Journal of Bioinformatics and Computational Biology
Vol. 14, No. 5 (2016) 1644004 (18 pages)

DOI: 10.1142/S0219720016440042

World Scientific
www.worldscientific.com

Spectra-first feature analysis in clinical proteomics — A case study in renal cancer

Wilson Wen Bin Goh[*,‡,¶] and Limsoon Wong[†,§,¶]

[*]*School of Pharmaceutical Science and Technology*
Tianjin University, 92 Weijin Road
Tianjin 300072, P. R. 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*

Accepted 24 August 2016
Published 30 September 2016

In proteomics, useful signal may be unobserved or lost due to the lack of confident peptide-spectral matches. Selection of differential spectra, followed by associative peptide/protein mapping may be a complementary strategy for improving sensitivity and comprehensiveness of analysis (spectra-first paradigm). This approach is complementary to the standard approach where functional analysis is performed only on the finalized protein list assembled from identified peptides from the spectra (protein-first paradigm). Based on a case study of renal cancer, we introduce a simple spectra-binning approach, MZ-bin. We demonstrate that differential spectra feature selection using MZ-bin is class-discriminative and can trace relevant proteins via spectra associative mapping. Moreover, proteins identified in this manner are more biologically coherent than those selected directly from the finalized protein list. Analysis of constituent peptides per protein reveals high expression inconsistency, suggesting that the measured protein expressions are in fact, poor approximations of true protein levels. Moreover, analysis at the level of constituent peptides may provide higher resolution insight into the underlying biology: Via MZ-bin, we identified for the first time differential splice forms for the known renal cancer marker MAPT. We conclude that the spectra-first analysis paradigm is a complementary strategy to the traditional protein-first paradigm and can provide deeper level insight.

Keywords: Mass spectrometry; data independent acquisition; SWATH; proteomics; feature-selection.

1. Introduction

Mass-spectrometry (MS)-based proteomics is a critical technology in high-throughput biological research as it is our primary means of proteome characterization, and

¶Corresponding authors.

has key clinical applications. Unfortunately, the data it generates is noisy, and difficult to analyze comprehensively (incomplete proteome coverage and inter-sample inconsistency) without incorporating other biological data sources as contextualization frameworks (e.g. networks and protein complexes.[1–4]

In a typical MS run, as low as 30% of acquired spectra are confidently mapped to known peptides.[5] On one hand, this means a large proportion of remaining spectra are left unused (due to low-confidence matches, noise, or unspecified post-translational modifications). On the other hand, since only a handful of peptides are identifiable per protein, the finalized protein expression may not be accurate.

Traditionally, functional analysis of proteomics data relies on the protein-first paradigm, where spectra are first matched to peptides peptide-spectra matches (PSMs), filtered based on some statistical threshold, and the finalized protein list assembled based on the statistically significant PSMs. In most cases, peptide-spectra matching is based on library search algorithms, which assign PSMs between theoretical spectra from known peptides and real spectra. This is an error-prone process. One may increase the statistical stringency to reduce the number of false positives but doing so increases the false negatives, resulting in lower proteome coverage. Alternatively, one may use the false discovery rate (FDR) e.g. via decoy sequences matches.[6] However, this is an inferential procedure, and an inferred 1% FDR based on the decoy, even if unbiased, does not guarantee the inclusion of all good quality PSMs.[7] Useful information is still lost nonetheless.

Although the protein-first paradigm is the *de facto* standard, and it is possible to work with the protein lists it generates, perhaps more or other means of analysis can be done with raw spectra. One alternative to the protein-first paradigm is to perform feature selection on raw spectra directly before mapping them to peptides/proteins. We refer to this as the spectra-first paradigm where differential spectra are first identified followed by associative peptide matching and functional analysis. Unlike the protein-first paradigm, spectra-first strategies are likely more sensitive although this may come at the cost of lower precision, as raw spectra are highly redundant and noisy. Spectra-first paradigm is not an entirely new concept, traditional methods in this class of approaches include PPC,[8] BinDa[9] and CAMS-RS.[5]

As proteomics data advances from the traditional data-dependent acquisition (DDA) to the new data-independent acquisition (DIA) methods, the spectra-first paradigm may still be a useful and interesting complement to traditional protein lists generated from the protein-first paradigm. Against an extensively analyzed renal cancer dataset based on DIA,[10] we have designed and tested a spectra-first heuristic, MZ-bin, and determined if the differential spectra features it selects map to phenotypically-relevant proteins. By focusing on differential spectra first, we are also interested to know if this may increase resolution at the level of proteome expression analysis.

2. Material and Methods

2.1. *MZ-bin design*

MZ-bin is a heuristic that deploys iterative spectral feature selection on three steps: A/Retention-time (RT)-Condensation, B/Binning and C/Deconvolution (Fig. 1).

In RT-condensation, spectra intensities with same mass-to-charge (m/z) ratio are summed and the RT is ignored (Fig. 1(A)).

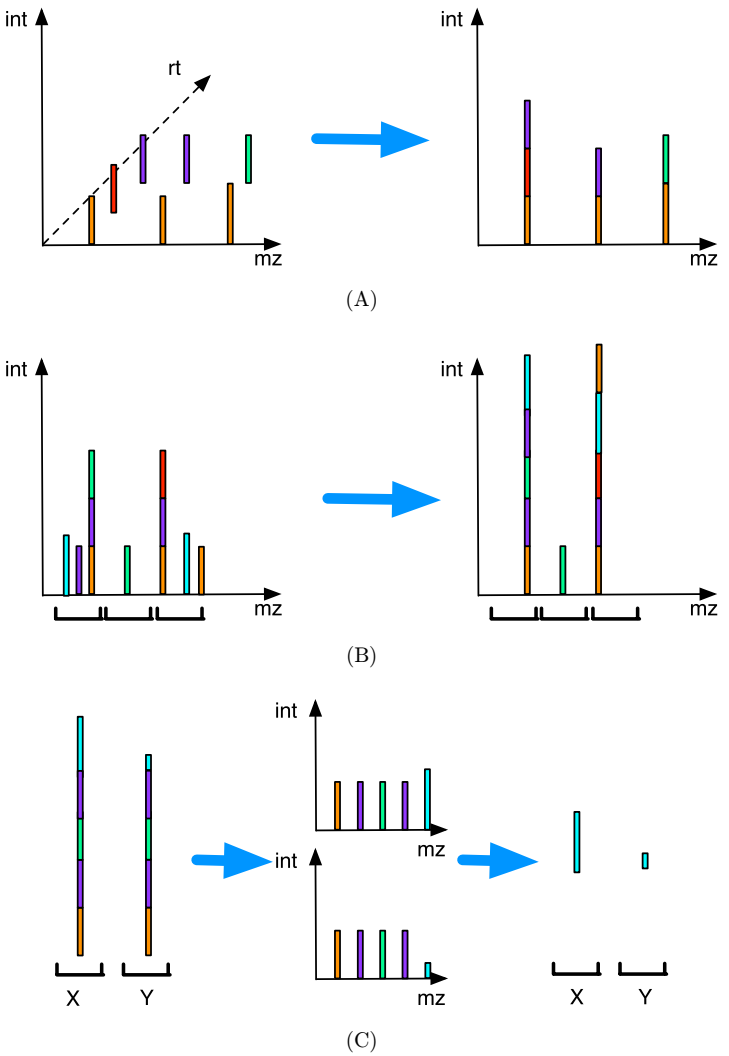


Fig. 1. Illustration of MZ-bin concepts. (A) RT-condensation. The intensity of spectra with same mass-to-charge (m/z) ratio is summed irrespective of retention time (rt). (B) MZ-binning. Given an m/z range, the intensities of overlapping spectra are combined together. (C) MZ-bin deconvolution. If an MZ-bin is differential, it is dismantled iteratively, until we reach the lowest-level MZ-bin(s) that can explain the differential result between X and Y.

Binning is the process of generating spectral bins based on m/z windows (which we also refer to as MZ-bins) (Fig. 1(B)). The m/z windows are nested to give several levels of MZ-bins (from level 1... level n) such that at level n, a spectrum with m/z x is mapped to the MZ-bin $\lfloor((10^{(n-1}\times x) + 0.5)/10^{(n-1)}\rfloor$, where $\lfloor \quad \rfloor$ is a function returning the largest integer less than or equal to a given number.

To see how this works: consider an MZ-bin at level 2. Given the formulation above, a level 2 MZ-bin can be expressed as 10.1, 10.2, etc. A level 2 MZ-bin 10.1, would include spectra with m/z values of 10.1, 10.12, 10.123, 10.14, etc. A deeper-level MZ-bin has higher resolution but smaller m/z range (since m/z ranges are nested by levels). Consider a level 3 MZ-bin 10.12. Spectral values with m/z of 10.12, 10.123 and 10.1234 fall within it, but not m/z of 10.10 or 10.14.

In deconvolution, for each MZ-bin (at the level being tested), we compare the summed intensity scores of this MZ-bin for samples across classes X and Y. If the comparison is significant (based on feature-selection test procedure; see below), we can deconvolute the MZ-bin by dismantling it iteratively into MZ-bins one level deeper (recall that MZ-bins are nested by levels), until we reach the deepest or lowest-level MZ-bin(s) that can explain the differential result. Figure 1(C) shows one MZ-bin dismantling event which revealed that the differential signal originates from the blue bars, which becomes increasingly isolated as we proceed down each MZ-bin level.

To elaborate further, for the set of significant MZ-bins selected at the $(n-l)$-th level, we iteratively work down to the nth level, and reselect the critical MZ-bins based on the t-test with an alpha of 0.01 (i.e. $p < 0.01$). For example, we can only select MZ-bin 10.123 provided we have earlier selected MZ-bin 10.12, and we can only select MZ-bin 10.12 provided we have even earlier selected MZ-bin 10.1, and we can only select MZ-bin 10.1 provided in the beginning we have selected MZ-bin 10. Since the limit of the machine resolution for m/z measurements is only up to four decimal places, the iterative MZ-bin expansion procedure can theoretically proceed up to four levels (e.g. 10, 10.1, 10.12 and 10.123).

MZ-bin can potentially deal with misalignments and stochastic variation issues: the spectra attributable to a given peptide are expected to shift depending on stochastic and experimental variation. Suppose a peak px at m/z $= x$ in subject sx actually corresponded to a peak py at $m/z = y$ (where $y \neq x$) in another subject sy. Without binning, the peaks px and py would be regarded as distinct peaks and thus there would be no peak (i.e. a hole) at m/z $= y$ in subject sx and at m/z $= x$ in subject sy. If subjects sx and sy belong to the same class, the hole at m/z $= y$ in sx would weaken the test statistic for the peak py, and the hole at mz $= x$ in sy would weaken the test statistic for the peak px. This increases the likelihood of peaks px and py being declared insignificant/noninformative. In contrast, suppose the peaks px and py were in the same MZ-bin, then there would not be a hole in sx and sy for this MZ-bin.

To demonstrate this, we considered the effects of spectral-feature selection without binning (Supplementary Fig. 1(A)) where we observed a large number of

missing data points. Furthermore, the unbinned spectra features are uninformative (Supplementary Fig. 1(B)). An analogous argument also applies for rt shifts.

In practice, we do not expect most proteins to be differential between sample classes, i.e. differential features are rare. Suppose in a comparison between classes A and B, an MZ-bin contains a peptide from a protein X and a peptide from a protein Y. Suppose also that protein X is differential in its abundance levels in the two phenotype classes, but protein Y is not. Thus, the MZ-bin's differential signal is dominated by the peptide from X Summing m/z intensities from X and Y wipes out the irrelevant Y but still maintains the signal. To demonstrate this, we plotted the distribution of signal intensities for each contributing spectra in each level 1 MZ-bin (Supplementary Fig. 1(C)). Most of the signal intensity in any MZ-bin is attributable to a very small number of spectra. On average, only 20 spectral features are needed to account for 25% of total intensity.

2.1.1. *Rules-based feature selection procedure*

For feature selection (i.e. selection of differential MZ-bins), we can compare the corresponding MZ-bins between samples from classes A and B using a simple feature-selection method, e.g. two-sample *t*-test at an alpha of 0.01 (see Supplementary methods).[11] However, this may not be robust enough due to potential issues with MS data quality. Hence, we introduced some refinement rules:

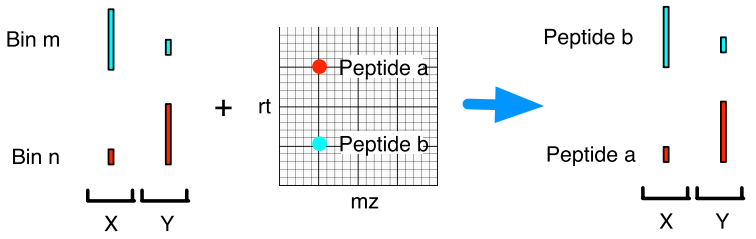
Rule 1. If an MZ-bin has nonzero intensity values in more than half of the samples in class A and nonzero in more than half of the samples in class B, it is kept (for further filtering by Rule 2 below). Otherwise, it is discarded.

Rule 2. For each class (A or B), the top 20% MZ-bins (ranked by summed intensity) supported by at least half of the samples in that class are kept.

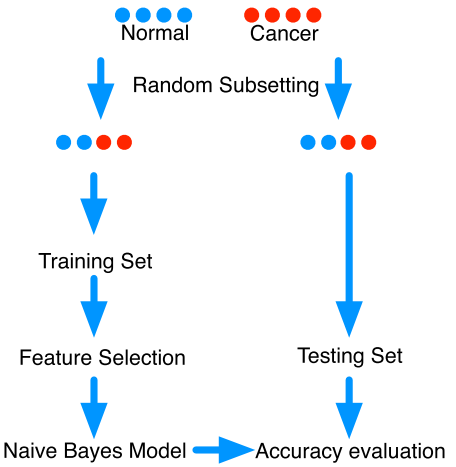
These rules are sensible and have been used for feature selection on protein-based expression data. For Rule 1, if the majority of intensity values are 0, then there is very little evidence (amongst samples) to support the existence or summed intensity value of that particular MZ-bin. For Rule 2, there are two lines of evidence from recent works: One comes from quantitative proteomics signature profiling (QPSP) where low-abundance proteins (and by implication, low-abundance MZ-bins) have very high coefficient of variation and thus rather unstable.[12] Another evidence comes from work on Paired Fuzzy SubNETs (PFSNET) analysis where restricting the top 20% proteins (and by implication the top 20% MZ-bins) improves stability and does not impact sensitivity.[13]

2.2. *Associative peptide mapping*

The MS setup has been pre-calibrated with known proteins such that the m/z and rt coordinates of their constituent peptides are known.[10] Since the rt coordinates are irrelevant due to the binning process, each differential MZ-bin can be mapped to known peptides (and the originating protein) based on m/z overlap (Fig. 2(A)).



(A)



(B)

Fig. 2. Associative peptide mapping and cross-validation procedure. (A) Associative peptide mapping. Significant MZ-bins are mapped to known peptides based on m/z overlap, i.e. the m/z coordinate of a known peptide falls within the m/z range of an MZ-bin. (B) Cross-validation evaluation. Normal and cancer samples are split randomly to form a training set and testing set. The training set is used for model building which is evaluated with the testing set. Accuracy is the proportion of a sample class which is correctly predicted over all predictions.

2.3. *Cross-validation*

Cross-validation (CV) evaluation tests the reproducibility of a method based on different subsets derived from the dataset. Since we do not have large numbers of samples for this MS screen, we repeated CV 10 times on 10 different equal random splits of the data into training and testing sets. Each split maintains the same proportion of cancer and normal tissues as in the original un-split data set. For each split, a Naïve Bayes classifier is trained on the differential MZ-bins and validated on the test set for accuracy. The CV accuracy is calculated as:

$$\text{CV Accuracy} = \frac{\text{Number of correct class assignments}}{\text{Size of testing set}}.$$

To determine if CV accuracy is meaningful, we randomly picked an equal number of features 1,000 times and retrained the Naïve Bayes classifier to produce a vector of null accuracy values. The CV accuracy p-value is the number of times null accuracy beats the actual accuracy divided over total number of simulations.

2.4. *False-positive analysis*

Samples from the normal sample class are randomly split 1,000 times into two pseudo-groups, followed by conversion into MZ-bins. Differential MZ-bins selected via the two-sample t-test here are considered false positives.

2.5. *Test proteomics data*

The SWATH-MS proteomics dataset of renal cancer is used for evaluating MZ-bin.[10] This well-characterized dataset contains 24 MS runs derived from six pairs of non-tumorous and tumorous clear-cell renal carcinoma (ccRCC) tissues, with two technical replicates each (12 normal, 12 cancer). For more details, refer to supplementary methods.

A spectral library containing 49,959 reference spectra from 4,624 reviewed SwissProt proteins[10] is compiled on the same MS setup. We use this for associative peptide mapping following MZ-bin.

3. Results and Discussions

3.1. *Evaluating MZ-bin as a feature-selection method*

Raw spectra cannot be directly used for feature selection. Preliminary checks revealed large numbers of data holes due to stochastic effects and misalignments during spectral acquisition. We find that raw spectral features are also poorly predictive (Supplementary Figs. 1(A) and 1(B)). Iterative deconvolution of differential level 1 MZ-bins reveals that differential signal is dominated by small number of constituent spectra in each bin. Thus, iterative deconvolution of the differential MZ-bin should isolate the relevant spectrum (and its associated peptides/proteins) (Supplementary Figs. 1(C) and 1(D)).

To prevent poor-quality MZ-bins from being selected as differential features, we introduce two refinement rules to exclude MZ-bins with many missing values or whose summed intensities are generally low. We compare this to feature selection without refinement rules, using MZ-bins level 2 and 3 as case example (Table 1). At level 2, rule-based feature selection identified less than half the number of features while at level 3, it was less than one-fifth. At both levels, features derived from rule-based filtering maintained similar if not higher CV accuracy. This suggests that rule-based feature-selection identifies relevant high-quality spectral bins.

We further demonstrate that the high CV accuracy of MZ-bin is meaningful via p-value estimation (Table 2). Here, the p-value is the proportion of simulations where randomly picked MZ-bins produce higher accuracy. It turns out that randomly

Table 1. CV accuracy of levels 2 and 3 MZ-bins with and without rule-based feature filtering (A and B respectively). The rule-based MZ-bin selection process expectedly predicts less features than nonrule based (standalone *t*-test), while maintaining similar CV accuracy. This observation is consistent, as shown for levels 2 and 3.

A				
	Rule-based feature selection			
	MS1 merged level 2		MS1 merged level 3	
Group	No. of significant features (0.01)	CV Accuracy	No. of significant features (0.01)	CV Accuracy
1	130	0.83	171	0.75
2	402	0.75	438	0.75
3	22	0.83	87	0.83
4	261	0.83	274	0.75
5	389	0.83	608	0.75
6	141	0.83	299	0.83
7	140	0.75	270	0.83
8	153	0.75	265	0.92
9	184	0.83	354	0.83
10	30	0.92	53	0.92
mean	185.20	0.82	281.90	0.82
s.d.	130.42	0.05	163.08	0.07
COV	0.70	0.06	0.58	0.08
B				
	t-test only selection			
	MS1 merged level 2		MS1 merged level 3	
Group	No. of significant features (0.01)	CV Accuracy	No. of significant features (0.01)	CV Accuracy
1	65	0.83	451	0.83
2	372	0.75	1290	0.75
3	198	0.83	887	0.75
4	694	0.75	1854	0.75
5	746	0.75	2717	0.75
6	608	0.83	2279	0.75
7	190	0.92	1354	0.92
8	370	0.83	1784	0.83
9	1199	0.83	3471	0.75
10	189	0.83	399	0.92
mean	463.10	0.82	1648.60	0.80
s.d.	348.22	0.05	983.85	0.07
COV	0.75	0.06	0.60	0.09

picked MZ-bins are almost never more accurate (Supplementary Fig. 2). Hence, the observed CV accuracy is meaningful.

In contrast, if feature selection was performed using the two-sample *t*-test solely at the level of individual proteins derived from the protein-first paradigm (Single Proteins, SP), an inordinately large number of proteins are selected as differential

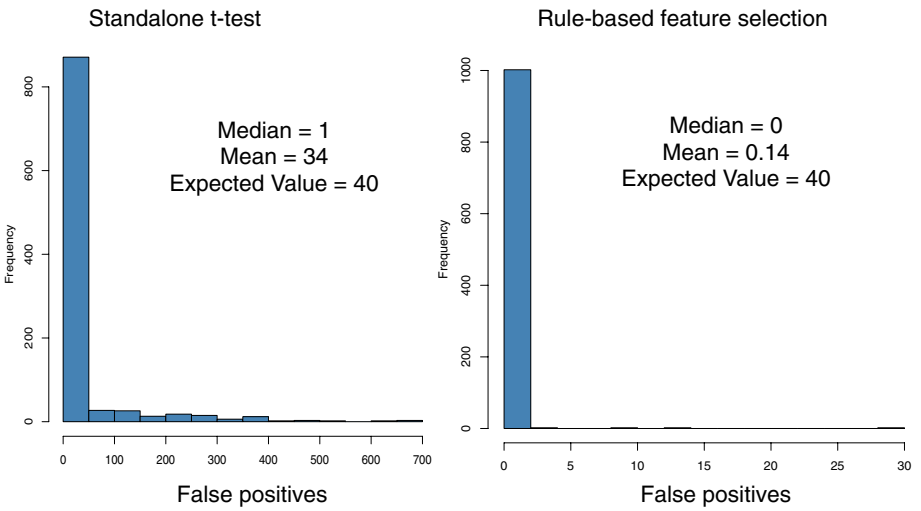
Table 2. CV accuracy comparing MZ-bin and protein-based expression (Single Proteins, SP). Although CV accuracy is high in SP, any random selection of proteins yields equally high CV accuracy (cf. supplementary Fig. 2). On the other hand, although CV accuracy of MZ-bin is lower, the results are statistically more meaningful; i.e. random selection of MZ-bin cannot produce higher CV accuracy.

	MS1 filtered merged level 1 spectra			Single protein		
Group	No. of significant features (0.05)	CV Accuracy	CV p-value	No. of significant features (0.05)	CV Accuracy	CV p-value
1	82.00	0.83	0.00	901	1.00	0.92
2	130.00	0.75	0.00	1213	1.00	0.911
3	179.00	0.75	0.00	908	1.00	0.919
4	234.00	0.75	0.00	1210	1.00	0.905
5	220.00	0.75	0.00	800	1.00	0.915
6	158.00	0.75	0.00	1825	0.75	0.892
7	96.00	0.92	0.00	1017	1.00	0.902
8	166.00	0.83	0.00	1284	1.00	0.925
9	230.00	0.75	0.00	1251	1.00	0.917
10	98.00	0.92	0.00	834	1.00	0.92
mean	159.30	0.80	0.00	1124.30	0.98	0.91
s.d.	167.89	0.80	0.00	306.48	0.08	0.01
COV	171.14	0.80	0.00	0.27	0.08	0.01

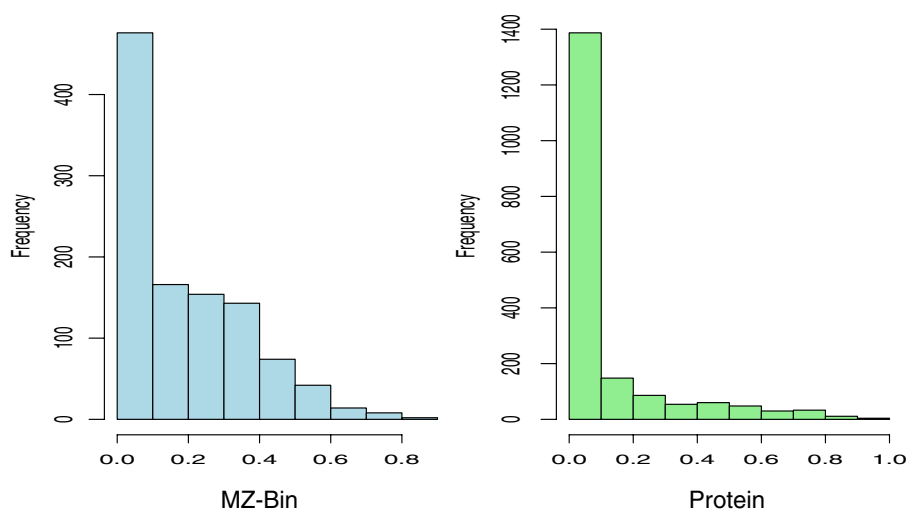
features each time (Table 2). Although CV accuracy appears high, almost any random subset of proteins (containing 5, 20 or 100 randomly picked proteins) performs equally or better (Supplementary Fig. 2). In contrast, random selection of features in MZ-bin did not generate skewed null distributions (i.e., very few randomly picked MZ-bins generate null CV accuracy > 0.95); cf. Supplementary Fig. 2.

In MZ-bin level 1, the number of false positives given the standalone and rules-based t-test fall within expectation (Fig. 3). However, the latter turns out more stringent. As with the CV accuracy results, rule-based refinement is beneficial.

Feature selection at the level of MZ-bins is more reproducible: To illustrate this, we used the feature-reproducibility analysis method described in QPSP.[12,14] Briefly, using the level 2 MZ-bin matrix following rules-based filtering, random samplings of size six (six from Normal and six from Cancer) were taken 1000 times, and feature selection was performed using the t-test at a p-value threshold of 0.05. An equal number of significant protein features are taken from the corresponding protein expression matrix each time (the top n based on ranking by p-value). Resampling was performed 1000 times. For each MZ-bin or protein feature that has been observed at least once, its selection reproducibility is taken as the proportion of times it was reported as significant over all 1,000 resampling. Feature-selection reproducibility can be discerned by plotting the frequency distribution of selection reproducibility where a right shift would suggest more features are reproducibly selected in spite of resampling (Fig. 3(B)). We observe a stronger right shift for MZ-bin with a median selection reproducibility of 0.133 whereas for its corresponding proteins, the median selection reproducibility is 0.022. In other words, the significant features



(A)



(B)

Fig. 3. MZ-bin false-positive rates and feature-selection reproducibility analysis. (A) MZ-bin false-positive rates. Using control samples derived from normal tissues only, and randomly splitting these into two groups, we tested the level 1 MZ-bins 1,000 times using standalone t-test and rule-based feature selection. The number of false positives is within expectation, with a median $= 1$ and mean $= 34$ (expected value $= 800 \times 0.05 = 40$), confirming that the MZ-bin approach does not generate overly high noise levels. However, the rule-based feature-selection strategy is even more stringent, with lower false positive rate. (B) Feature-selection reproducibility analysis. (x-axis: proportion of time a significant feature is observed over 1000 simulations, y-axis: frequency). Following MZ-bin feature selection with rules at MZ-bin level 2, random sampling of size six was performed 1000 times using MZ-bin, and an equal number of protein features selected on the corresponding protein expression matrix. Feature selection at the level of MZ-bins is more reproducible, with a median feature-selection reproducibility of 0.133 against 0.022 for protein-based feature-selection.

selected by MZ-bin are about 10 times more reproducible in different subsamples than those selected by SP.

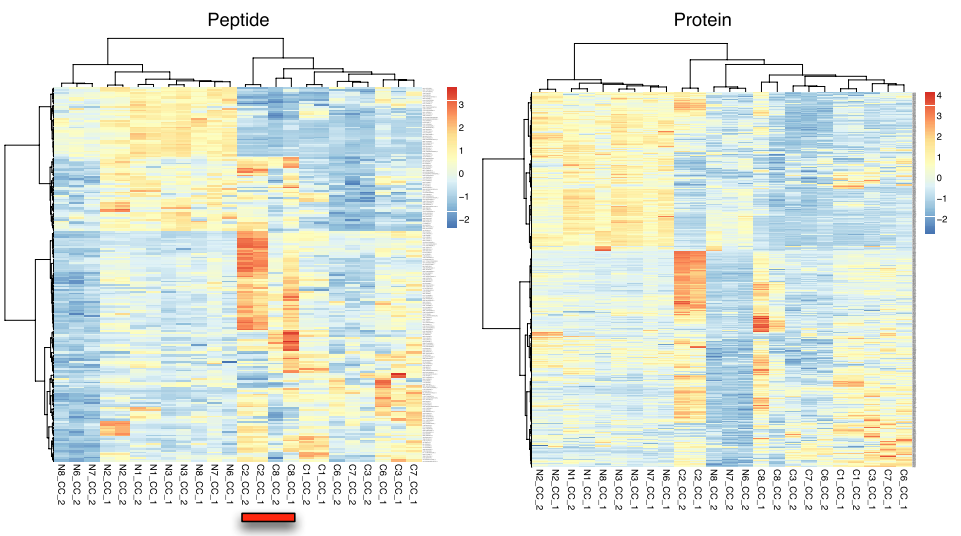
Iterative deconvolution cannot proceed indefinitely given the limits of instrument detection. Here, we determine the limit is level 3 as none of the level 4 bins can satisfactorily pass rule-based feature selection (see supplementary results: *Analytical limits of MZ-bin*).

3.2. *MZ-bin associated peptides have strong class discrimination signal*

Across MZ-bin levels 1–3, we check the peptide groups and proteins associated with the differential MZ-bins (Fig. 4(A)). As MZ-bin level progresses from 1 to 3, the

Level	Sig_bins	pep_groups	proteins
1	127	8249	2389
2	182	3018	1470
3	337	1044	688

(A)



(B)

Fig. 4. Peptide/protein features associated with significant MZ-bins from levels 1 to 3. (A) Numbers of associated peptides/proteins. With each MZ-bin iteration, the number of significant features generally increases, with a concomitant decrease in the number of associated peptides/proteins. (B) Class segregation for peptides and proteins. Hierarchical clustering (Euclidean distance; Ward's Linkage) shows that the significant peptides selected based on level 3 MZ-bins can clearly separate the phenotype classes (Notation: N7_CC_2 refers to normal sample 7, clear-cell renal carcinoma, replicate 2). It is particularly interesting that patients C2 and C8, who suffered from a severe form of the disease, are grouped together. In contrast, the proteins corresponding to these peptides seem to have poorer discrimination, since patients N6, N7 and N8 seem to have been misclassified in the cancer branch. This suggests there is some loss of information in the peptide-to-protein transition.

number of differential MZ-bins increases alongside concomitant decrease in associated proteins. This is unsurprising since MZ-bins are nested by levels. For example, if MZ-bin 10 (level 1) is significant, it might be due to several of its nested levels 2 or 3 MZ-bins being significant, but not necessarily all of them. Thus, while peptides whose m/z values were in MZ-bin 10 are declared significant when the analysis was performed at level 1, they may not be declared significant when analyzed at level 3 if their m/z values are not in any of the differential nested level 3 MZ-bins. At level 3,337 differential MZ-bins are associated with 1,044 peptide groups corresponding to 688 unique proteins.

Via hierarchical clustering (Euclidean distance; Ward's linkage), class-discrimination based on peptide intensities is strong. Technical replicates are also closely grouped together. Patients 2 and 8 (severe renal cancer subgroup) are also banded together (Fig. 4(B) left). This is consistent with previous observation.[12] Together, the results suggest that the MZ-bins are relevant.

Translating peptide expression level to protein expression level is not straightforward. For example, one may take the mean or median of all unique constituent peptides. Alternatively, in this dataset, the authors used the top two constituent peptides per protein for quantitation. However, we are concerned the top two peptides per protein may not be the same across different samples. Furthermore, the two most-abundant peptides need not be differential between sample classes. This may lead to potential loss-of-signal. It turns out that this concern is valid: class discrimination is less pronounced at the corresponding protein expression level (Fig. 4(B) right).

3.3. *Proteins associated with differential MZ-bins are more biologically coherent*

MZ-bin associated proteins have good class-discrimination power. But they are not the same set of differential proteins if a t-test is performed on the protein expression list derived from the protein-first paradigm (SP). 1,247 out of 1,649 proteins are SP-unique, and not associated with significant level 3 MZ-bins. It is useful to determine which set of differential proteins (from MZ-bin or t-test selection directly from the protein list) are more biologically coherent.

To do this, we check which differential protein set tends to cluster together in same networks, with the reasoning that the more coherent list of differential proteins are more likely to work together, and therefore will be closely located on the reference biological network.[15–19] We may use biological complexes to check this. Using 1,363 protein complexes,[20] we generate all possible pairs of differential MZ-bin proteins and determine the fraction of these pairs that hit the same complexes. This is repeated for SP-unique proteins.

After accounting for the smaller number of MZ-bin proteins, the log odds for MZ-bin proteins against SP-unique proteins is $\sim 1.22x$. Therefore, there is stronger propensity for MZ-bin proteins to co-locate in the same complexes. Hence, we determine the MZ-bin differential protein list is more biologically coherent.

3.4. *Using MZ-bin to search for splice variants — novel splice forms of MAPT differentially associated with good and poor renal-cancer outcomes*

We may use the associated peptides in differential level 3 MZ-bins to check for splice variants in corresponding proteins. We first isolate all peptides that can be unambiguously mapped to the 688 level 3 proteins. For any of these, if all constituent peptides are similarly up- or down-regulated, then the abundance of the entire protein is likely to be regulated at the transcriptional level. However, if the constituent peptides are inconsistently expressed, this may suggest the presence of alternative splice events.

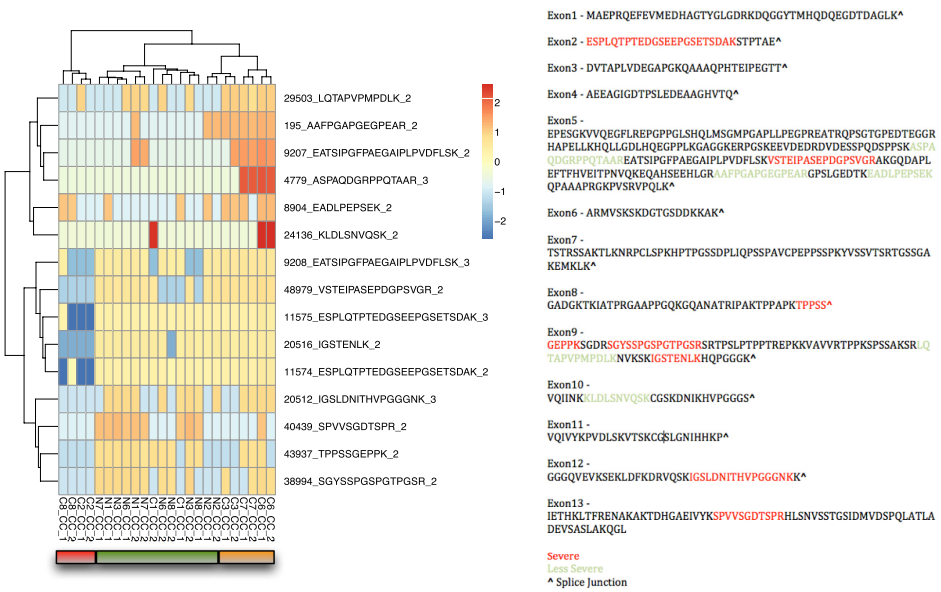
We first check for peptides with poor quantitation stability (missing data in more than half of either sample class) and flag them (Supplementary Fig. 3(A)). Interestingly, most of the 688 proteins do not have consistent constituent peptide abundance (supplementary data 1), which suggests that in protein-first paradigm, the final reported protein expression is really very rough approximation of true expression level. Furthermore, while these proteins are expected to have at least one peptide associated with a differential MZ-bin, it turns out that many of the other constituent peptides are in fact nondifferential. To increase stringency and refine the search for alternatively spliced proteins, we introduced the following rules:

(1) There must be at least 10 constituent unique peptides (to ensure reasonable coverage of the entire length of the protein);
(2) The peptides must be unambiguously mapped to the corresponding protein;
(3) at least 30% constituent peptides are over-expressed, i.e. > 1.25; and
(4) at least 30% constituent peptides are repressed, i.e. < 0.8.

Although four proteins fulfilled these criteria (Supplementary Fig. 3(B)), microtubule-associated protein tau (MAPT, P10636) is particularly interesting, separating severe from other less severe cancer patients.

MAPT is differentially enriched for different peptides in severe cancer (C2 and C8 — highlighted in red) (Fig. 5(A)). MAPT is commonly associated with neurological diseases such as dementia and is known to have large numbers of splice forms.[21] Interestingly, MAPT forms part of a predictive gene signature in severe renal cancer and has been reported as down-regulated.[22] But the existence of specific differential splice forms able to distinguish between severe and nonsevere renal cancer is not known.

The peptides discriminative for severe and less severe renal cancer are evenly distributed across the full MAPT protein sequence (Supplementary Fig. 4). To determine if the discriminative peptides are localized within the splice junctions, we used genewise to map the MAPT protein sequence against the MAPT unspliced DNA sequence (supplementary data 2) where we predicted 13 exons (Fig. 5(B)).[23] Mostly, peptides discriminative for severe and less severe respectively are located on different exons with the exception of exons 5 and 9. This suggests that certain



(A) (B)

Fig. 5. Peptide features associated with MAPT. (A) Hierarchical clustering using MAPT peptides. MAPT peptides are differential between severe (red) and less severe cancers (orange). This suggests that these peptides may be useful as markers for prognosis. (B) Localization of MAPT differential peptides within exon junction. For the most part, severe and less severe peptides are located within different exons, except for exons 5 and 9. This suggests that there may be patients with mutations within these regions that may generate novel splice sites within these exons.

splicing junctions are disrupted during tumorigenic progression, possibly due to mutations that generates novel splice sites.

To see if our differential peptides might correspond to any known splice forms, we picked the two most discriminatory peptides from "severe" and "less severe" groups (ASPAQDGRPPQTAAR and KLDLSNVQSK for the less severe group, and ESPLQTPTEDGSEEPGSETSDAK and IGSTENLK to represent the severe group) (Fig. 5). We compared these sequences to eight known splice forms of MAPT (UniprotKB, supplementary data 3) using T-Coffee (default parameters).[24]

ESPLQTPTEDGSEEPGSETSDAK is found on splice forms 4–9, while IGSTENLK is found across all splice forms. While ASPAQDGRPPQTAAR is found across all splice forms and KLDLSNVQS is found only in splice forms 6–9. Forms 7–9 are common to both peptide groups. Forms 4 and 5 are unique to severe cancer associated peptides. Form 6 is unique to less severe cancer. Perhaps it is these splice forms themselves that are differentially expressed.

We discuss first ASPAQDGRPPQTAAR (less severe) and IGSTENLK (severe), which are found in all splice forms. In our opinion, these two peptides make perfect

biomarkers as: (1) they can be detected in everyone and (2) their abundance is completely distinct between less severe, severe and normal tissue.

Using InterPro[25,26] to predict domain information, ASPAQDGRPPQTAAR is on exon 5 (Fig. 5(B)), and corresponds to Microtubule associated protein MAP2/MAP4/Tau (IPR027324). This domain has a net negative charge and exerts a long-range repulsive force. This provides a mechanism that can regulate microtubule spacing which facilitate organelle transport.[27] IGSTENLK is found on exon 9 (Fig. 5(B)) and corresponds to several domains — IPR027324, MAPT (IPR002955) associated with microtubule binding, and Microtubule associated protein, tubulin-binding repeat (IPR001084) which is implicated in tubulin-binding and has a stiffening effect on microtubules.

Exon 5 enriched for peptides associated with less severe cancer (3), which implies its over-expression may be a positive prognosis. Disrupting this, as seen with VSTEIPASEPDGPSVGR, could potentially reverse this either by mutating part of the exon, or by overall down-regulation of this region. Exon 9 on the other hand, is enriched for peptides associated with severe cancer, its overall downregulation can be a poor prognosis indicator. Within exon 9, a single peptide, TAPVPMPDLK is found to be overexpressed and associated with less severe phenotype. Perhaps, increasing the expression of IPR001084 stabilizes the microtubules, and makes it harder for the cancer to undergo metastasis.

To discover more interesting domain-specific information associated with the alternative splice forms, we aligned the sequences of splice forms 4, 5 (severe) and 6 (less severe) using T-Coffee. We then extracted two representative domain sequences — ESPLQTPTEDGSEEPGSETSDAKSTPTAEDVTAPLVDEGAPG-KQAAAQPHTEIPEGTT for severe (severe domain) and QIINKKLDLSNV-QSKCGSKDNIKHVPGGGSV for less severe (less severe domain) and checked if these corresponded to any known domains using InterPro.[25,26] The severe domain is found on exon 2 and corresponds to IPR027324 while the less severe domain is found on exon 10 and maps to both IPR027324 and IPR001084. As before, the repression of exon 2 suggests this region is potentially impaired in severe cancer. Similarly, overexpression of exon 10 may increase microtubule stability, impeding metastasis.

Upregulated MAPT is a known good-prognosis indicator in renal cancer while its down-regulation means the opposite.[28] Based on MS analysis, we report here that the dysregulation is, in fact, inconsistent across its entire length. Certain peptide regions are specifically over-expressed for less severe renal cancer while other non-overlapping regions are repressed for severe renal cancer. Consideration of these specific peptide regions is more useful as prognostic markers than using the entire protein length as this will dilute diagnostic signal.

4. Conclusions

MZ-bin is a rule-based heuristic for iteratively identifying relevant features from complex spectra without the need for prior spectra clustering or peptide-spectra

assignments. Despite the simplicity of the rules, the selected differential features — i.e. the condensed spectra — have good predictive power and reproducibility. We further demonstrate that the selected features are phenotypically relevant, on a renal cancer dataset derived from SWATH-MS.

Furthermore, careful consideration of constituent peptides reveal that reported expression levels at the level of proteins cannot be fully trusted. Abundance inconsistencies amongst constituent peptides and presence of splice forms may mislead.

Lastly, we should highlight a technical aspect that we have not pursued here. Raw spectra are noisy. Stringent statistical thresholds are thus used in determining PSMs. As a result, many spectra are discarded. Nonetheless, considering only differential spectra (rather than all raw spectra) potentially requires less stringent the statistical thresholds in the mapping of this subset of spectra to peptides. For example, if one corrects the p-value of PSMs using the Bonferroni method, given 100,000 raw spectra, the p-value threshold is 5×10^{-7} to achieve a false-positive rate of 5%. Suppose only 10% of the spectra are differential, and PSMs are sought only for these 10%, the Bonferroni-corrected p-value threshold is 5×10^{-6} to achieve a false-positive rate of 5%. Thus, one can potentially gain sensitivity without increasing false-positive rate of the PSMs. That said, we have conservatively stuck to PSMs determined using statistical thresholds based on the entire raw spectra, rather than based on differential spectra. The effect described in this example may be worth investigating in a follow-up study.

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

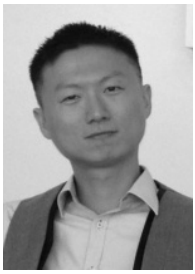
The authors declare that they have no competing interests.

References

1. Goh WW, Wong L, Networks in proteomics analysis of cancer, *Curr Opin Biotechnol* **24**:1122–1128, 2013.
2. Goh WW, Wong L, Computational proteomics: Designing a comprehensive analytical strategy, *Drug Discov Today* **19**(3):266–274, 2013.
3. Goh WW, Wong L, Sng JC, Contemporary network proteomics and its requirements, *Biology* **3**:22–38, 2013.
4. Sajic T, Liu Y, Aebersold R, Using data-independent, high-resolution mass spectrometry in protein biomarker research: Perspectives and clinical applications, *Proteomics Clin Appl* **9**:307–321, 2015.
5. Saeed F, Hoffert JD, Knepper MA, CAMS-RS: Clustering algorithm for large-scale mass spectrometry data using restricted search space and intelligent random sampling, *IEEE/ACM Trans Comput Biol Bioinform* **11**(1):128–141, 2014.

6. Granholm V, Kall L, Quality assessments of peptide-spectrum matches in shotgun proteomics, *Proteomics* **11**:1086–1093, 2011.
7. Keich U, Kertesz-Farkas A, Noble WS, Improved false discovery rate estimation procedure for shotgun proteomics, *J Proteome Res* **14**:3148–3161, 2015.
8. Tibshirani R, Hastie T, Narasimhan B *et al.*, Sample classification from protein mass spectrometry, by 'peak probability contrasts', *Bioinformatics* **20**:3034–3044, 2004.
9. Gibb S, Strimmer K, Differential protein expression and peak selection in mass spectrometry data by binary discriminant analysis, *Bioinformatics* **31**:3156–3162, 2015.
10. 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.
11. Raju TN, Gosset WS, Silverman WA, Two "students" of science, *Pediatrics* **116**:732–725, 2005.
12. Goh WW, Guo T, Aebersold R, Wong L, Quantitative proteomics signature profiling based on network contextualization, *Biol Direct* **10**:71, 2015.
13. Goh WW, Wong L, Evaluating feature-selection stability in next-generation proteomics, *J Bioinform Comput Biol* **14**(5):16500293, 2016.
14. Goh WW, Wong L, Design principles for clinical network-based proteomics, *Drug Discov Today* **21**:1130–1138, 2016.
15. Goh WW, Lee YH, Zubaidah RM *et al.*, Network-based pipeline for analyzing MS data: An application toward liver cancer, *J Proteome Res* **10**:2261–2272, 2011.
16. 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.
17. Goh WW, Lee YH, Ramdzan ZM, Chung MC, Wong L, Sergot MJ, A network-based maximum link approach towards MS identifies potentially important roles for undetected ARRB1/2 and ACTB in liver cancer progression, *Int J Bioinform Res Appl* **8**:155–170, 2012.
18. Goh WW, Lee YH, Chung M, Wong L, How advancement in biological network analysis methods empowers proteomics, *Proteomics* **12**:550–563, 2012.
19. 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.
20. Ruepp A, Waegele B, Lechner M *et al.*, CORUM: The comprehensive resource of mammalian protein complexes — 2009, *Nucleic Acids Res* **38**:D497–D501, 2010.
21. Garcia-Blanco MA, Baraniak AP, Lasda EL, Alternative splicing in disease and therapy, *Nat Biotechnol* **22**:535–546, 2004.
22. Kosari F, Parker AS, Kube DM *et al.*, Clear cell renal cell carcinoma: Gene expression analyses identify a potential signature for tumor aggressiveness, *Clin Cancer Res* **11**:5128–5139, 2005.
23. Birney E, Clamp M, Durbin R, GeneWise and Genomewise, *Genome Res* **14**:988–995, 2004.
24. Notredame C, Higgins DG, Heringa J, T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J Mol Biol* **302**:205–217, 2000.
25. Mitchell A, Chang HY, Daugherty L *et al.*, The InterPro protein families database: The classification resource after 15 years, *Nucleic Acids Res* **43**:D213–D221, 2015.
26. Apweiler R, Attwood TK, Bairoch A *et al.*, The InterPro database, an integrated documentation resource for protein families, domains and functional sites, *Nucleic Acids Res* **29**:37–40, 2001.

27. Mukhopadhyay R, Hoh JH, AFM force measurements on microtubule-associated proteins: The projection domain exerts a long-range repulsive force, *FEBS Lett* **505**:374–378, 2001.
28. Brooks SA, Brannon AR, Parker JS *et al.*, ClearCode34: A prognostic risk predictor for localized clear cell renal cell carcinoma, *Eur Urol* **66**:77–84, 2014.



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 B.Sc. (Biology) in 2005 from the National University of Singapore and his M.Sc./Ph.D. 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, inducted 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 B.Sc. (Eng) in 1988 from Imperial College London and his Ph.D. in 1994 from University of Pennsylvania.