Journal of Bioinformatics and Computational Biology
Vol. 14, No. 5 (2016) 1650029 (23 pages)

DOI: 10.1142/S0219720016500293

World Scientific
www.worldscientific.com

Evaluating feature-selection stability in next-generation proteomics

Wilson Wen Bin Goh* and Limsoon Wong†

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

Department of Computer Science
National University of Singapore
13 Computing Drive, Singapore 117417
**wilson.goh@tju.edu.cn; goh.informatics@gmail.com*
†*wongls@comp.nus.edu.sg*

Received 29 July 2016
Accepted 29 July 2016
Published 19 September 2016

Identifying reproducible yet relevant features is a major challenge in biological research. This is well documented in genomics data. Using a proposed set of three reliability benchmarks, we find that this issue exists also in proteomics for commonly used feature-selection methods, e.g. *t*-test and recursive feature elimination. Moreover, due to high test variability, selecting the top proteins based on *p*-value ranks — even when restricted to high-abundance proteins — does not improve reproducibility. Statistical testing based on networks are believed to be more robust, but this does not always hold true: The commonly used hypergeometric enrichment that tests for enrichment of protein subnets performs abysmally due to its dependence on unstable protein pre-selection steps. We demonstrate here for the first time the utility of a novel suite of network-based algorithms called ranked-based network algorithms (RBNAs) on proteomics. These have originally been introduced and tested extensively on genomics data. We show here that they are highly stable, reproducible and select relevant features when applied to proteomics data. It is also evident from these results that use of statistical feature testing on protein expression data should be executed with due caution. Careless use of networks does not resolve poor-performance issues, and can even mislead. We recommend augmenting statistical feature-selection methods with concurrent analysis on stability and reproducibility to improve the quality of the selected features prior to experimental validation.

Keywords: Proteomics; networks; biostatistics; translational research.

1. Introduction

Next-generation mass spectrometry (MS)-based proteomics is indispensable to current biological and clinical research.[1] It is the primary means of observing

*Corresponding author.

protein changes relating to the phenotype and, thus, provides direct information on druggable targets or biomarkers. Recent proteomic advancements have led to brute-force acquisition approaches such as the new data-independent acquisition (DIA) and high-resolution data-dependent acquisition (DDA) paradigms, which yield more complete datasets.[2,3] However, noise and proteome-coverage issues remain prevalent.

A common application in biostatistical analysis is to compare and identify which variables are strongly discriminative between two groups. This process is known as feature selection. In proteomics, the measured variables (or features) are the proteins and their measured expression levels. Discriminative proteins may be causal or merely correlated to the differences between the groups in a particular dataset. Discriminative proteins that are causal are useful for building predictive models that may be used for diagnostics and prognostics. However, irrelevant proteins that are merely correlated would exhibit poor generalizability, and are of no practical value in real settings. Unfortunately, given current feature-selection approaches, they are not easy to tell apart from each other. Also, ranking by p-values and taking the top $n\%$ features does not guarantee signatures of better quality.[4] Amongst the most popular feature-selection methods at the level of individual proteins (e.g. t-test, SAM,[5] and DESEQ,[6] no approach strongly outperforms the others with respect to effect size, proportion size (i.e., how large one sample class is over the other) and influence on the false-discovery rate (FDR).[7]

Networks can be combined with proteomics synergistically.[8–13] to overcome its idiosyncratic coverage and consistency issues.[14,15] They can also be adapted for feature selection (where features are subnets or complexes instead of individual proteins) with high stability; i.e. similar features are selected given any random subsampling of the original data.[16] Recent advances in network-based approaches have led to the development of a new class of methods, rank-based network algorithms (RBNAs), which have been extensively tested on genomics data and shown to be extremely powerful for identification of relevant subnets, producing unparalleled prediction reliability and reproducibility.[17,18] RBNAs include SNET (SubNETworks).[18] and its successors, FSNET (Fuzzy SNET) and PFSNET (Paired FSNET).[17] As they have never been studied in the context of proteomics (as high-quality proteomics datasets of sufficient sample sizes have only become recently available), it is useful (for users) to know how well they fare, and how much better relative to conventional approaches.

This case-study paper also illustrates the importance of incorporating other means of evaluating feature-selection stability and relevance as alternative indicators in addition to the statistical p-value. For real data, we introduce a set of three benchmarking approaches here. While networks are potentially powerful approaches, we do not think they are all born equal as some underlying assumptions or algorithm design may be flawed. Hence, in addition to protein features, we also consider a multitude of network approaches that select subnets/protein complexes as features.

2. Material and Methods

2.1. *Simulated proteomics data — D1.2 and D2.2*

Two simulated datasets from the study of Langley and Mayr[7] is used. D.1.2 is from LC-MS/MS study of proteomic changes resulting from addition of exogenous matrix metallopeptidase (3 control, 3 test). D2.2 is from a study of hibernating arctic squirrels (4 control, 4 test). Quantitation in both studies is based on spectral counts.

For both D1.2 and D2.2, 100 simulated datasets each with 20% randomly generated significant features are used. For D1.2 and D2.2, this works out as 177 and 710 significant proteins, respectively. Effect size of these 20% features is not even as they are randomly generated from one out of five possibilities or p (20%, 50%, 80%, 100%, and 200%) and is expressed as:

$$\overline{\mathrm{SC}}_{i,j} = \overline{\mathrm{SC}}_{i,j} * (1 + p), \tag{1}$$

where $\mathrm{SC}_{i,j}$ is the simulated spectral count from the jth sample of protein i.

2.2. *Proteomics dataset 1 based on DDA — colorectal cancer*

The colorectal cancer (CR) study contains 90 CR samples derived through the TCGA Biospecimen Core Resource (BCR).[19] 30 normal (non-matched) samples are obtained from screening colonoscopies.

To ensure data quality: For every five CR samples, benchmark quality controls (QCs) from one basal and one luminal human breast tumor xenograft are analyzed. Both standard search (Myrimatch v2.1.87) and spectral search (Pepitome) were used. Peptide identification stringency is set at FDR of 2% for higher sensitivity. For protein assembly, a minimum of two unique peptides per protein is essential for a positive identification (3899 proteins with a protein-level FDR of 0.43%). To limit data holes, only proteins supported by 95% of samples are kept (3609 proteins).

Proteins are quantified via spectral count, which is the total number of MS/MS spectra acquired for peptides from a given protein.

2.3. *Proteomics dataset 2 based on DIA — renal cancer*

For the second case study, a SWATH-derived renal cancer (RC) dataset is used.[20] This contains 12 normal and cancer samples which have been captured in duplicates. Here, we opt to analyze the technical duplicates together: If technical noise is present, it will impede the ability of less-robust feature-selection methods to consistently select the same features anyway.

All spectral maps are analyzed using OpenSWATH[21] against a spectral library containing 49,959 reference spectra for 41,542 proteotypic peptides from 4624 reviewed SwissProt proteins.[20] The library is compiled using DDA data of the kidney tissues in the same mass spectrometer. Protein isoforms and protein groups are excluded.

In the original dataset, 2375 proteins are quantified across all samples with a precursor FDR below 0.1% and 32% sparsity. The high level of sparseness is likely due to higher noise and poorer alignments between features. To improve consistency in the data, we relax the retention time (RT) alignment criteria using transition of identification confidence (TRIC) (version r238) where we allow a wider maximal RT difference (max_RT = 30) but set the precursor FDR to 0.001% to limit false matches. The two most intense peptides are used to quantify proteins. Finally, protein FDR is set to 1% with 3123 reported proteins.

2.4. *Reference complexome*

For network-based feature-selection methods, the features are subnets/complexes. Currently, the gold-standard protein-complex dataset is the CORUM database, which contains manually annotated protein complexes from mammalian organisms.[22] In earlier studies, real complexes are demonstrated to be superior to predicted ones from protein-interaction networks[23]; so we use real complexes.

2.5. *Two-sample t-test of single proteins*

As a control, the two-sample t-test for selection of single proteins (SP) is performed. Briefly, a t-statistic (T_p) and its corresponding nominal p-value are calculated for each protein p by comparing the expression scores between classes C_1 and C_2, with the assumption of unequal variance between the two classes[24]:

$$T_p = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{2}$$

where $\bar{x}_j$ is the mean expression level of the protein p, s_j is the standard deviation, and n_j is the sample size, in class C_j.

2.6. *Hypergeometric enrichment*

The hypergeometric enrichment (HE) is a frequently used form of subnet evaluation and consists of two steps[9]: First, differential proteins are identified using the two-sample t-test. This is followed by a hypergeometric test where given a total of N proteins (with B of these belonging to a complex) and n test-set proteins (i.e. differential proteins), the exact probability that b or more proteins from the test set are associated by chance with the complex is given by:

$$p(X \geq b) = \sum_{i=b}^{\min(n,B)} \frac{\binom{n}{i}\binom{N-n}{B-i}}{\binom{N}{B}}. \tag{3}$$

The sum $P(X \geq b)$ is the p-value of the hypergeometric test.

2.7. *SNET/FSNET/PFSNET*

The RBNAs (SNET, FSNET, and PFSNET) are in principle, improvements on top of each other. Implementations of these RBNAs are available at http://www.comp.nus.edu.sg/∼ wongls/projects/drug-pathway/pfsnet-v2.zip.

We begin with a description of SNET: Given a protein g_i and a tissue p_k, let $fs(g_i, p_k) = 1$, if the protein g_i is among the top alpha percent (default = 10%) most-abundant proteins in the tissue p_k; and = 0 otherwise. Given a protein g_i and a class of tissues C_j, let

$$\beta(g_i, C_j) = \sum_{pk \epsilon c_j} \frac{fs(g_i, p_k)}{|C_j|}. \tag{4}$$

That is, $\beta(g_i, C_j)$ is the proportion of tissues in C_j that have g_i among their top alpha percent most-abundant proteins.

Let score(S, p_k, C_j) be the score of a protein complex S and a tissue p_k weighted based on the class C_j. It is defined as:

$$\text{score}(S, p_k, C_j) = \sum_{g_i \in S} fs(g_i, p_k) * \beta(g_i, C_j). \tag{5}$$

The function $f_{SNET}(S, X, Y, C_j)$ for some complex S is a t-statistic defined as:

$$f_{\text{SNET}}(S, X, Y, C_j) = \frac{\text{mean}(S, X, C_j) - \text{mean}(S, Y, C_j)}{\sqrt{\frac{\text{var}(S,X,C_j)}{|X|} + \frac{\text{var}(S,Y,C_j)}{|Y|}}}, \tag{6}$$

where mean(S,#,C_j) and var(S,#,C_j) are respectively the mean and variance of the list of scores {score(S,p_k,C_j)|p_k is a tissue in #}.

The complex S is considered differential (weighted based on C_j) in X but not in Y if $f_{\text{SNET}}(S,X,Y,C_j)$ is at the largest 5% extreme of the Student t-distribution, with degrees of freedom determined by the Welch–Satterwaite equation.

Given two classes C_1 and C_2, the set of significant protein complexes returned by SNET is the union of $\{S|f_{\text{SNET}}(S,C_1,C_2,C_1)$ is significant$\}$ and $\{S|f_{\text{SNET}}(S,C_2,C_1,C_2)$ is significant$\}$; the former being complexes that are significantly consistently highly abundant in C_1 but not C_2, the latter being complexes that are significantly consistently highly abundant in C_2 but not C_1.

FSNET is identical to SNET, except in one regard:

For FSNET, the definition of the function $fs(g_i, p_k)$ is replaced such that $fs(g_i,p_k)$ is assigned a value between 1 and 0 as follows: $fs(g_i, p_k)$ is assigned the value 1 if g_i is among the top alpha1 percent (default = 10%) of the most-abundant proteins in p_k. It is assigned the value 0 if g_i is not among the top alpha2 percent (default = 20%) most-abundant proteins in p_k. The range between alpha1 percent and alpha2 percent is divided into n equal-sized bins (default $n = 4$), and $fs(g_i, p_k)$ is assigned the value 0.8, 0.6, 0.4, or 0.2 depending on which bin g_i falls into p_k.

A test statistic f_{FSNET} is then defined analogously to f_{SNET}. Given two classes C_1 and C_2, the set of significant complexes returned by FSNET is the union of $\{S|f_{\mathrm{FSNET}}(S, C_1, C_2, C_1)$ is significant$\}$ and $\{S|f_{\mathrm{FSNET}}(S, C_2, C_1, C_2)$ is significant$\}$.

For PFSNet, the same $fs(g_i, p_k)$ function as in FSNet is used. But it defines a score $\mathrm{delta}(S, p_k, X, Y)$ for a complex S and tissue p_k with respect to classes X and Y as the difference of the score of S and tissue p_k weighted based on X from the score of S and tissue p_k weighted based on Y. More precisely: $\mathrm{delta}(S, p_k, X, Y) = \mathrm{score}(S, p_k, X) - \mathrm{score}(S, p_k, Y)$.

If a complex S is irrelevant to the difference between classes X and Y, the value of $\mathrm{delta}(S, p_k, X, Y)$ is expected to be around 0. So PFSNet defines the following one-sample t-statistic:

$$f_{\mathrm{PFSENT}}(S, X, Y, Z) = \frac{\mathrm{mean}(S, X, Y, Z)}{\mathrm{se}(S, X, Y, Z)}, \tag{7}$$

where $\mathrm{mean}(S, X, Y, Z)$ and $\mathrm{se}(S, X, Y, Z)$ are respectively the mean and standard error of the list $\{\mathrm{delta}(S, p_k, X, Y)|p_k$ is a tissue in $Z\}$. The complex S is considered significantly consistently highly abundant in X but not in Y if $f_{\mathrm{PFSNet}}(S, X, Y, X \cup Y)$ is at the largest 5% extreme of the Student t-distribution.

Given two classes C_1 and C_2, the set of significant complexes returned by PFSNet is the union of $\{S|f_{\mathrm{PFSNet}}(S, C_1, C_2, C_1 \cup C_2)$ is significant$\}$ and $\{S|f_{\mathrm{PFSNet}}(S, C_2, C_1, C_1 \cup C_2)$ is significant$\}$; the former being complexes that are significantly consistently highly abundant in C_1 but not C_2, and vice versa.

2.8. *Recursive feature elimination*

Recursive feature elimination (RFE) is another popular feature-selection method and it follows a different paradigm (viz machine learning) from t-test.[25] In RFE, a machine-learning approach (e.g. support vector machines (SVMs), Naïve Bayes, or random forest (RF)) is used to build a classifier first. This is followed by iteratively eliminating features having the least impact on accuracy. This allows the features to be ranked and selected based on their overall impact on model accuracy.

We used the R module "caret" to perform RFE.[26] The machine-learning approach we used here is the popular ensemble method, RF.[27] Model accuracy is evaluated based on 10-fold cross-validation. Two feature-selection thresholds are used: The minimum number of features required to reach 100% model accuracy (designated RFE in our tests) and the top 500 features to match approximately the number of proteins selected in PFSNET (designated RFE500).

2.9. *Performance benchmarks (simulations)*

There is no universally accepted standard for generating randomized data/complexes for evaluating network-based feature-selection methods. Using D1.2 and 2.2, we propose two related approaches: In the first, we randomly partition the differential proteins into non-overlapping pseudo-complexes of sizes 6 to 7 (20 complexes)

and 7 to 10 for D2.2 (101 complexes). An equal number of non-significant proteins are randomly selected and partitioned into same number of pseudo-complexes. The significant and non-significant pseudo-complexes are combined into a single complex vector and evaluated.

To create a more realistic simulation, we can incorporate expression information as a constraint. In this second approach, for both differential and non-differential proteins, a Euclidean distance is calculated for all protein pairs across the samples. These are then clustered via Ward's linkage. The proteins are then reordered such that those with similar expression pattern are adjacent to each other. This reordered list is then split at regular intervals to generate 20 and 101 complexes for D1.2 and D2.2, respectively. The remaining steps are similar to the first method.

Since the significant and non-significant complexes are known *a priori*, we can determine the distribution of precision and recall as well as the F-score across all simulation datasets (see next section).

2.10. *Performance benchmarks (real data)*

To evaluate stability and relevance on real data (where the real differential features are not known prior), we propose benchmarks based on three criteria: Precision/recall, feature-selection stability, and normalized cross-validation prediction accuracy. These benchmarks are designed to evaluate reproducibility in the form of sensitivity of a method to variations in the training set.[28]

Precision/recall — The set of significant complexes c, from each subsampling is benchmarked against the total set of significant complexes, C, derived from an analysis of the complete dataset. This makes the assumption that the complete dataset is representative of the population. Thus, a completely precise method based on a subsampling should report a subset c of C ($c \subseteq C$) as significant, and no more (considered false positives). Similarly, perfect recall should report all complexes in C as significant.

Precision and recall are defined as:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

where TP, FP, and FN are the True Positives, False Positives, and False Negatives, respectively.

Precision and Recall can be combined using the harmonic mean. This is also known as the F-score (F_S):

$$F_s = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{9}$$

Repeated subpopulation sampling provides insight into the variability and approximation of precision/recall. However, a caveat is that it may mislead when a feature-selection algorithm is unstable, as it is likely to return an entirely different "gold standard" set of features when applied on a different dataset.

Feature-selection stability — Random subsamplings is performed 1000 times to generate a binary matrix, where rows and columns represent samplings and complexes, respectively. A value of 1 denotes statistical significance and 0 otherwise.

To evaluate the stability of each complex, each column can be summed and normalized by the number of subsamplings such that a value close to 1 indicates very high stability. The distribution of the protein-complex stabilities provides an overall indication of the stability of the feature-selection approach. For simplicity, the mean of the complex stabilities denotes the feature-selection stability.

Normalized cross-validation accuracy — If the selected features are relevant, then they should be able to correctly predict sample classes (e.g. normal versus disease) from new data. In the absence of suitable independent data, cross-validation is an, albeit less satisfying, alternative. In cross-validation, the data is split into training and validation sets. Features selected from the training set are used to train a machine-learning classifier (in this case, Naïve Bayes). The classifier is then presented with the validation set (of known class assignments), and a cross-validation accuracy (CVAccuracy) determined,

$$\text{CVAccuracy} = \frac{\text{Number of correct class assignments}}{\text{Total size of validation set}}. \tag{10}$$

The CVAccuracy in itself is not particularly meaningful as random features may also have predictive accuracy comparable to the selected features themselves.[29] Therefore, to determine whether CVAccuracy is meaningful, an equal number of random features is used for retraining the Naïve Bayes classifier. This is repeated 1000 times to generate a null distribution. Akin to Monte Carlo resampling statistics, the CVAccuracy p-value is the number of times null accuracy $\geq$ CVAccuracy divided by 1000.[30]

The normalized CVAccuracy considers both the CVAccuracy and its corresponding p-value simultaneously. It is expressed as the ratio, CVAccuracy/p-value, and is $\gg 1$ if meaningful.

3. Results and Discussions

3.1. *Network-based methods, especially RBNAs excelled in the simulation studies*

The original RBNA papers base their evaluations on subnet and gene agreements between related real data where the true positive features are not known *a priori*.[17,18] While subnet/gene agreements provide an indication of feature reproducibility, it is not exactly a clear measure of feature-selection performance. As the preponderance of false positives and negatives are not known, evaluation based on reproducibility is a reasonable compromise. This unfortunately stems from a lack of gold-standard simulation approaches for evaluating network-based approaches.

Here, we try two approaches that may be useful — the first generates pseudo-complexes via random incorporation while the second utilizes expressional

correlations (given that complex constituents are more likely to be co-expressed). It turns out that regardless of simulation approach, the RBNAs — in particular, PFSNET — excelled, especially at recall while maintaining reasonable precision (Figs. 1 and 2). Moreover, the correlation-based simulation approach does not immediately appear to be superior. These results are consistent both at the protein-complex as well as the individual-protein level (Figs. 1 and 2; top and bottom rows).

As the RBNAs have never been evaluated based on simulation studies, these results further demonstrate the power of this class of approaches. The high recall performance is especially noteworthy, given that only the top 20–30% of proteins across samples are actually being considered. In contrast, the standard t-test (SP) is severely lacking in both precision and recall. For the hypergeometric test (HE), while precision tends to be very high, its recall tends to be very poor. We expect that this simulation tends to favor HE: All the incorporated proteins in the significant complexes are differential; so we expect better intersection rates and hence, better p-values. But this may be counteracted by the poor precision and recall of the t-test step (which is the same as SP) prior to over-representation testing. In this regard, HE appears to be a good means of filtering false positives reported by the t-test: When the complex is reported to be significant in HE, it is likely correct (high precision). This is within expectation, and quite frequently why the hypergeometric test is used as a post-hoc filter following the t-test. However, its very low recall leaves room for concern.

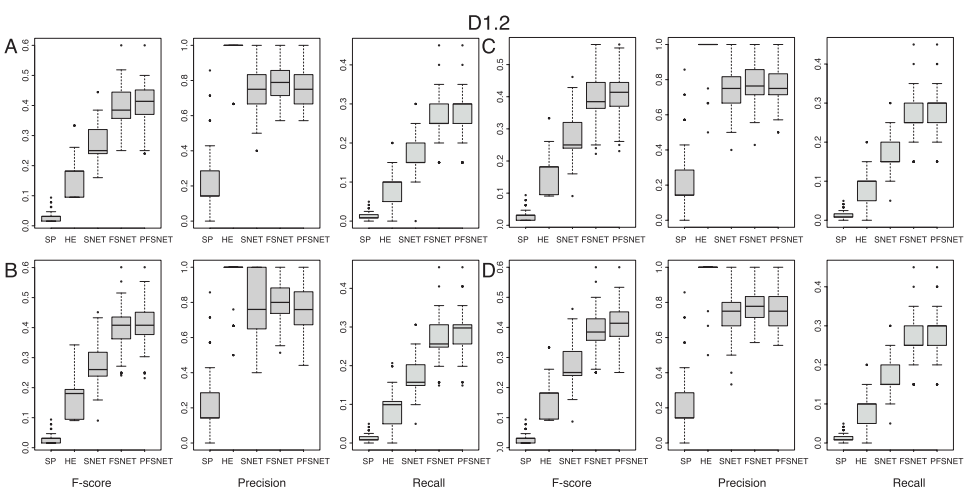


Fig. 1. F-score, precision, and recall distributions for simulated data D1.2. A/B and C/D show the F-score, precision, and recall distributions across 100 simulated datasets based on the random pseudo-complex generation method and correlations-based pseudo-complex generation method, respectively. The top row (A/C) shows the distributions for HE and the RBNAs based on complexes while the bottom row (B/D) show the same distributions in terms of the proteins embedded within the complexes. Note that SP does not change in either row, as it does not utilize pseudo-complexes. The RBNAs, particularly PFSNET, dominate, excelling in recall.

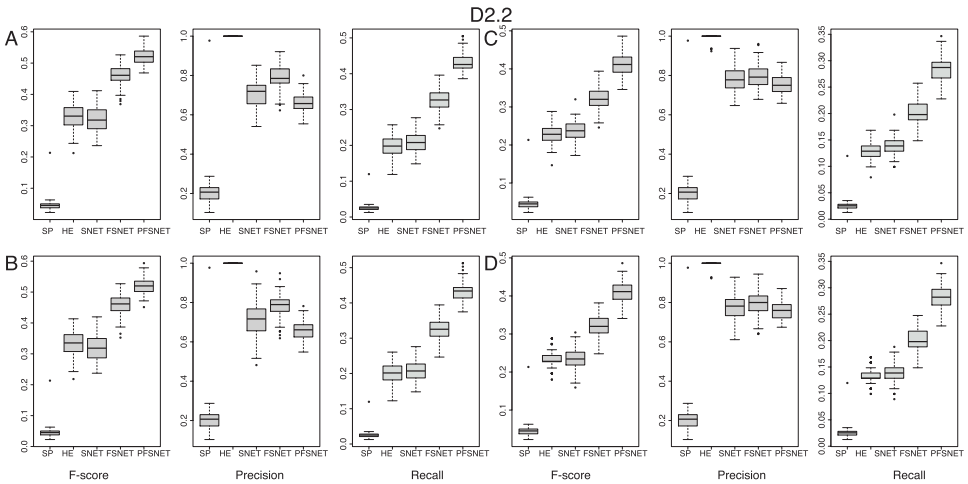


Fig. 2. F-score, precision, and recall distributions for simulated data D2.2. A/B and C/D show the F-score, precision, and recall distributions across 100 simulated datasets based on the random pseudo-complex generation method and correlations-based pseudo-complex generation method, respectively. The top row (A/C) shows the distributions for HE and the RBNAs based on complexes while the bottom row (B/D) show the same distributions in terms of the proteins embedded within the complexes. Note that SP does not change in either row, as it does not utilize pseudo-complexes. As with Fig. 1, the RBNAs, particularly PFSNET, dominate, excelling in recall.

The simulation studies provide a surprising glimpse into the ability of RBNAs to recover significant features (subnets and complexes), as it only considers a small subset of the differential proteins (based on the alpha filters). It also reveals some rather interesting aspects regarding the individual RBNAs. For SNET, FSNET, and PFSNET, their precision is largely comparable, although PFSNET lags relatively in this respect. However, PFSNET is far more sensitive, allowing it to achieve considerably higher F-score than the other two.

Incorporation of correlations for pseudo-complex generation has subtle effects. It appears to have consistent homogenizing effects on the RBNAs' accuracy, making them closer to each other as well as elevating it slightly. It appears to have minimal beneficial effects on HE (SP does not use the pseudo-complexes). We acknowledge that this may not be the perfect way of generating pseudo-complexes, as these pseudo-complexes may not be biologically coherent. However, if the differential proteins are simply grouped together, then the RBNAs do a good job of identifying these groups anyway. The proposed three benchmarks (e.g. feature-selection stability and CVAccuracy) for real data may also be performed on the simulated data, but these require large numbers of resamplings, for which the simulated dataset is unsuitable due to small n. Moreover, despite these promising results, simulation studies do not capture true biological reality, hence it is essential, and more important, to test on real data.

3.2. *On real data, network-based approaches do not necessarily do better*

Tables 1 and 2 summarize the performance of the three benchmarks across the standard SP t-test, HE, and the RBNAs (SNET, FSNET, PFSNET) for the CR and RC datasets, respectively. As a different class of techniques, the results from RFE with the RF are discussed in Sec. 3.7. Three resampling sizes (4, 6, and 8) are used to

Table 1. Summary score tables for CR. A. The individual and average of the F-scores across resamplings of sizes 4 to 8. B. The individual and average of the feature-selection stability scores across resamplings of sizes 4 to 8. C. The normalized CVAccuracy is the cross-validation accuracy (CV Accuracy) divided over its accompanying p-value. Hence, it is meaningful when it is $\gg 1$. D. Overall rank sums across the three evaluation benchmarks. The smaller the rank sum, the better the performance. Here, PFSNET > FSNET > SNET > SP > HE.

A

	F-scores			
Method/Sampling_size	4	6	8	Avg
SP	0.73	0.81	0.81	0.79
HE	0.11	0.17	0.17	0.15
SNET	0.65	0.76	0.76	0.72
FSNET	0.73	0.80	0.80	0.78
PFSNET	0.84	0.89	0.89	0.88

B

	Feature-stability scores			
Method/Sampling_size	4	6	8	Avg
SP	0.38	0.47	0.53	0.46
HE	0.04	0.07	0.09	0.07
SNET	0.45	0.56	0.61	0.54
FSNET	0.57	0.65	0.69	0.64
PFSNET	0.74	0.83	0.86	0.81

C

Method	Number features	CVaccuracy	CV p-val	CVAccuracy/CV $p$val
SP	2662	1.00	1.00	1.00
HE	570	0.75	1.00	1.00
SNET	130	0.98	0.61	1.61
FSNET	141	1.00	0.67	1.49
PFSNET	200	1.00	0.72	1.39

D

Method	F-score	Feature-stability	CVaccuracy/CVp-val	Total Ranks
SP	2	4	4.5	10.5
HE	5	5	4.5	14.5
SNET	4	3	1	8
FSNET	3	2	2	7
PFSNET	1	1	3	5

Table 2. Summary score tables for RC. A. The individual and average of the F-scores across resamplings of sizes 4 to 8 B. The individual and average of the feature-selection stability scores across resamplings of sizes 4 to 8 C. The normalized CVAccuracy is the cross-validation accuracy (CV Accuracy) divided over its accompanying p-value. Hence, it is meaningful when it is $\gg 1$ D. Overall rank sums across the three evaluation benchmarks. The smaller the rank sum, the better the performance. Here, as with CR (c.f. Table 1), PFSNET > FSNET > SNET > SP > HE.

A

	F-scores			
Method/Samplin g_size	4	6	8	Avg
SP	0.61	0.76	0.76	0.71
HE	0.19	0.36	0.47	0.34
SNET	0.46	0.77	0.77	0.67
FSNET	0.59	0.77	0.77	0.71
PFSNET	0.87	0.94	0.94	0.92

B

	Feature-stability scores			
Method/Samplin g_size	4	6	8	Avg
SP	0.27	0.36	0.44	0.36
HE	0.08	0.13	0.17	0.13
SNET	0.28	0.60	0.72	0.53
FSNET	0.36	0.52	0.63	0.50
PFSNET	0.79	0.92	0.95	0.88

C

Method	Number features	CVaccuracy	CV p-val	CV accuracy/$p$val
SP	1124	0.98	0.91	1.08
HE	162	0.98	0.91	1.08
SNET	21	0.84	0.06	14.00
FSNET	36	0.96	0.06	16.00
PFSNET	65	0.92	0.06	15.33

D

Method	F-score	Feature stability	CVAccuracy/CV p-val	Total ranks
SP	3	5	5	13
HE	6	6	5	17
SNET	5	3	4	12
FSNET	3	4	1	8
PFSNET	2	2	3	7

simulate small to moderate data scenarios. In both CR and RC, SP has better feature-selection reproducibility and stability than HE. However, they perform similarly where normalized CVAccuracy is concerned. As with the observation of Venet *et al.*,[29] which is based on genomics, proteomic features selected by SP and HE also do no better than randomly selected features. On the same benchmarks, the RBNAs do extremely well and are highly ranked (Tables 1D and 2D). The network

features selected by RBNAs are also more relevant, with normalized CVAccuracy consistently greater than 1 (Tables 1C and 2C).

It is important to also consider the distribution of the scores; hence we plotted the global pairwise similarities of selected features (Figs. 3(a) and 4(a)) and the stability of individual selected features (Figs. 3(b) and 4(b)). While SP benefits from increased sampling size, it is noteworthy that on the contrary, HE does not, further

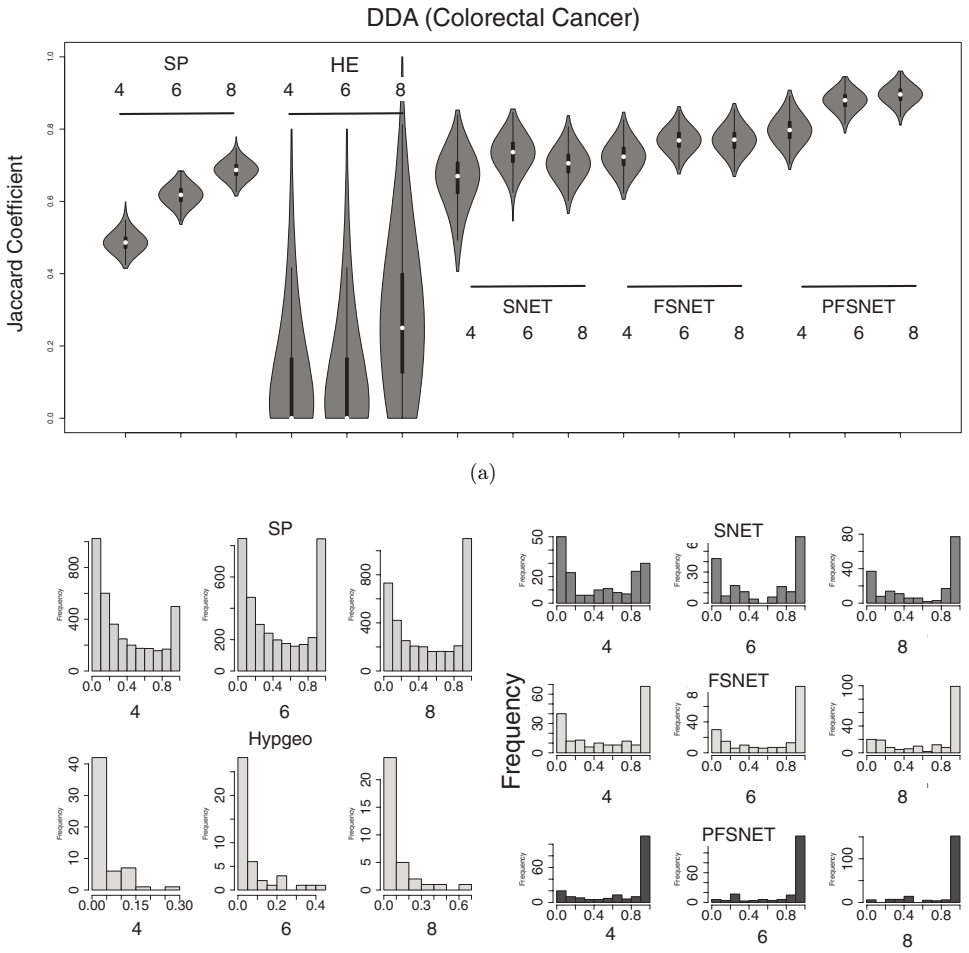


Fig. 3. Pairwise global similarity of selected features across random samplings and feature-selection stability for CR data. (a) Pairwise feature vector similarity across random samplings. All methods were evaluated 1000 times on random subsets of sizes 4, 6, and 8. Simulations were compared pairwise for reproducible features using the Jaccard Coefficient (SP — standard protein-based t-test, HE — Hypergeometric Enrichment). (b) Feature-selection stability across 1000 simulations. Histograms show distribution of feature-stability. Histograms with high density towards the right are more stable. Unlike SNET and FSNET, PFSNET is very stable at small sample sizes (x-axis: feature stability scores, y-axis: frequency), with the majority of features consistently reproducible across simulations.

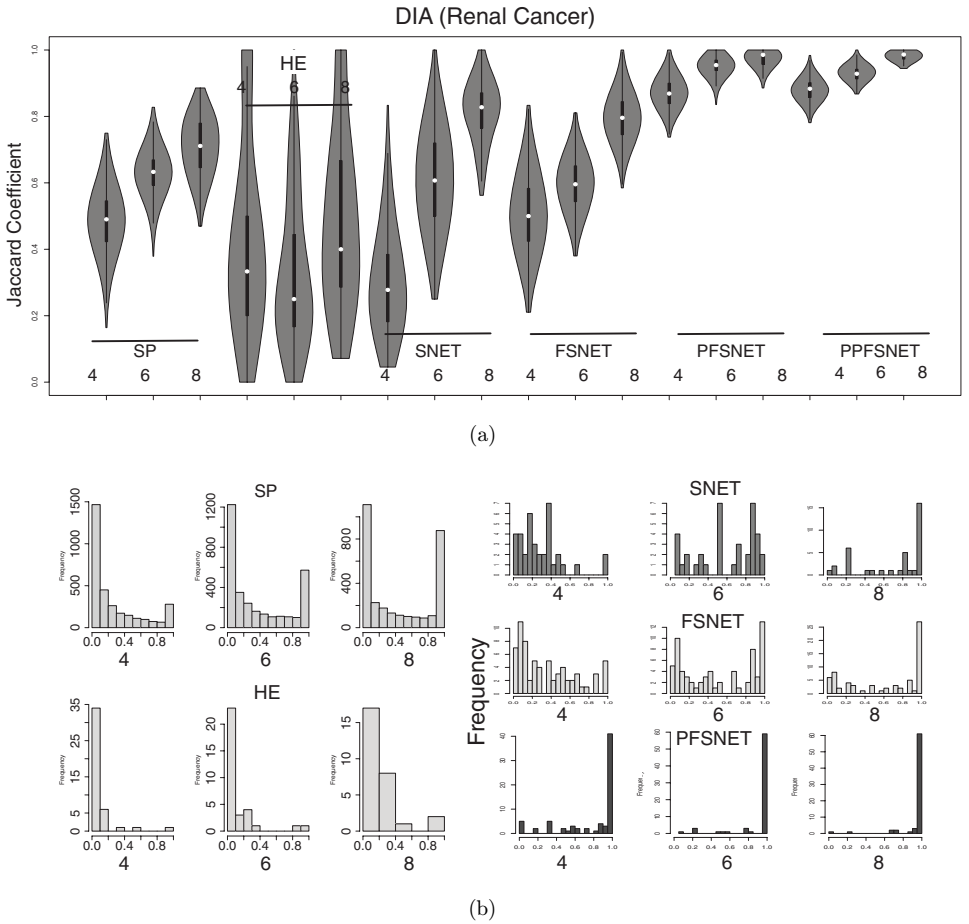


Fig. 4. Pairwise global similarity of selected features across random samplings and feature-selection stability for RC data. (a) Pairwise feature vector similarity across random samplings. All methods were evaluated 1000 times on random subsets of sizes 4, 6, and 8. Simulations were compared pairwise for reproducible features using the Jaccard Coefficient (SP — standard protein-based t-test, HE — Hypergeometric Enrichment). (b) Feature-selection stability across 1000 simulations. Histograms show distribution of feature stability. Histograms with high density towards the right are more stable. Unlike SNET and FSNET, PFSNET is very stable at small sample sizes (x-axis: feature stability scores, y-axis: frequency), with the majority of features consistently reproducible across simulations.

illustrating its innate instability. As with the improved performances observed in genomic data,[17,18] the more advanced RBNAs perform better, in particular PFSNET.

On real data, SP appears to have good precision/recall (Tables 1A and 2A). However, a closer inspection reveals this is an artifact due to an excessive number of features reported as significant. SP's feature stability is low relative to the F-scores. This confirms that the precision/recall of SP is inflated — individual features are highly unstable but because SP tends to report most features as significant (C), any

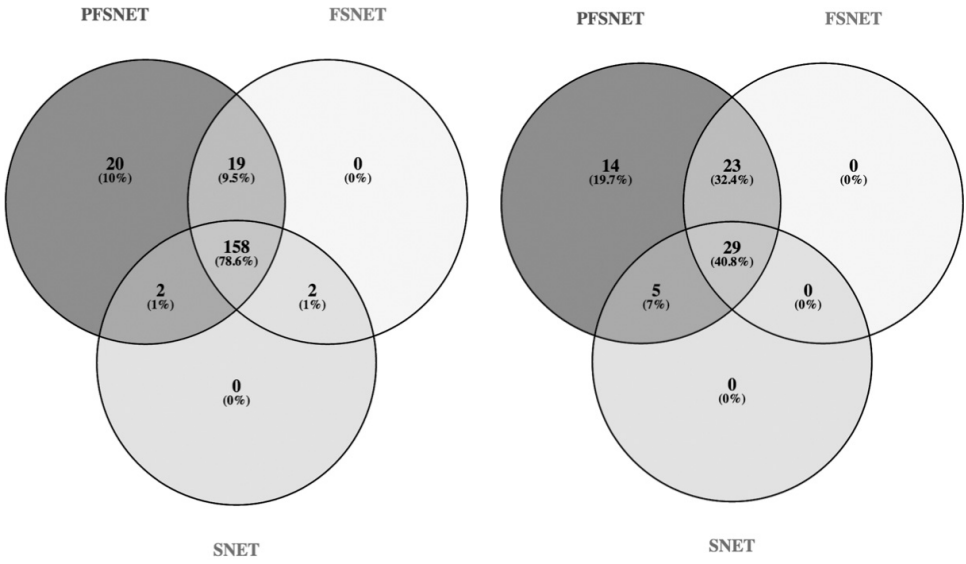


Fig. 5. Feature agreements between the RBNAs. (a) Significant feature overlaps for CR. (b) Significant feature overlaps for RC. All RBNAs have good overlaps with each other.

random resampling will generate a subset of C. Indeed, when the number of SP-significant features is restricted to the top n most (where n = number of unique proteins in complexes selected by PFSNET), the precision/recall and feature-stability score of SP drop precipitously (discussed below).

Despite its common usage, HE's poor showing is not completely unexpected since it uses the unstable t-test to pre-select differential protein features.[4] HE's instability implies that, in different subsamplings, the pre-selected proteins tend to enrich different complexes. Thus the subsets of pre-selected proteins that correspond to complexes are themselves unstable, and those pre-selected proteins that appear stable mostly do not enrich protein complexes. This is suggestive that many of these pre-selected proteins are biologically irrelevant. On the other hand, the corresponding values between F-scores and feature-stability scores are well aligned amongst the RBNAs (Tables 1A/B and 2A/B). PFSNET does particularly well in this regard. The RBNAs also largely agree with each other where feature selection is concerned (Fig. 5).

3.3. *In proteomics, standard feature-selection approaches do no better than random signatures*

In genomics, at least for breast cancer outcome, it is observed that random gene signatures do as well, if not better, in class prediction.[29] This suggests that many of

the selected features are merely correlated but not relevant to the phenotype. It is hypothesized that networks may act as useful constraints in improving the signal-to-noise ratio. Unfortunately, this cannot be naively assumed to be true.[31]

As with Venet *et al.*'s work on genomic signatures for breast cancer outcome,[29] we observe a similar phenomenon in proteomics (Tables 1C and 2C). By the CVAccuracy alone, it would appear that any method is good for class prediction. However, the inferred CVAccuracy p-value tells a different story. The traditional approaches, SP and HE, do no better than random protein signatures.

To examine this issue further, we find that the CVAccuracy null distribution generated from SP and HE is strongly biased (Fig. 6) — a small random selection of any 10 features already gives very high CVAccuracy. It seems that this bias may stem from the use of direct protein expression information as, in contrast, the RBNAs score each complex based on weighted ranks. Conversion of expression in-

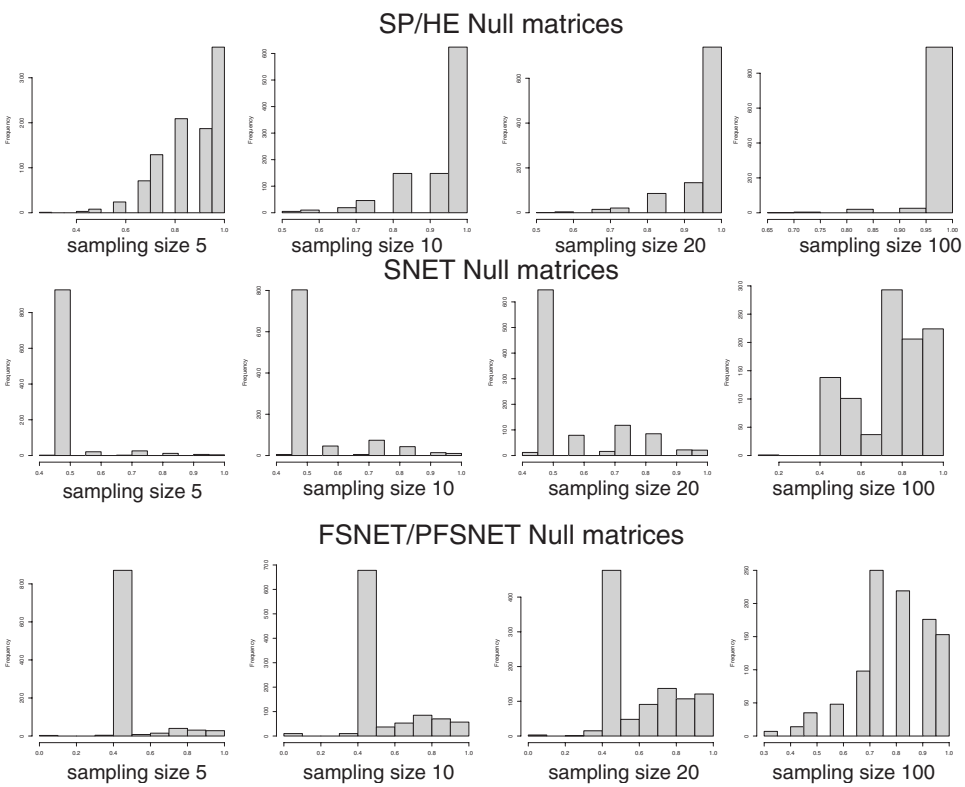


Fig. 6. CVAccuracy null models. Shown are the frequency distributions of cross-validation accuracies (x-axis) for different sampling sizes on different methods. For SP and HE, only a small number of randomly selected features are needed to generate highly accurate Naïve Bayes models. This set of histograms suggests offhand that the high CVAccuracy observed in SP (and therefore, HE) has no better predictive accuracy than a cardinal set of randomly picked protein. On the other hand, the null models for the RBNAs are more balanced. SNET and FSNET/PFSNET's null models are plotted separately due to different scoring system.

formation into weighted rank sums appears to be rewarding, with substantially more significant CVAccuracy p-values for the RBNAs. This finding deepens recent examination into ranks as more robust and unbiased than expression values[32]: Simple rank conversion is insufficient; network contextualization is still the key factor for improved performance (see Sec. 3.6).

3.4. *SP-based and hypergeometric-based enrichment (HE) are unsuitable for clinical applications*

The high variability of the t-test p-value is well described, particularly when sample size is small.[4,33] Thus, the poor performance of SP is expected. HE, on the other hand, is commonly used as a salve to improve confidence in SP-selected proteins, by taking advantage of the rich information in biological networks. HE's poor performance therefore is less expected.

Comparisons of the protein expression distributions revealed no major differences between the normal and cancer classes (Supplementary Figs. 1(a)/1(b)). Therefore, expressional disparity cannot explain why half the proteins are significant given the t-test.

We believe the benchmark statistics (e.g. F-score) generated for SP are inflated due to the high numbers of selected features but are not convinced that multiple testing corrections, e.g. Bonferroni, can solve this. We adjust the p-value threshold to restrict the number of allowed features in SP to the top 500 proteins (which would approximate the number of corresponding proteins in PFSNET). With this restriction, we observe a substantial drop in the proportion of stable features (Supplementary Fig. 1(c)), which implies high rank fluctuations among selected features and concomitant drop in pairwise feature similarity (Supplementary Fig. 1(d)).

The drastic drop in precision/recall (top 500 proteins) confirms that the perceived stability and consistency produced by SP is an artifact due to feature inflation (Supplementary Fig. 1(e)). This also highlights a problem associated with multiple-testing correction: Correction methods such as Benjamini–Hochberg and Bonferroni restrict the number of reported features to the top n but do not resolve rank instability. Moreover, restricting to the top n features post-multiple testing correction are even more likely to generate fluctuative results. It follows that the hypergeometric test, which is sensitive to the instability of SP, may be even more affected following multiple-testing correction.

There is a positive relationship between stability, t-test p-values and Bayes weights amongst protein features selected by both SP and HE (Supplementary Fig. 2. But there is a qualification — this only works when HE is performed on a sufficiently large dataset. The 302 HE-selected proteins (HE-SP intersect) are clearly enriched for stable SP-selected proteins (Supplementary Figs. 3(a)/3(b)). Unfortunately, this beneficial effect of HE is strongly dependent on sample size. We have already established that HE has high instability. Supplementary Fig. 4(a) confirms this as HE feature-selection stability remains low at the complex and corresponding

protein levels even as sampling size increases. There are 302 proteins associated with HE-significant complexes (when tested on the full dataset). Although most of these are enriched for stable SP proteins (Supplementary Fig. 3(a)/3(b)), reasonable enrichment for stable SP proteins can only be achieved at a large sample size (Supplementary Fig. 4(b)). Even so, HE cannot capture all stable SP proteins, nor is the SP subset it captures the most statistically significant (Supplementary Fig. 3(d)).

Ideal feature-selection methods should be highly accurate, stable, and identify relevant features. It appears that neither SP nor HE is able to meet these requirements. If analysts wish to use these tests on their data anyway, we advocate benchmarking for reproducibility and stability (and not to simply rank based on p-value) before attempting to interpret the results biologically.

3.5. *Stability issues associated with adjustment of alpha in RBNAs*

For RBNAs, we have maintained the default alpha parameters (top 10% + next 10%) based on prior genomics studies. Results from using default alphas indicate that the selected features are quite stable.[17,18] However, it is a valid concern that parameterization of alphas in RBNAs is not fixed: Increasing alpha from top 10% onwards can increase sensitivity, but comes at the cost of introducing more false positives as signals from lower ranked proteins are introduced into the complex scores.

As a simple demonstration, we repeated PFSNET using different alphas — top 10%, 20%, 30%, and 40%, and checked the rank stability (Spearman correlation) and overlaps (A∩B)/min(A,B) amongst differential complexes identified at $p < 0.05$ against the original parameters. Expectedly (Table 3), as alpha is increased (thus admitting more low-ranked proteins), more complexes are selected. But the variability of the PFSNET p-values (as measured by the coefficient of variation — standard deviation/mean) also rises correspondingly (Table 3). This means that the additional complexes have less significant p-values.

When alpha is low ($< 20\%$), protein-complex agreements are fairly stable (Table 3). However, when alpha rises above 20% more complexes are selected but the overall overlaps with the default settings remain fairly high ($> 75\%$). Although more complexes are selected when alpha is raised, the rank correlations for the selected

Table 3. Feature-selection overlaps and rank correlation of PFSNET significant features selected at various alphas (compared against the default parameters).

PFSNET (alphas)	Number of predicted complexes	Overlaps relative to top 10 (A∩B)/min (A,B)	Spearman rank correlation	p-values variability (s.d/mean)
10	42	1.00	1.00	2.47
20	53	1.00	0.84	2.81
30	76	0.79	0.86	3.08
40	83	0.75	0.93	4.05

complexes remain conserved and stable — i.e. overlapping complexes tend to maintain fairly similar ranks even at different alphas. However, we do not recommend setting alpha too high at first pass ($> 20\%$), as this will introduce many poorer-quality complexes early into analysis.

3.6. *Network constraints are the critical factor for deriving stability*

To confirm that network contextualization is the critical factor for obtaining stability, and not just restriction to the top 20% of proteins, for each sample, we replace the expression values by ranks. Then, we keep just the top $n\%$ proteins ($n = 20\%$ to match the RBNAs) per sample; i.e. different n proteins is allowed to be chosen per sample. The proteins outside the top n in each sample are set to the same rank (100,000). This is followed by a t-test and a Wilcoxon rank-sum test based on the rank values assigned as above. In this test scenario, the non-network-based feature-selection approaches such as the t-test and the Wilcoxon test both still fare worse than RBNAs (Supplementary Fig. 5). Unsurprisingly, fewer proteins are selected following rank restriction (Supplementary Fig. 5(c)) but these top $n\%$ proteins do not exhibit the same degree of stability as the RBNAs (Supplementary Fig. 5(d)). This demonstrates that restriction to the top $n\%$ most highly abundant proteins is insufficient for obtaining the stability observed in RBNAs. Thus, introducing network-based constraints (e.g. scoring against a protein complex) is a key factor for the performance improvement.

3.7. *Machine-learning approaches are powerful but the feature selection is unstable*

We have discussed some of the most commonly used deterministic feature-selection methods, e.g. the t-test and hypergeometric test. But machine-learning approaches such as the SVM and ensemble methods such as the RF and AdaBoost are important and can generate extremely robust models with good class-prediction accuracy.

Machine-learning methods incorporate feature ranking and model evaluation internally and iteratively. An example is the popular RFE method. Here, we opted to perform RFE with the RF and 10-fold cross-validation for model evaluation. Two feature-selection thresholds are used — the minimum number of features required for the model to reach 100% accuracy (default), and the top 500 (to approximate the number of corresponding proteins in PFSNET). We examine RFE's performance on CR and RC based on our stability benchmarks.

As powerful as machine-learning methods are, the default option leads to extreme instability issues (as demonstrated by the first two benchmarks on feature stability and precision/recall): A selected predictive feature here is as good as any other across the resamplings (Supplementary Figs. 6(a)/(c), and 7(a)/(c); Supplementary Tables 1 and 2). Here, predictive signatures range from one to five irreproducible proteins. Obviously, this also makes functional interpretation difficult. The result

also mirrors our observation that random selection of any five proteins can generate very accurate predictive models (Fig. 4).

Increasing the number of selected proteins to 500 improves stability (Supplementary Figs. 6(b)/(d), and 7(b)/(d); Supplementary Tables 1 and 2). However, while RFE is now superior to HE, it still trails far behind SP and the other RBNAs in terms of feature-selection reproducibility. As RFE does poorly on the first two benchmarks, and given known issues with use of SP as predictors (Fig. 6), it is not worthwhile to further evaluate its CVAccuracy performance.

3.8. *Making a case for RBNAs in proteomics*

On simulated data, RBNAs dominate, excelling particularly at recall. On real data, RBNAs display high feature-selection stability, precision/recall, and high normalized CVAccuracy. PFSNET is particularly powerful and can work even in the small-sample-size scenario, which is a big advantage given the high cost of generating large proteomics datasets. RBNAs easily outperform conventional SP expression and HE techniques. However, there is one poignant limitation: Because RBNAs work on the rank system, low-abundance proteins are occluded, even if they are informative. Although this is a critical limitation, low-abundance proteins generally have poor quantitation reliability. Thus, their rank shifts are less informative.[16] Earlier transcriptomic analysis results have proven this point, which resulted in the development of the alpha cut-offs.[17,18] Moreover, low-abundance proteins associated with significant yet stable complexes may be recoverable.[23]

4. Conclusions

Statistical feature selection in proteomics shares similar irreproducibility problems as in genomics. On simulated data, we demonstrate that RBNAs excelled, particularly at recall, even though we may not be able to fully capture biological coherence in our pseudo-complexes. On real data, we have introduced a suite of three reliability benchmarks that may be used for evaluating the quality of a given feature-selection method. We have shown that not all network-based approaches are born equal — the fundamental assumptions of the method (and how sound these are) are far more important. We have also demonstrated, for the first time, the beneficial utility of the RBNAs on proteomics data. Although rank conversion in RBNAs is important, network contextualization is the key factor for their superior performance.

Acknowledgments

This work was supported by an education grant from Tianjin University, China to WWBG and a Singapore Ministry of Education tier-2 grant, MOE2012-T2-1-061 to LW.

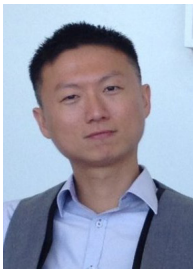
Competing Interests

The authors declare that they have no competing interests.

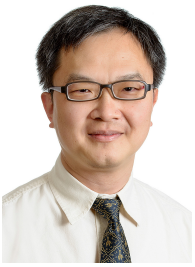
References

1. Altelaar AF, Munoz J, Heck AJ, Next-generation proteomics: Towards an integrative view of proteome dynamics, *Nat Rev Genet* **14**:35–48, 2013.
2. Carvalho PC, Han X, Xu T *et al.*, XDIA: Improving on the label-free data-independent analysis, *Bioinformatics* **26**:847–848, 2010.
3. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR, Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra, *Nat Methods* **1**:39–45, 2004.
4. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB, The fickle P value generates irreproducible results, *Nat Methods* **12**:179–185, 2015.
5. Larsson O, Wahlestedt C, Timmons JA, Considerations when using the significance analysis of microarrays (SAM) algorithm, *BMC Bioinformatics* **6**:129, 2005.
6. Love MI, Huber W, Anders S, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biol* **15**:550, 2014.
7. Langley SR, Mayr M, Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics, *J Proteomics* **129**:83–92, 2015.
8. Bensimon A, Heck AJ, Aebersold R, Mass spectrometry-based proteomics and network biology, *Annu Rev Biochem* **81**:379–405, 2012.
9. Goh WW, Lee YH, Chung M, Wong L, How advancement in biological network analysis methods empowers proteomics, *Proteomics* **12**:550–563, 2012.
10. Goh WW, Wong L, Networks in proteomics analysis of cancer, *Curr Opin Biotechnol* **24**:1122–1128, 2013.
11. Goh WW, Wong L, Computational proteomics: Designing a comprehensive analytical strategy, *Drug Discov Today* **19**:266–274, 2014.
12. Goh WW, Wong L, Sng JC, Contemporary network proteomics and its requirements, *Biology (Basel)* **3**:22–38, 2013.
13. Gstaiger M, Aebersold R, Applying mass spectrometry-based proteomics to genetics, genomics and network biology, *Nat Rev Genet* **10**:617–627, 2009.
14. Goh WW, Fan M, Low HS, Sergot M, Wong L, Enhancing the utility of Proteomics Signature Profiling (PSP) with Pathway Derived Subnets (PDSs), performance analysis and specialised ontologies, *BMC Genomics* **14**:35, 2013.
15. Goh WW, Lee YH, Ramdzan ZM, Sergot MJ, Chung M, Wong L, Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics, *J Proteome Res* **11**:1571–1581, 2012.
16. Goh WW, Guo T, Aebersold R, Wong L, Quantitative proteomics signature profiling based on network contextualization, *Biol Direct* **10**:71, 2015.
17. Lim K, Wong L, Finding consistent disease subnetworks using PFSNet, *Bioinformatics* **30**:189–196, 2014.
18. Soh D, Dong D, Guo Y, Wong L, Finding consistent disease subnetworks across microarray datasets, *BMC Bioinformatics* **12**(Suppl 13):S15, 2011.
19. Zhang B, Wang J, Wang X *et al.*, Proteogenomic characterization of human colon and rectal cancer, *Nature* **513**:382–387, 2014.
20. Guo T, Kouvonen P, Koh CC *et al.*, Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps, *Nat Med* **21**:407–413, 2015.

21. Rost HL, Rosenberger G, Navarro P *et al.*. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data, *Nat Biotechnol* **32**:219–223, 2014.
22. Ruepp A, Brauner B, Dunger-Kaltenbach I *et al.*, CORUM: The comprehensive resource of mammalian protein complexes, *Nucl Acids Res* **36**:D646–D650, 2008.
23. Goh WW, Sergot MJ, Sng JC, Wong L, Comparative network-based recovery analysis and proteomic profiling of neurological changes in valproic acid-treated mice, *J Proteome Res* **12**:2116–2127, 2013.
24. Raju TN, William Sealy Gosset and William A. Silverman: Two "students" of science, *Pediatrics* **116**:732–735, 2005.
25. Isabelle G, Jason W, Stephen B, Vladimir V, Gene selection for cancer classification using support vector machines, *Mach Learn* **46**:389–422, 2002.
26. Kuhn M, Building predictive models in R using the caret package, *Journal of Statistical Software* **28**:1–26, 2008.
27. Liaw A, Wiener M, Classification and regression by randomforest, *R News* **2**:18–22, 2002.
28. Alexandros K, Julien P, Melanie H, Stability of feature selection algorithms: A study on high-dimensional spaces, *Knowl Inf Syst* **12**:95–116, 2007.
29. Venet D, Dumont JE, Detours V, Most random gene expression signatures are significantly associated with breast cancer outcome, *PLoS Comput Biol* **7**:e1002240, 2011.
30. Phipson B, Smyth GK, Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn, *Stat Appl Genet Mol Biol* **9**:Article39, 2010.
31. Allahyar A, de Ridder J, FERAL: Network-based classifier with application to breast cancer outcome prediction, *Bioinformatics* **31**:i311–i319, 2015.
32. Patil P, Bachant-Winner PO, Haibe-Kains B, Leek JT, Test set bias affects reproducibility of gene signatures, *Bioinformatics* **31**:2318–2323, 2015.
33. Lim K, Li Z, Choi KP, Wong L, A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small, *J Bioinformatics Comput Biol* **13**:1550018, 2015.



Wilson Wen Bin Goh is an Associate Professor of Bioinformatics at the School of Pharmaceutical Science and Technology, and the Department of Bioengineering, Tianjin University. He works on multiple applications in bioinformatics including network biology and clinical proteomics. He received his BSc (Biology) in 2005 from the National University of Singapore and his MSc/PhD in 2014 from Imperial College London.



Limsoon Wong is KITHCT 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, named in 2013 for his contributions to database theory and computational biology. Some of his other recent 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. He received his BSc(Eng) in 1988 from Imperial College London and his PhD in 1994 from University of Pennsylvania.