nonmissing values are assigned a value of 1, and missing values (NAs) are assigned a value of 0. If users want to consider only the top *n*% proteins per sample, then use gnfs	PSP
Class-Proportion Weighting		
subset_weights	*Applies a class-based weight refinement based on the degree of supporting evidence from other samples of the same class	SNET/FSNET/ PFSNET/ PPFSNET
Complex-based Scoring		
internal_subtraction	*For each gene, generates a series of deltas by subtracting the expression scores of every sample in class A against a sample in class B	ESSNET
fsnet	*For each complex, calculates a set of complex-scores via reciprocal comparisons (class A→ class B and class B → class A)	SNET/FSNET/ PFSNET
Statistical Test		
standard_t_test	*performs a standard two sample *t* test based on the reciprocal matrices generated by fsnet	SNET/FSNET
pfsnet_theoretical_t_test	*performs a one-sample *t* test based on the reciprocal matrices generated by fsnet	PFSNET
ppfsnet_theoretical_t_test	*performs a paired *t* test based on the reciprocal matrices generated by fsnet	PPFSNET
essnet	*performs a one-sample *t* test based on the delta matrix generated by internal_substration	ESSNET
Meta-Methods		
generate_proteomics_sim	*Generates simulated two class data with randomly insert differential features at the protein level	any
generate_pseudo_cplx	*Takes the output from generate_proteomics_sim, and converts the protein features into complex-based features	any

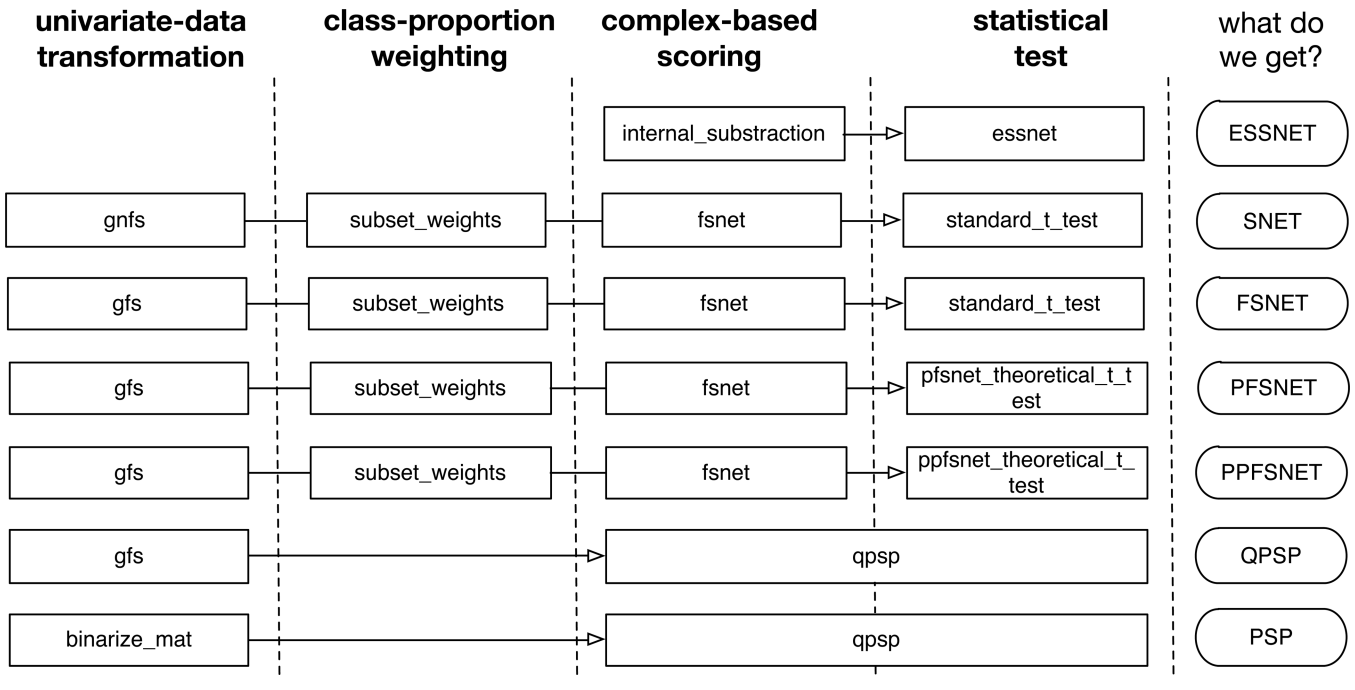


Figure 4. Chaining NetProt functions to obtain each complex-based feature-selection method. NetProt functions (rectangular boxes; cf. Table 2) are systematically grouped along the categories of data transformation to statistical test in a linear fashion (directed arrow). Many protein complex-based feature-selection methods use common functions. NetProt's modular approach makes it easy for users to mix and match with functions from other R packages or to design their own to create new methods. (HE and GSEA are excluded here; also not shown are functions from meta-methods.)

bioconductor.org/packages/release/bioc/html/biomaRt.html) is a viable option.

Extension 1 Clustering, Classification, and Visualizing Output

The R implementation of NetProt does not provide options for clustering, classification, and visualization as users can directly use many R packages for such tasks. With the exception of ESSNET, because the data output is typically a data matrix (sample by complexes), many analytical tasks are possible (using R's well-developed libraries).

A simple but common task is to test whether PCBFS-selected features correlate well with known sample classes. This is achievable using any variety of unsupervised clustering methods. We particularly liked R's **pheatmap** package, which combines unsupervised hierarchical clustering directly with

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112

heatmap visualization. In Figure 5A, we applied SNET on the renal cancer data set of Guo et al.[48] and CORUM complexes[5,6]

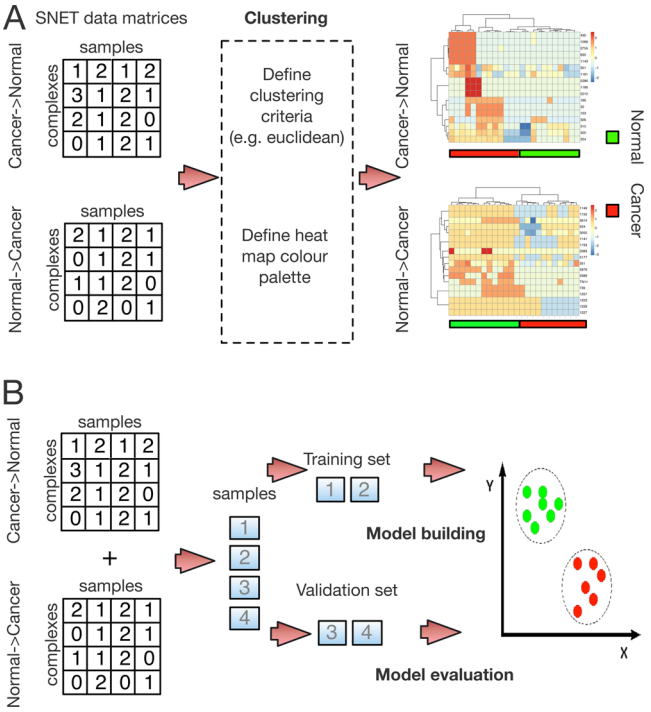


Figure 5. Clustering and classification with NetProt (SNET) data matrices. (A) Clustering and visualization using R's pheatmap package. (B) Standard cross-validation procedure for determining whether the selected features have any diagnostic power.

to obtain a pair of reciprocal matrices. These matrices, in turn, can be clustered to determine whether the samples segregate perfectly into their known respective classes. If so, then we may claim that PCBFS has been somewhat successful, as the selected complexes exhibit good class-discriminatory power. We may further evaluate if the selected features provide stable discrimination via R's bootstrap clustering tool, pvclust.[49]

Alternatively, if we are interested to know whether the selected features have any diagnostic power, we may combine and input the SNET data matrices, feed part of it into a machine learning method (e.g., R's eBayes package is an implementation of the naïve Bayes model) for training, and confirm its predictive power (classification) on the remaining samples not used in model training (Figure 5B).

Extension 2 Functional Characterization

Often, functional analysis requires linking functionalities to the collective set of selected features. On individual features, this is typically achieved by performing the hypergeometric test on a given set of differential proteins against a background of known proteins for each known functional term. Unfortunately, functional annotation strategies based on protein complexes are lacking. Gene Ontology is organized along the lines of individual proteins, not complexes. A crude workaround is to dismantle significant complexes into a protein list and feed it into a protein functional annotation platform, for example, DAVID[50] or GO-Term-Finder.[51]

Currently, CORUM does provide basic protein complex functional annotations, but these are quite limited, and there is room to develop computational approaches for automatic complex-based annotation. It is surprising that although protein complexes are physical assemblies, they are quite poorly annotated, but PCBFS does not absolutely require protein complexes as input. GSEA gene groups are well-annotated assemblies of genes; they may be used in conjunction or in place of protein complexes if functional characterization is the main objective.

There is further development in complexome research, where protein complex-based networks (the physical and functional interactions between complexes instead of individual proteins) are being developed. This is a logical generalization

A	Dataset	Advantages	Disadvantages
Simulated	D1.2/D2.2/	Sound effect size simulation strategy Generated from real data Large variety of simulation scenarios Limited/no batch effects	Small sample sizes (n= 6 and 8) Only five effect size possibilities are incorporated Incorporation of effect sizes do not obey protein-protein correlations Protein labels are removed
	DRCC	Larger sample size (n=12) Sound effect size simulation strategy Generated from real data Protein labels present	Only five effect size possibilities are incorporated Incorporation of effect sizes do not obey protein-protein correlations Batch effects present
Real	RCC	Moderate sized control (n=12) Useful for evaluating false positive rates	May have mild batch effects due to 3 technical replicates Sampling size too small for effective resampling statistics
	RC	Moderate sized (n=24) Technical duplicates ---useful for reproducibility test Well-controlled batch effects High quality control checks Sample classes are pairable	Sample size is too small for effective resampling statistics Quantitation platform not tried and tested
	CR	Large sample size (n=120) Quantitation platform well-established High quality control checks (internal checks for each run)	Controls (n=30) are much less than test states (n=90) Multiple confounding factors may be present; Sample classes are not pairable Strong batch effects are likely present

B	Dataset	Advantages	Disadvantages
Real complexes	CORUM	Enriched for biological signal	Incomplete set hence limited functional representation Complex-complex interactions not examined
	Complex Census	Larger coverage (includes predicted complexes) Wider functional representation	Predicted complexes may not be real Complex-complex interactions examined

Figure 6. Considerations on potential benchmarking data set. (A) Simulated and real protein-expression data sets. (B) Real complex databases.

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112

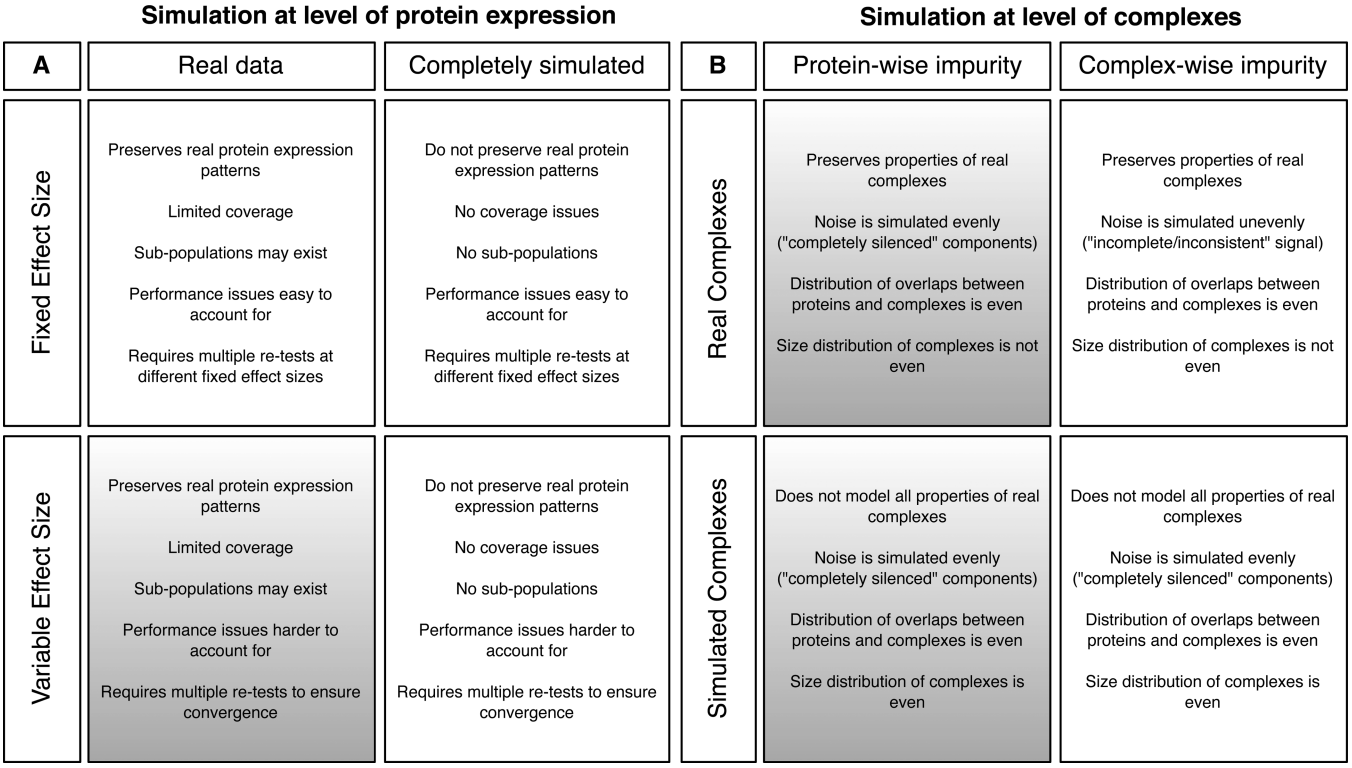


Figure 7. Considerations for differential feature simulation at the protein and complex levels.

(as proteins work in assemblies rather than individually) and helps to eliminate a high amount of redundancies associated with the analysis of individual proteins. Unfortunately, this is currently better developed for model organisms, for example, *Drosophila*,[52] than in humans. Regardless, such advances are invaluable for helping to understand the interconnections and relationships among a set of differential complexes and can help in functional/mechanistic investigations. As human complex-based networks become available, we may use PCBFS for helping identify the components of the complex-based networks that are altered in response to a disease.

DISCUSSIONS

Data Sets for Benchmarking

NetProt can be used for analysis of real data sets, but we think it also provides a shared opportunity for developing new benchmarks and PCBFS methods. Previously, we proposed that a PCBFS method can be broken up along the lines of a data-preprocessing transformation, weight assignment, complex-scoring method, and a statistical test (Figure 4). Although not every method needs to follow this design exactly, we expect most approaches can be broadly organized as such (even ESSNET and QPSP; Figure 4).

Another key consideration is which data sets are suitable for use as benchmark data. On proteomics data, a simple broad categorization would be simulated or real data (Figure 6A). NetProt includes three real data sets, RCC, RC, and CR. Simulated data sets D1.2/D2.2 and DRCC are available online at https://github.com/gohwils/NetProt/releases/tag/0.1/Data.zip. Every data set has a unique set of pros and cons, and we have listed these in Figure 6A. It is important that users understand these idiosyncrasies before using the data sets (especially the presence of batch effects in some of the data). Incidentally, batch-effect resistance is a useful property of some PCBFS methods. The complexome (or complex list) does change over time, and so NetProt does not include this internally. While an example complex vector is available internally in the package, we advise users to get the latest versions from CORUM itself. If CORUM is deemed too limiting, then another alternative is to derive the protein complex list from the complex census,[53] which offers wider coverage but may be less clean due to the incorporation of predicted protein complexes. Annotated gene sets may be obtained from GSEA.[4] Finally, if interest is in testing against predicted protein complexes, there is a wide variety of biological network databases (e.g., HPRD,[54−56] BioGrid,[57] etc.) and protein complex-prediction algorithms.[20]

Considerations for Data Simulations

With simulated data, it is necessary to generate differential features for the purpose of method evaluation. At the level of protein expression, there are two dimensions to consider. The first is the origin of the data: completely simulated data or building up from real data (Figure 7A). The second is concerned with how to simulate the magnitude of the differential effects, whether the effect size is fixed or variable. Completely simulated data with fixed effect sizes have very conserved behavior and are easily tractable but fall short of resembling real data. Alternatively, the more complex the simulations, building on top of the behaviors of real data, with variable effect sizes inserted, the more likely unforeseen confounding factors may creep in, with indeterminate effects on the performance of the feature-selection method. Figure 7A lists primary considerations along these two dimensions.

Because we are interested in PCBFS, it is also important to think how to simulate protein complexes. There also are two dimensions to consider. The first is whether to use real or simulated complexes. The second is how to simulate noise/impurity in the complex, making it more challenging for PCBFS

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112

methods to detect differential effects, if they exist (Figure 7B). It is also important to note that completely simulated data sets would not work with real complexes, as the selection of proteins is random and not based on inherent correlations among same-complex proteins. If the decision is to use real complexes, then the insertion of differential effects must be conserved at the level of complexes; that is, a complex is randomly selected to be differential, and the effect size inserted into its constituent proteins, but such approaches must also take into account the fact that many complexes share proteins, and so this approach can also artificially cause nonselected complexes to also become significant consequently. An alternative approach is to simulate effect sizes at the level of proteins first and then assemble them into differential complexes with no overlaps; however, such pseudocomplexes may lack realism and do not have many of the inherent properties of true complexes. An extensive listing of the considerations involved in protein complex-based simulations is discussed in Figure 7B.

Limitations of NetProt and Future Work

PCBFS is one of three important quantitative applications in network-based proteomics, the other two being class prediction (for diagnostics) and coverage expansion (for predicting missing proteins).[12] For PCBFS to become more powerful, two further advances are required: novel annotation strategies at the level of protein complexes and reference protein complex-based networks for understanding the functional relationships among differential protein complexes.

Currently, the R implementation of NetProt does not provide proprietary visualization options and dynamic output. It requires users to be familiar with the manipulation of data matrices as inputs to various data visualization packages, for example, pheatmaps. However, gaining even basic familiarity with R's basic data structures opens the way toward its powerful library of bioinformatics software and tools.

NetProt is currently in its first version of development and includes a limited set of representative methods from each PCBFS paradigm. We expect future upgrades to incorporate more PCBFS methods. We also hope to successfully develop and incorporate appropriate complex-wise functional annotation tests in the near future.

AUTHOR INFORMATION

Corresponding Authors

*W.W.B.G.: E-mail: goh.informatics@gmail.com.
*L.W.: E-mail: wongls@comp.nus.edu.sg. Tel: +65-65162902.

ORCID

Wilson Wen Bin Goh: 0000-0003-3863-7501

Notes

The authors declare no competing financial interest.
NetProt is free software and distributed under the terms of the GNU General Public License as published by the Free Software Foundation, version 3. The R source file is available at https://github.com/gohwils/NetProt/releases/. Online documentation is available at http://rpubs.com/gohwils/204259.

ACKNOWLEDGMENTS

W.W.B.G. and L.W. thank Abha Belorkar, Kevin Lim, and Donny Soh for ideas contributed towards elements of this work. W.W.B.G. is supported by an education grant (290-0819000002), Tianjin University, China.

REFERENCES

(1) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. 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* (6), O111.016717.

(2) Wang, W.; Sue, A. C.; Goh, W. W. Feature selection in clinical proteomics: with great power comes great reproducibility. *Drug Discovery Today* **2017**, *22* (6), 912−918.

(3) Goh, W. W.; Wong, L. Design principles for clinical network-based proteomics. *Drug Discovery Today* **2016**, *21* (7), 1130−1138.

(4) Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (43), 15545−50.

(5) Ruepp, A.; Brauner, B.; Dunger-Kaltenbach, I.; Frishman, G.; Montrone, C.; Stransky, M.; Waegele, B.; Schmidt, T.; Doudieu, O. N.; Stumpflen, V.; Mewes, H. W. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* **2007**, *36* (Database issue), D646−50.

(6) Ruepp, A.; Waegele, B.; Lechner, M.; Brauner, B.; Dunger-Kaltenbach, I.; Fobo, G.; Frishman, G.; Montrone, C.; Mewes, H. W. CORUM: the comprehensive resource of mammalian protein complexes−2009. *Nucleic Acids Res.* **2010**, *38* (suppl_1), D497−501.

(7) Rolland, T.; Tasan, M.; Charloteaux, B.; Pevzner, S. J.; Zhong, Q.; Sahni, N.; Yi, S.; Lemmens, I.; Fontanillo, C.; Mosca, R.; Kamburov, A.; Ghiassian, S. D.; Yang, X.; Ghamsari, L.; Balcha, D.; Begg, B. E.; Braun, P.; Brehme, M.; Broly, M. P.; Carvunis, A. R.; Convery-Zupan, D.; Corominas, R.; Coulombe-Huntington, J.; Dann, E.; Dreze, M.; Dricot, A.; Fan, C.; Franzosa, E.; Gebreab, F.; Gutierrez, B. J.; Hardy, M. F.; Jin, M.; Kang, S.; Kiros, R.; Lin, G. N.; Luck, K.; MacWilliams, A.; Menche, J.; Murray, R. R.; Palagi, A.; Poulin, M. M.; Rambout, X.; Rasla, J.; Reichert, P.; Romero, V.; Ruyssinck, E.; Sahalie, J. M.; Scholz, A.; Shah, A. A.; Sharma, A.; Shen, Y.; Spirohn, K.; Tam, S.; Tejeda, A. O.; Trigg, S. A.; Twizere, J. C.; Vega, K.; Walsh, J.; Cusick, M. E.; Xia, Y.; Barabasi, A. L.; Iakoucheva, L. M.; Aloy, P.; De Las Rivas, J.; Tavernier, J.; Calderwood, M. A.; Hill, D. E.; Hao, T.; Roth, F. P.; Vidal, M. A proteome-scale map of the human interactome network. *Cell* **2014**, *159* (5), 1212−26.

(8) Yook, S. H.; Oltvai, Z. N.; Barabasi, A. L. Functional and topological characterization of protein interaction networks. *Proteomics* **2004**, *4* (4), 928−42.

(9) Bensimon, A.; Heck, A. J.; Aebersold, R. Mass spectrometry-based proteomics and network biology. *Annu. Rev. Biochem.* **2012**, *81*, 379−405.

(10) Goh, W. W.; Lee, Y. H.; Chung, M.; Wong, L. How advancement in biological network analysis methods empowers proteomics. *Proteomics* **2012**, *12* (4−5), 550−63.

(11) Goh, W. W.; Wong, L. Networks in proteomics analysis of cancer. *Curr. Opin. Biotechnol.* **2013**, *24* (6), 1122−8.

(12) Goh, W. W.; Wong, L. Integrating Networks and Proteomics: Moving Forward. *Trends Biotechnol.* **2016**, *34* (12), 951−959.

(13) Mewes, H. W.; Frishman, D.; Mayer, K. F.; Munsterkotter, M.; Noubibou, O.; Pagel, P.; Rattei, T.; Oesterheld, M.; Ruepp, A.; Stumpflen, V. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* **2006**, *34* (90001), D169−72.

(14) Mewes, H. W.; Amid, C.; Arnold, R.; Frishman, D.; Guldener, U.; Mannhaupt, G.; Munsterkotter, M.; Pagel, P.; Strack, N.; Stumpflen, V.; Warfsmann, J.; Ruepp, A. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic acids research* **2004**, *32* (90001), D41−4.

(15) Fraser, H. B.; Plotkin, J. B. Using protein complexes to predict phenotypic effects of gene mutation. *Genome biology* **2007**, *8* (11), R252.

(16) Michaut, M.; Baryshnikova, A.; Costanzo, M.; Myers, C. L.; Andrews, B. J.; Boone, C.; Bader, G. D. Protein complexes are central in the yeast genetic landscape. *PLoS Comput. Biol.* **2011**, *7* (2), e1001092.

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112

(17) Goh, W. W.; Sergot, M. J.; Sng, J. C.; Wong, L. Comparative network-based recovery analysis and proteomic profiling of neurological changes in valproic Acid-treated mice. *J. Proteome Res.* **2013**, *12* (5), 2116−27.
(18) Yong, C. H.; Wong, L. From the static interactome to dynamic protein complexes: Three challenges. *J. Bioinf. Comput. Biol.* **2015**, *13* (2), 1571001.
(19) Yong, C. H.; Wong, L. Prediction of problematic complexes from PPI networks: sparse, embedded, and small complexes. *Biol. Direct* **2015**, *10* (1), 40.
(20) Srihari, S.; Yong, C. H.; Patil, A.; Wong, L. Methods for protein complex prediction and their contributions towards understanding the organisation, function and dynamics of complexes. *FEBS Lett.* **2015**, *589* (19), 2590−2602.
(21) Goh, W. W.; Wong, L. Class-paired Fuzzy SubNETs: A paired variant of the rank-based network analysis family for feature selection based on protein complexes. *Proteomics* **2017**, *17*, 1700093.
(22) Goh, W. W.; Wang, W.; Wong, L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* **2017**, *35* (6), 498−507.
(23) Goh, W. W.; Wong, L. Protein complex-based analysis is resistant to the obfuscating consequences of batch effects — A case study in clinical proteomics. *BMC Genomics* **2017**, *18* (Suppl 2), 142.
(24) Leek, J. T.; Scharpf, R. B.; Bravo, H. C.; Simcha, D.; Langmead, B.; Johnson, W. E.; Geman, D.; Baggerly, K.; Irizarry, R. A. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **2010**, *11* (10), 733−9.
(25) Johnson, W. E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8* (1), 118−27.
(26) Jaffe, A. E.; Hyde, T.; Kleinman, J.; Weinbergern, D. R.; Chenoweth, J. G.; McKay, R. D.; Leek, J. T.; Colantuoni, C. Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis. *BMC Bioinf.* **2015**, *16*, 372.
(27) Parker, H. S.; Leek, J. T.; Favorov, A. V.; Considine, M.; Xia, X.; Chavan, S.; Chung, C. H.; Fertig, E. J. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics* **2014**, *30* (19), 2757−63.
(28) Leek, J. T.; Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* **2007**, *3* (9), e161.
(29) Venet, D.; Dumont, J. E.; Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* **2011**, *7* (10), e1002240.
(30) Goh, W. W. B.; Wong, L. Evaluating feature-selection stability in next-generation proteomics. *J. Bioinf. Comput. Biol.* **2016**, *14* (5), 1650029.
(31) Goh, W. W. B.; Wong, L. Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms. *J. Proteome Res.* **2016**, *15* (9), 3167−3179.
(32) Huttlin, E. L.; Ting, L.; Bruckner, R. J.; Gebreab, F.; Gygi, M. P.; Szpyt, J.; Tam, S.; Zarraga, G.; Colby, G.; Baltier, K.; Dong, R.; Guarani, V.; Vaites, L. P.; Ordureau, A.; Rad, R.; Erickson, B. K.; Wuhr, M.; Chick, J.; Zhai, B.; Kolippakkam, D.; Mintseris, J.; Obar, R. A.; Harris, T.; Artavanis-Tsakonas, S.; Sowa, M. E.; De Camilli, P.; Paulo, J. A.; Harper, J. W.; Gygi, S. P. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **2015**, *162* (2), 425−40.
(33) Muller, C. S.; Bildl, W.; Haupt, A.; Ellenrieder, L.; Becker, T.; Hunte, C.; Fakler, B.; Schulte, U. Cryo-slicing BN-MS - a novel technology for high-resolution complexome profiling. *Mol. Cell. Proteomics* **2016**, *15* (2), 669−81.
(34) Giese, H.; Ackermann, J.; Heide, H.; Bleier, L.; Drose, S.; Wittig, I.; Brandt, U.; Koch, I. NOVA: a software to analyze complexome profiling data. *Bioinformatics* **2015**, *31* (3), 440−1.
(35) Wu, M.; Yu, Q.; Li, X.; Zheng, J.; Huang, J. F.; Kwoh, C. K. Benchmarking human protein complexes to investigate drug-related systems and evaluate predicted protein complexes. *PLoS One* **2013**, *8* (2), e53197.
(36) Goh, W. W.; Guo, T.; Aebersold, R.; Wong, L. Quantitative proteomics signature profiling based on network contextualization. *Biol. Direct* **2015**, *10* (1), 71.
(37) Ideker, T.; Krogan, N. J. Differential network biology. *Mol. Syst. Biol.* **2012**, *8*, 565.
(38) Srihari, S.; Ragan, M. A. Systematic tracking of dysregulated modules identifies novel genes in cancer. *Bioinformatics* **2013**, *29* (12), 1553−61.
(39) Belorkar, A.; Wong, L. GFS: Fuzzy preprocessing for effective gene expression analysis. *BMC Bioinf.* **2016**, *17*, 540.
(40) Lim, K.; Wong, L. Finding consistent disease subnetworks using PFSNet. *Bioinformatics* **2014**, *30* (2), 189−96.
(41) Goh, W. W. Fuzzy-FishNET: A highly reproducible protein complex-based approach for feature selection in comparative proteomics. *BMC Med. Genomics* **2016**, *9* (Suppl 3), 67.
(42) Lim, K.; Li, Z.; Choi, K. P.; Wong, L. A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small. *J. Bioinf. Comput. Biol.* **2015**, *13* (4), 1550018.
(43) Khatri, P.; Draghici, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **2005**, *21* (18), 3587−95.
(44) Goh, W. W.; Lee, Y. H.; Ramdzan, Z. M.; Sergot, M. J.; Chung, M.; Wong, L. Proteomics signature profiling (PSP): a novel contextualization approach for cancer proteomics. *J. Proteome Res.* **2012**, *11* (3), 1571−81.
(45) Soh, D.; Dong, D.; Guo, Y.; Wong, L. Finding consistent disease subnetworks across microarray datasets. *BMC Bioinf.* **2011**, *12* (Suppl13), S15.
(46) Webb-Robertson, B.-J. M.; Wiberg, H. K.; Matzke, M. M.; Brown, J. N.; Wang, J.; McDermott, J. E.; Smith, R. D.; Rodland, K. D.; Metz, T. O.; Pounds, J. G.; Waters, K. M. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J. Proteome Res.* **2015**, *14* (5), 1993−2001.
(47) Langley, S. R.; Mayr, M. Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. *J. Proteomics* **2015**, *129*, 83−92.
(48) Guo, T.; Kouvonen, P.; Koh, C. C.; Gillet, L. C.; Wolski, W. E.; Rost, H. L.; Rosenberger, G.; Collins, B. C.; Blum, L. C.; Gillessen, S.; Joerger, M.; Jochum, W.; Aebersold, R. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat. Med.* **2015**, *21* (4), 407−13.
(49) Suzuki, R.; Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **2006**, *22* (12), 1540−2.
(50) Dennis, G., Jr.; Sherman, B. T.; Hosack, D. A.; Yang, J.; Gao, W.; Lane, H. C.; Lempicki, R. A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome biology* **2003**, *4* (5), P3.
(51) Boyle, E. I.; Weng, S.; Gollub, J.; Jin, H.; Botstein, D.; Cherry, J. M.; Sherlock, G. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **2004**, *20* (18), 3710−5.
(52) Guruharsha, K. G.; Rual, J. F.; Zhai, B.; Mintseris, J.; Vaidya, P.; Vaidya, N.; Beekman, C.; Wong, C.; Rhee, D. Y.; Cenaj, O.; McKillip, E.; Shah, S.; Stapleton, M.; Wan, K. H.; Yu, C.; Parsa, B.; Carlson, J. W.; Chen, X.; Kapadia, B.; VijayRaghavan, K.; Gygi, S. P.; Celniker, S. E.; Obar, R. A.; Artavanis-Tsakonas, S. A protein complex network of Drosophila melanogaster. *Cell* **2011**, *147* (3), 690−703.
(53) Havugimana, P. C.; Hart, G. T.; Nepusz, T.; Yang, H.; Turinsky, A. L.; Li, Z.; Wang, P. I.; Boutz, D. R.; Fong, V.; Phanse, S.; Babu, M.; Craig, S. A.; Hu, P.; Wan, C.; Vlasblom, J.; Dar, V. U.; Bezginov, A.; Clark, G. W.; Wu, G. C.; Wodak, S. J.; Tillier, E. R.; Paccanaro, A.; Marcotte, E. M.; Emili, A. A census of human soluble protein complexes. *Cell* **2012**, *150* (5), 1068−81.
(54) Gandhi, T. K.; Zhong, J.; Mathivanan, S.; Karthick, L.; Chandrika, K. N.; Mohan, S. S.; Sharma, S.; Pinkert, S.; Nagaraju,

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112

S.; Periaswamy, B.; Mishra, G.; Nandakumar, K.; Shen, B.; Deshpande, N.; Nayak, R.; Sarker, M.; Boeke, J. D.; Parmigiani, G.; Schultz, J.; Bader, J. S.; Pandey, A. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* **2006**, *38* (3), 285−93.

(55) Mathivanan, S.; Periaswamy, B.; Gandhi, T. K.; Kandasamy, K.; Suresh, S.; Mohmood, R.; Ramachandra, Y. L.; Pandey, A. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinf.* **2006**, *7* (Suppl 5), S19.

(56) Prasad, T. S.; Kandasamy, K.; Pandey, A. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.* **2009**, *577*, 67−79.

(57) Stark, C.; Breitkreutz, B. J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **2006**, *34* (Database issue), 90001.

DOI: 10.1021/acs.jproteome.7b00363
J. Proteome Res. 2017, 16, 3102−3112