Statistics and Its Interface Volume 5 (2012) 47–59

Bayesian false discovery rates for post-translational modification proteomics*

Yan Fu

Tandem mass spectrometry-based proteomics enables high throughput analysis of post-translational modifications (PTMs) on proteins. In current researches of shotgun proteomics, peptides with various PTMs and those without PTMs are often identified together and an overall false discovery rate (FDR) is estimated. However, it is often the case that only a subset of identifications, e.g. those with specific PTMs, are emphasized or reported. In doing so, the risk arises that the FDR of reported results is seriously under- or overestimated, based on which unreliable conclusions may be drawn. But unfortunately, this has not been widely realized in the field, and there is still no agreement on the right way to control the FDR of PTM identifications. As a result, the ostrich policy is commonly adopted wittingly or unwittingly, i.e., a simplistic overall estimate is assumed. This paper, for the first time, proves that the FDRs of various PTM identifications are in theory not equivalent to the overall FDR and quantifies several major factors influencing their relationships. Elaborate simulation experiments are carried out to empirically verify the theoretical conclusions. Strategies are suggested for better control of PTM FDRs.

Keywords and phrases: False discovery rate, Group structure, Protein identification, Post-translational modification, Proteomics.

1. INTRODUCTION

When multiple hypotheses are tested simultaneously, small *p-values* may just occur by chance if the number of tests is large enough. Researchers have realized early the potential risk in real applications that only the results with small *p-values* are selected for reporting, making the actual proportion of type I errors much higher than thought. Therefore, methods for controlling the global error rate in multiple hypothesis testing were proposed. False discovery rate (FDR) is the one that has been most intensively investigated in theory and most widely utilized in practice. FDR is defined as the expected proportion of incorrectly rejected null hypotheses among all rejected ones [3]. Nowadays, FDR control has become indispensable in throughput biological experiments such as microarray-based genomics [11] or mass spectrometry-based proteomics [6].

However, just as one can misuse *p-values*, one can also misuse FDRs. Imagine that in a study we have tested a large number of hypotheses and estimated an overall FDR for the rejected null hypotheses. Now, since only a subgroup of these rejected hypotheses are of our current interest, we only report them as discoveries. In doing so, the risk arises that the false part of the claimed discoveries may be significantly more or less than what is suggested by the estimated overall FDR, as pointed out by Efron [12]. This is because the *p-values* of the hypotheses of interest may tend to be smaller or larger than those of other hypotheses, even though all null hypotheses are true. In more formal words, if the hypotheses have some inherent group structure and this structure leads to different *p-value* sub-distributions, then the subgroups of hypotheses partitioned by the structure would have their FDRs significantly different from each other [21]. To remedy this, *p-value* weighting methods have been proposed [12, 21, 4, 20]. The most important thing is to first recognize the existence of potential structure in hypotheses. Afterward, one can avoid selecting subgroups of the hypotheses in line with the structure or use a remedy method to compensate for the bias introduced by the structure. The microarray data from different locations of an organism or from genes involved in different biological processes is a well-known example with such group structures [12, 21]. This paper shows that the same problem is faced in the mass spectrometry-based proteomics.

The proteome refers to the protein complement of a genome, and proteomics is the large-scale study of proteins expressed in an organism or system [31]. Currently, tandem mass spectrometry is the mainstay technology used for proteomics [1]. Most proteins in cells carry some post-translational modifications (PTMs) [36], and PTMs often play an essential role in protein functions and are involved in many pathological processes [37]. For example, fucosylated glycoproteins have been demonstrated to be related to several serious human diseases, such as hepatocellular carcinoma [5], pancreatic cancer [30] and lung cancer [19]. PTMs are not directly encoded in the DNA sequences and thus have to be studied in the protein level. Tandem mass spectrometry-based proteomics provides a powerful tool for large-scale analysis of PTMs [26, 38].

*This research was supported by the National Natural Science Foundation of China under Grant No. 30900262, the CAS Knowledge Innovation Program under Grant No. KGGX1-YW-13, and the National Key Basic Research & Development Program (973) of China under Grant No. 2010CB912701.

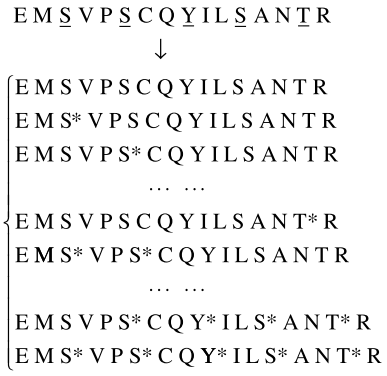

Figure 1. An illustration of generating candidate peptides with variable modifications (underlined amino acids are possible sites and modifications are denoted by stars).

In shotgun proteomics, proteins are digested into peptides and the latter are analyzed by tandem mass spectrometry. Simply speaking, a tandem mass spectrum of a peptide is a mass histogram of the fragment ions produced from peptide fragmentation. Besides, the mass of the whole peptide is also measured. When the peptide has been modified by one or more PTMs, the mass of the intact peptide and the masses of some fragment ions in the spectrum will be shifted by the mass of the PTM(s) they are carrying. Identification of peptides and the PTMs on them is usually accomplished by searching the tandem mass spectra against a database of protein sequences [17]. In this approach, proteins in the database are first digested in silico into peptides, and the peptides with similar masses to the peptide mass of the input spectrum are fragmented in silico to generate theoretical spectra. Then, the theoretical spectra are compared with the input spectrum, and candidate peptides are ranked by a scoring function that measures the similarity between the theoretical spectra and input spectrum. At last, the top-scored peptide is reported as the hypothetical identification for the input spectrum.

To identify the peptides with PTMs, the so-called variable modification search mode is employed [39]. Variable modifications are those that may or may not be present on the peptides. During the database search, each of the variable modification sites has two possible states: modified or unmodified, and all possible modified forms of peptides are enumerated for matching with the input spectrum. For example, if there are five variable modifications sites on a peptide, then there will be 2^5 modified forms for this peptide, as illustrated in Figure 1. Therefore, the search space of candidate peptides would be geometrically expanded if there are many variable modification sites on the protein sequences, for example, when many types of variable modifications are specified in a single search.

In a typical proteomic laboratory, millions of spectra can be produced each day, and analysis of these spectra via database search leads to a great number of hypothetical identifications of peptides. These identifications, however, are not always correct (often incorrect in fact) because of multiple possible reasons, e.g. incompleteness of the protein database, unanticipated PTMs, poor spectrum quality, inaccuracy of spectrum prediction, or imperfectness of the scoring function. Therefore, it is extremely important to perform post-search filtering of identifications and control the error rate of accepted identifications [27]. In the early years of proteomics, empirical thresholds of identification scores were used and no estimate of error rate was provided. Today, estimating the FDR of peptide identifications from tandem mass spectra has become one of the minimal requirements for publications in proteomics.

Database search engines for peptide identification usually report scores instead of *p-values*. But fortunately, estimating FDR from scores is theoretically feasible according to the two-groups model for FDR estimation [13].The model says that all FDR estimation methods rely on some underlying assumptions about the distributions of *p-values* associated with true and false null hypotheses. For example, it is generally assumed that *p-values* from true null hypotheses are uniformly distributed in the interval $[0, 1]$, while for false null hypotheses, distributions skewed toward 0 and away from 1, e.g. beta distributions [2, 33], were assumed. Importantly, with the two-groups model, the FDR can be recast in a Bayes framework [14, 34, 35]. Further, the two-groups model implies that it is the distributions of *p-values* instead of the values themselves that are important. Actually, we do not have to work with *p-values* for FDR estimation. Any significance measures such as z-values or arbitrary scores can also be used to derive FDR as long as their distributions are known [29]. This is the very way used in proteomics to estimate the FDR of peptide identifications.

To estimate the FDR of peptide identifications with scores above a given threshold, the score distributions are usually obtained in either a supervised manner or an unsupervised manner [28]. Keller et al. [25] proposed to model the scores of false and correct identifications with a Gamma and a Gaussian distributions, respectively, and use an expectation-maximization algorithm to fit the mixture model to the empirical data on the fly in an experiment. In a more popular and more robust approach, the empirical null distribution is obtained by searching against a decoy protein database with equal size to the target one, and the number of false identifications is simply estimated as the number of decoy matches at the same score threshold [16]. These two approaches can also be viewed as parametric and non-parametric approaches, respectively. Following these pioneering works, various semi-supervised, semi-parametric or nonparametric approaches were proposed [7, 23, 8, 24, 40].

In current proteomics, peptides with various PTMs are usually identified together with those without PTMs and an overall FDR is estimated for all identifications. This has been due to several reasons. First, most if not all proteomic

samples are mixtures of unmodified and variously modified peptides, even if proteins with specific PTMs have been enriched. Second, mass spectrometers are unable to distinguish modified peptides from unmodified ones and generate all their mass spectra together. Third, in the variable-modification database search mode, candidate peptides in all PTM forms are enumerated and tested in the same round of search. Last, estimating an overall FDR for all identifications is computationally convenient, and is the common feature provided by current proteomic software.

As we will see in following sections, the FDRs of identifications with different PTMs, or PTM FDRs for short, may be very different from each other and from the overall FDR. However, in contemporary literature, it is very common that an overall FDR is estimated but only the identifications of specific PTMs are emphasized or reported. This represents a kind of abuse of the FDR concept, which apparently contradicts the true aspiration of FDR, as explained in the beginning of this section. Unfortunately, the question of whether or not the PTM FDRs differ from the overall FDR has not drawn much attention from the field, much less why and how.

This paper makes the first attempt at theoretical modeling of the PTM and overall FDRs in the tandem mass spectrometry-based proteomics. Several important factors influencing the relationships between the two kinds of FDRs are discussed. Elaborate simulation experiments are carried out to empirically verify the theoretical conclusions. Strategies are suggested for better control of PTM FDRs.

2. THEORETICAL MODELING

The Bayesian FDR is the posterior probability that a rejected null hypothesis is true [15]. In the problem of peptide identification, a null hypothesis is that the peptide identified for a spectrum is false. Therefore, the FDR of a group of selected identifications is the probability of a spectrum being falsely identified given that its identification passes the selection criterion. In this section, the FDRs of peptide and PTM identifications are modeled based on the mixture distributions of identification scores and probabilities of other events. As we will see, although these probabilities (distributions) are in general unknown in practice, the relationship between the overall FDR and PTM FDR can be derived, and insightful conclusions can be made from the pure theoretical analysis.

2.1 Notations and assumptions

Suppose all peptides and tandem mass spectra can be grouped into $m+1$ categories: $\{c_0, c_1, c_2, \ldots, c_m\}$. Peptides belonging to category c_0 are those that are not in the search space under given database search conditions. Peptides belonging to $c_i (i \geq 1)$ are those that are in the search space and are with a particular PTM form (indexed by i). Non-existence of PTMs is regarded as a special PTM form. The categories of spectra are directly transferred from the peptides from which the spectra are produced. The identification for a spectrum is an assignment of a peptide (along with potential PTMs).

Note that the categorization of peptides/spectra is not the purpose or a necessary step of the peptide identification problem at all. It is introduced here for discussion purposes. The category of a spectrum is unknown unless the peptide producing it is correctly identified. The way to categorize peptides depends on what we are interested in and may be not unique. For example, if we are looking for phosphorylated peptides in a proteomic experiment, we may group peptides into three categories: i) peptides out of the search space, ii) unmodified peptides in the search space and iii) phosphorylated peptides in the search space. If we want to discriminate the number of phosphorylation sites per peptide, the phosphorylated peptides can be further divided into, say, the singly phosphorylated and the multiply phosphorylated. Researchers in current proteomics would not do such categorizations, although they may indeed focus on some category of identified peptides (e.g. those with a certain PTM form). The purpose of this paper is to analyze the relationships between the commonly reported overall FDR and the actual category-specific FDRs at the same score threshold.

The following notations and assumptions are used for theoretical modeling of FDRs of peptide and PTM identifications.

- π_i: the prior probability that a spectrum in a data set belongs to category c_i. We have $\sum_i \pi_i = 1$.
- $P(T|c_i)$: the probability that the identification for a spectrum from category c_i is true.
- $P(F|c_i)$: the probability that the identification for a spectrum from category c_i is false.
- $f(x|T, c_i)$: the conditional probability density function (pdf) of the identification score of a spectrum given that the spectrum is from category c_i and is correctly identified.
- $f(x|F, c_i)$: the conditional pdf of the identification score of a spectrum given that the spectrum is from category c_i and is incorrectly identified.
- $S(x|T, c_i) \doteq P(X > x|T, c_i) = \int_x^\infty f(X|T, c_i)dX$: the survival function of the identification score of a spectrum given that the spectrum is from category c_i and is correctly identified.
- $S(x|F, ci) \doteq P(X > x|F, ci) = \int_x^\infty f(X|F, c_i)dX$: the survival function of the identification score of a spectrum given that the spectrum is from category c_i and is incorrectly identified.
- $S(x, T|c_i) \doteq P(T|c_i)S(x|T, c_i)$.
- $S(x, F|c_i) \doteq P(F|c_i)S(x|F, c_i)$.
- $\gamma_{i,k}(x) \doteq P(k|X > x, F, c_i)$: the probability that a spectrum is identified as a peptide from category c_k given that the spectrum is from category c_i, the identification is false and the score is greater than x.

γ_k: the proportion of candidate peptides belonging to category $c_k (k > 0)$ in the search space.

$FDR(x) = P(F|X > x)$: the overall FDR of identifications with scores greater than x.

$FDR_k(x) = P(F|X > x, k)$: the category-specific FDR of the identifications with scores greater than x and with assigned peptides belonging to category $c_k (k > 0)$.

2.2 Overall versus category-specific FDRs

With the above notations and assumptions, the overall FDR of identifications with scores above a threshold x is:

$$FDR(x) = \frac{\sum_{i=0}^{m} \pi_i S(x, F|c_i)}{\sum_{i=0}^{m} \pi_i (S(x, F|c_i) + S(x, T|c_i))}. \quad (1)$$

Similarly, the FDR of the subgroup of identifications with assigned peptides belonging to category $c_k (k = 1, 2, \ldots, m)$ is:

$$FDR_k(x) = \frac{\sum_{i=0}^{m} \pi_i \gamma_{i,k}(x) S(x, F|c_i)}{\sum_{i=0}^{m} \pi_i \gamma_{i,k}(x) S(x, F|c_i) + \pi_k S(x, T|c_k)}. \quad (2)$$

Note that the summation in Equation 2 spans over all categories including c_k, because a spectrum of category c_k could also be identified as some false peptide of category c_k. The category-specific FDR in Equation 2 is particularly important, because in PTM-centric proteomics only the identifications with some specific PTM forms are of interest to biologists. Currently, as the overall FDR is easy to estimate, it is often used as the category-specific FDR of PTM identifications. As shown by Equations 1 and 2, these two kinds of FDRs are apparently different in general. However, the quantitative relationship between them cannot be clearly seen from the equations. To make the problem tractable, we make the following assumption.

Assumption 1. *If a spectrum is incorrectly identified, then the category of the assigned peptide is independent of the category of the spectrum and the identification score, and is only related to the proportions of different categories of candidate peptides in the search space, that is, for any $k = 1, 2, \ldots, m$, we have*

$$\gamma_{0,k}(x) = \gamma_{1,k}(x) = \cdots = \gamma_{m,k}(x) = \gamma_k. \quad (3)$$

This is not a strong assumption. In fact, it can be considered as a truth for general search engines, because of the consistent nature of theoretical spectra predicted from candidate peptides of different categories. With currently used models for theoretical spectrum prediction, the PTM(s) on a candidate peptide only cause mass shifting of predicted fragment ions carrying the PTM(s) and do not change the statistical characteristics of the theoretical spectrum. As a result, the theoretical spectra predicted from different categories of peptides will show no difference sensible to the search engine and the query spectrum. Meanwhile, a candidate peptide will not tend to be higher or lower scored just because it has a specific PTM form. All candidate peptides, either modified or unmodified, are born equal, and they have equal probabilities of being mismatched.

Therefore, if a spectrum is incorrectly identified, the category of the assigned peptide is only related to the frequencies that different categories of candidate peptides are compared to the input spectrum in the searching process. For sufficiently large search spaces, which are common in practice, the category frequencies of compared peptides are expected to be the proportions of candidate peptides of various categories, i.e. γ_k, in the search space. Given a protein sequence database, the values of γ_k are exclusively determined by the specificities of PTMs. Assumption 1 leads us to the following important theorem on the relationship between the overall and category-specific FDRs.

Theorem 1. *Under Assumption 1, the following relationship between $FDR_k(x)$ and $FDR(x)$ holds for $k = 1, 2, \ldots, m$:*

$$FDR_k(x) = \frac{FDR(x)}{FDR(x) + \rho(x)(1 - FDR(x))}, \quad (4)$$

where

$$\rho(x) = \frac{\pi_k S(x, T|c_k)}{\gamma_k \sum_{i=1}^{m} \pi_i S(x, T|c_i)}. \quad (5)$$

See Appendix A for proof of Theorem 1. From Theorem 1, it is straightforward to obtain Corollaries 1–8.

Corollary 1. *The relative relationship between the category-specific FDR and the overall FDR is completely determined by $\rho(x)$ given in Equation 5:*

$$\begin{cases} FDR_k(x) > FDR(x), & \rho(x) < 1 \\ FDR_k(x) = FDR(x), & \rho(x) = 1 \\ FDR_k(x) < FDR(x), & \rho(x) > 1 \end{cases} \quad (6)$$

Corollary 2. *Given other conditions fixed, the larger the γ_k is, or the more candidate peptides belonging to category c_k in the search space, the relatively larger the specific FDR for category c_k is than the overall FDR, and vice versa.*

Corollary 2 indicates the key role of amino acid specificities of PTMs in PTM FDRs. The amino acids where a PTM can occur basically determine the proportion of candidate peptides with this PTM in a database search. If the occurrence frequency of the specific amino acids of a PTM is large in the searched sequence database, there would be many candidate peptides with this PTM. The specific positions on peptides or proteins are also important. PTMs allowed to occur in arbitrary positions would surely lead to more modified peptides in the search space than those that are only allowed on termini of peptides or proteins. In addition, even for the same type of PTM, the proportions of candidate peptides with different numbers of PTM sites may also be different, e.g. singly phosphorylated or doubly phosphorylated peptides.

Corollary 3. *Given other conditions fixed, the smaller the proportion of a category of spectra is, the relatively larger the specific FDR for this category is than the overall FDR, and vice versa.*

Corollary 3 is very important, because it tells us that different protein samples will have different biases in the PTM FDR estimation, depending on the abundance(s) of studied PTM(s). According to Corollary 3, a whole-cell lysate would tend to have an underestimated PTM FDR if the overall FDR is used, since natural PTMs in cells are expected to be present at substoichiometric amounts. On the contrary, a protein sample enriched for a certain PTM would have an overestimated PTM FDR. However, this has been hardly realized in the field of proteomics.

Corollary 4. *When the proportion of a category of spectra goes to zero, the specific FDR for this category approaches one.*

For proteomic experiments without enrichment of PTMs, many PTMs are present in extremely small amounts, and thus very few spectra will be produced for peptides carrying these PTMs. In such circumstances, we can expect that the majority of the identifications for these PTMs are false even if we control the overall FDR at a very low level, e.g. 0.01.

Corollary 5. *The relationship between the overall and category-specific FDRs is independent of π_0, the proportion of spectra of peptides that are out of the search space.*

Corollary 6. *Given other conditions fixed, the more likely the spectra from a category are to be correctly identified, the relatively smaller the specific FDR for this category is than the overall FDR at the same score threshold, and vice versa.*

Corollary 7. *Given other conditions fixed, the larger the scores of the correct identifications for spectra from a category are, the relatively smaller the specific FDR for this category is than the overall FDR at the same score threshold, and vice versa.*

Corollary 8. *The overall and category-specific FDR relation is independent of the score distributions of false identifications.*

Above we have made no assumption about the identification score distributions associated with each category. Below, we make a strong assumption that is not true in practice but is helpful for our analysis.

Assumption 2. *For spectra of peptides that exist in the search space, they have equal probabilities of being correctly identified and their identification scores are identically distributed, that is, $P(T|c_i) = P(T|c_j)$ and $f(x|T, c_i) = f(x|T, c_j)$ for any $i, j = 1, 2, \ldots, m$.*

Actually, factors from the spectra or the scoring function of the search engine may bias the score distributions of different categories. For example, some PTMs can significantly change the patterns of peptide fragmentation and result in mass spectra that tend to be poorly scored. However, since our focus here is on the influences of peptide search space and PTM concentrations on the PTM FDR, it is worthy taking one more step to simplifying the model. This will lead us to theoretical predictions that are easy to verify with simulation experiments.

With Assumption 2, we have

$$S(x, T|c_1) = S(x, T|c_2) = \cdots = S(x, T|c_m). \tag{7}$$

Taking it into Equations 4 and 5 gives the following simplified version of the overall and category-specific FDR relationship given in Corollary 1.

Corollary 9. *Under Assumptions 1 and 2, the following relationship holds:*

$$\begin{cases} FDR_k(x) > FDR(x), & \gamma_k > \pi_k / \sum_{i=1}^{m} \pi_i \\ FDR_k(x) = FDR(x), & \gamma_k = \pi_k / \sum_{i=1}^{m} \pi_i \\ FDR_k(x) < FDR(x), & \gamma_k < \pi_k / \sum_{i=1}^{m} \pi_i \end{cases} \tag{8}$$

Corollary 9 is useful in that it enables us to design well controlled simulation experiments to empirically verify these theoretical conclusions. As we will see in the next section, almost the same category-specific and overall FDRs can be observed when γ_k and $\pi_k / \sum_{i=1}^{m} \pi_i$ are set to equal values.

2.3 False discoveries and other properties

Above we have focused on the relationship between the category-specific and overall FDRs. In this subsection, we pay some attention to the composition of false discoveries and properties shared by both types of FDRs.

Theorem 2. *Under Assumption 1, for any given identification score threshold x, the expected proportion of false identifications with assigned peptides belonging to category c_k is γ_k.*

See Appendix A for proof of Theorem 2.

Corollary 10. *If the search space is dominated by a certain category of candidate peptides, then all false identifications are expected to be peptides of this category.*

According to Corollary 10, if many types of PTMs are considered in a search, then all false identifications are probably peptides with PTMs. The refinement search strategy employed by some search engines is an instance of such circumstances, in which tens or hundreds of PTMs are simultaneously considered in a single search [9]. In combination with Corollary 4, we can conclude that if many types of PTMs are considered but few spectra in a data set are from peptides with these PTMs, then probably all PTM identifications would be false and meanwhile all false identifications would be peptides with PTMs.

Theorem 3. *Given the score threshold fixed, the overall and the category-specific FDRs increase with the proportion of spectra of peptides out of the search space.*

See Appendix A for proof of Theorem 3.

Theorem 4. *Given the score threshold fixed, expanding the search space with irrelevant peptides increases the overall and the category-specific FDRs.*

See Appendix A for proof of Theorem 4.

Searching a larger database, setting a wider tolerance window of peptide masses, or considering more types of variable PTMs are typical ways to expand the search space.

Corollary 11. *For search space-independent scoring functions, there does not exist a score threshold large enough to guarantee a given upper bound of FDR of identifications for all search conditions.*

Here, the term *search space-independent* means that the score is completely determined by the input spectrum and the peptide being scored. The Xcorr score in SEQUEST [17] and the KSDP score in pFind [18] are search space-independent scoring functions. For these scores, the largest score that can be observed by chance depends on the size of search space. The *E-value* used in many search engines, e.g. X!Tandem [10], pFind and Mascot [32], is a score normalized by the size of search space and thus is search space-dependent.

3. SIMULATION EXPERIMENTS

To validate the theoretical conclusions in Section 2, this section presents some results of simulation experiments. The purpose of using simulated data is that the PTM forms and the number of spectra in each category can be freely controlled. Moreover, the potential effects of PTMs on the characteristics of spectra can be avoided.

3.1 PTMs

Three types of PTMs were considered including phosphorylation (+79.966 Da) on amino acids S, T and Y, carbamylation (+43.006 Da) on peptide N-termini, and acetylation (+42.011 Da) on protein N-termini. Additionally, three forms of phosphorylation were considered: one, two or three phosphorylation site(s) per peptide. These three types of PTMs may not all be the most biologically important ones, but they represent three classes of PTMs that expand the candidate peptide search space in varying degrees. Phosphorylation can occur on three amino acids (S, T and Y) in arbitrary positions of a peptide. Therefore, an exponential number of phosphorylated peptides would be enumerated in the search space, far more than the unphosphorylated peptides. Carbamylation occurs on the N-terminus of each peptide, thus there are exactly the same number of carbamylated peptides and un-carbamylated peptides. Acetylation only occurs on the N-terminus of each protein, thus acetylated peptides would be much fewer than un-acetylated peptides in the search space (because many peptides can be digested from a single protein, e.g. several hundred peptides for a protein of normal length). In addition, the proportions of different forms of phosphorylated peptides are also different. Peptides with two or three phosphorylation sites should be more than those with one phosphorylation site. The more candidate peptides a category consists of in the search space, the more probable that a spectrum will get matched to a random peptide of this category.

3.2 Data and database

A Markov chain model was used to generate random protein sequences. The model was trained on the Uniprot protein sequence database. Besides the one-step transition probabilities, the frequencies of the first and the second amino acids of proteins as well as the length distribution of proteins were also considered. A total of 100,000 protein sequences were sampled from the model to generate a target database, against which the spectra were to be searched. Besides, 1,000 extra protein sequences were generated for use in simulation of spectra of category c_0.

Tandem mass spectra were simulated from some peptide sequences that were theoretically digested from the random protein sequences. To simulate spectra of peptides with PTMs, PTM masses were added to the corresponding specific amino acid sites in the peptides. A total of twelve subsets of spectra were simulated, as summarized in Table 1. Ten of them were composed of spectra of peptides from the target database. For peptides without PTMs and peptides with each of the following three PTM forms: single-site phosphorylation (one phosphorylation site per peptide), carbamylation, and acetylation, a subset of 10,000 spectra and a subset of 1,000 spectra were generated, respectively. In addition, 500 and 200 spectra were generated for the double-site and triple-site phosphorylation forms, respectively. To mimic the spectra that were from peptides out of the search space, two subsets (also 10,000 and 1,000 respectively in size) of spectra were generated from peptides digested from the extra protein sequences. See Appendix B for the details of spectrum simulation process.

Table 1. Summary of simulated data

	In search space										Out of search space	
PTM	Phosphorylation				Carbamylation		Acetylation		None		None	
Sites	1	1	2	3	1	1	1	1	0	0	0	0
Size	10,000	1,000	500	200	10,000	1,000	10,000	1,000	10,000	1,000	10,000	1,000
Subset	$S_{ph1,10k}$	$S_{ph1,1k}$	$S_{ph2,5h}$	$S_{ph3,2h}$	$S_{car,10k}$	$S_{car,1k}$	$S_{ace,10k}$	$S_{ace,1k}$	$S_{non,10k}$	$S_{non,1k}$	$S_{out,10k}$	$S_{out,1k}$

Table 2. Database search settings

	Data subsets	Variable modifications
Search 1	$S_{ph1,1k} \cup S_{non,10k} \cup S_{out,10k}$	Phosphorylation (S, T, Y)
Search 2	$S_{ph1,10k} \cup S_{non,10k} \cup S_{out,10k}$	Phosphorylation (S, T, Y)
Search 3	$S_{ph1,10k} \cup S_{non,1k} \cup S_{out,10k}$	Phosphorylation (S, T, Y)
Search 4	$S_{car,1k} \cup S_{non,10k} \cup S_{out,10k}$	Carbamylation (peptide N-terminus)
Search 5	$S_{car,10k} \cup S_{non,10k} \cup S_{out,10k}$	Carbamylation (peptide N-terminus)
Search 6	$S_{car,10k} \cup S_{non,1k} \cup S_{out,10k}$	Carbamylation (peptide N-terminus)
Search 7	$S_{act,1k} \cup S_{non,10k} \cup S_{out,10k}$	Acetylation (protein N-terminus)
Search 8	$S_{act,10k} \cup S_{non,10k} \cup S_{out,10k}$	Acetylation (protein N-terminus)
Search 9	$S_{act,10k} \cup S_{non,1k} \cup S_{out,10k}$	Acetylation (protein N-terminus)
Search 10	$S_{non,10k} \cup S_{out,10k}$	Ten unrelated PTMs
Search 11	$S_{ph1,1k} \cup S_{car,1k} \cup S_{act,1k} \cup S_{non,1k} \cup S_{out,1k}$	Phosphorylation (S, T, Y) Carbamylation (peptide N-terminus) Acetylation (protein N-terminus)
Search 12	$S_{ph1,1k} \cup S_{ph2,5h} \cup S_{ph3,2h} \cup S_{non,1k} \cup S_{out,1k}$	Phosphorylation (S, T, Y)
Search 13	$S_{ph1,10k} \cup S_{non,10k} \cup S_{out,1k}$	Phosphorylation (S, T, Y)
Search 14	$S_{ph1,10k} \cup S_{non,10k} \cup S_{out,10k}$	Phosphorylation (S, T, Y) Oxidation (M)

3.3 Database searches

The database search engine used in this paper is pFind, a publicly available software tool that we have developed for protein identification [18], and has been used for practical proteomic researches, e.g. identification of core fucosylated glycoproteins [22]. Here, the simulated spectra were searched against the random protein sequence database using pFind in different settings, as summarized in Table 2. In each setting, different combinations of subsets of spectra were searched and/or different PTMs were set as variable modifications. The common search parameters for all searches were as follows: the precursor and fragment mass matching tolerances were ±3 Da and ±0.5 Da, respectively; trypsin was used for in silico protein digestion, and up to two missed cleavages were allowed; the specificity of phosphorylation was on amino acids S, T and Y, while carbamylation was on peptide N-termini and acetylation was on protein N-termini; a maximum of three PTM sites per candidate peptide were allowed. As the peptide sequences and PTM forms were *a priori* known for each of the simulated spectra, the real FDR of an arbitrary set of identifications could be exactly calculated without the need of estimation.

3.4 Results

In the first experiment (searches 1–9), only one PTM type was included in each search. For each of the three PTM types, three combinations of the spectrum subsets were separately searched against the random sequence database with the PTM specified as the variable modification parameter. The subsets were selected so that the ratio between the PTM-containing and the PTM-free in-search-space spectra was increased from 1:10 to 1:1 and 10:1. In all nine searches, the same subset (10K in size) of out-of-search-space spectra were included. After each search, the overall FDR and the PTM FDR were calculated at varying identification score thresholds. The curves of the PTM FDR versus the overall FDR for the nine searches are given in Figure 2.

Three trends are clearly demonstrated by Figure 2. First, for all three PTM types, as the ratio between PTM-containing spectra and PTM-free spectra increases, the PTM FDR becomes smaller and smaller than the overall FDR. Second, given the spectrum ratio fixed, the FDR of phosphorylation identifications is the largest among the three types of PTMs, while the FDR of protein N-terminal acetylation identifications is the smallest. Third, when the ratio is 1:1 for carbamylation, the PTM FDR is approximately equal to the overall FDR. These results verified Corollaries 1, 2, 3 and 9 in Section 2.2.

In the second experiment (search 10), a search was performed in an extreme setting, in which there were few spectra with PTMs but most candidate peptides in the search space were in PTM forms. This setting mimics the circumstance of a refinement search for PTM identification. Here, PTM-free spectra were searched against the database with ten types of unrelated PTMs specified as the variable modification parameters. It turned out that nearly all false identifications were with PTMs and the real FDR of PTM-free identifications was extremely low, in comparison with the overall FDR, as shown by Figure 3.

The above two experiments indicate that the FDR of PTM identifications should not be evaluated together with the PTM-free identifications. But can we calculate an overall PTM FDR for all PTM types? The answer is obviously no. We have seen the difference of FDRs between PTM types in the first experiment. For a clearer demonstration,

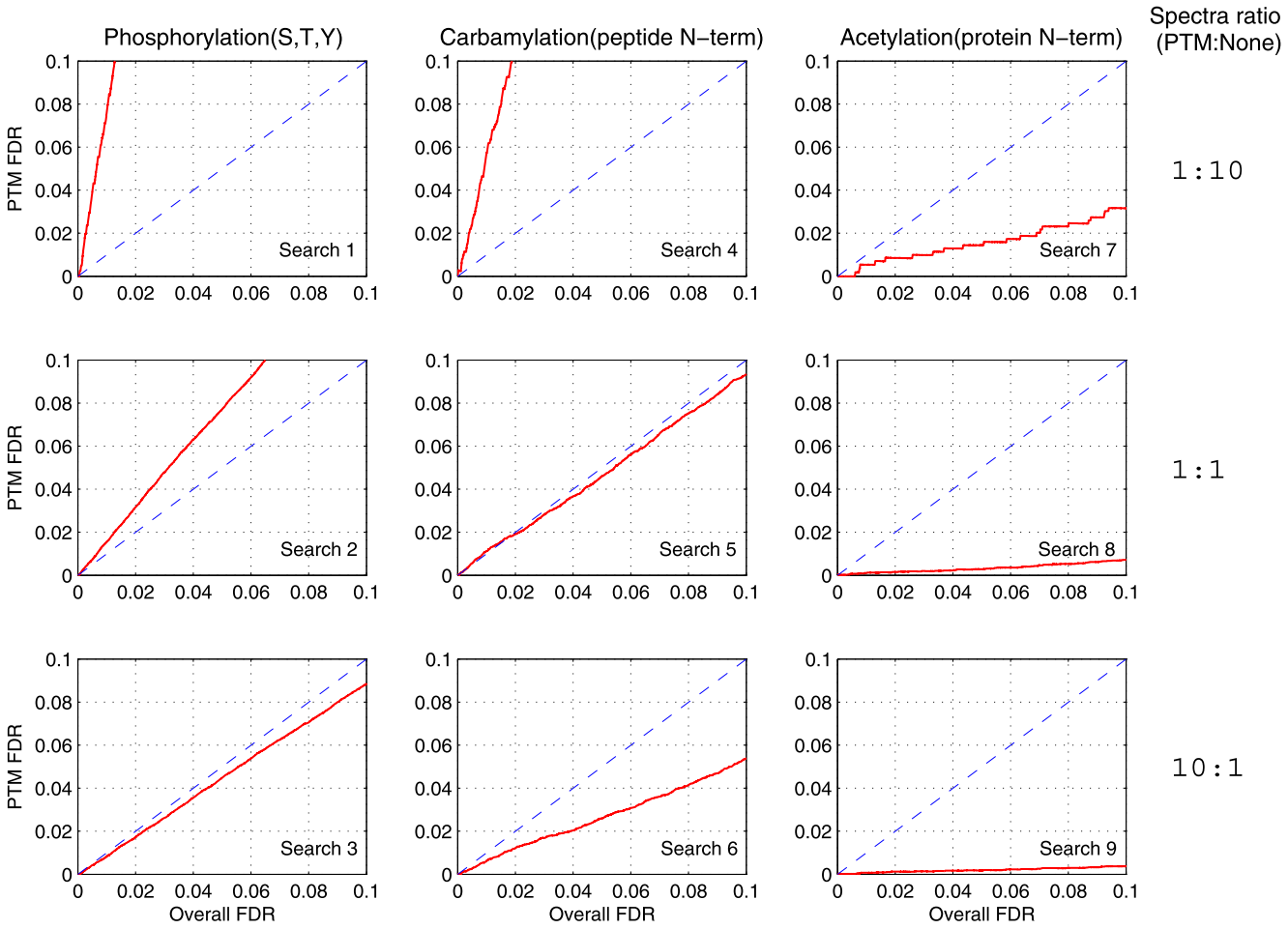

Figure 2. PTM FDR versus overall FDR under varying search settings (three types of PTMs and three levels of in-search-space PTM-containing to PTM-free spectra ratios). The PTM FDR increases with the proportion of PTM-containing peptides in the search space, and decreases with the proportion of PTM-containing spectra.

an experiment (search 11) was done by including all the three PTM types into a single search and comparing their FDRs directly. Figure 4 gives the type-specific PTM FDRs in comparison with the overall PTM FDR. The overall PTM FDR was calculated on all the identifications with PTMs. The three type-specific PTM FDRs were calculated on the identifications with each type of PTM.

However, separate FDR control for each type of PTM is not yet the end of the story. If one is interested in more specific PTM forms, e.g. two phosphorylations per peptide, then the PTM identifications should be further divided for more accurate FDR estimation. Otherwise, the FDR of the reported identifications may still be over- or underestimated, as revealed by an experiment (search 12, Figure 5). To what level that one should divide the identifications into subgroups totally depends on what are of interest. The principle of Occam's Razor applies here: whenever possible, exclude identifications that are unrelated to the final claims.

To verify Theorem 3, a comparative search (search 13) was conducted, which was the same as search 2 but the small subset (1k in size) of out-of-search-space spectra was used. Figure 6 shows the overall and PTM FDRs as functions of the score threshold. It is clear that both kinds of FDRs were greatly increased by the larger proportion of out-of-search-space spectra.

The last experiment (search 14) was to verify Theorem 4. Based on the search setting in search 2, five unrelated PTM types were added to expand the size of search space. Figure 7 compares the results of the two searches. It shows that due to search space expansion, the FDR was significantly increased at the same score threshold and identifications became much fewer at the same level of FDR.

The above experimental results are not specific to pFind. A popular commercial search engine, Mascot, was used to repeat all the experiments and results very similar to those of pFind were observed (data not shown).

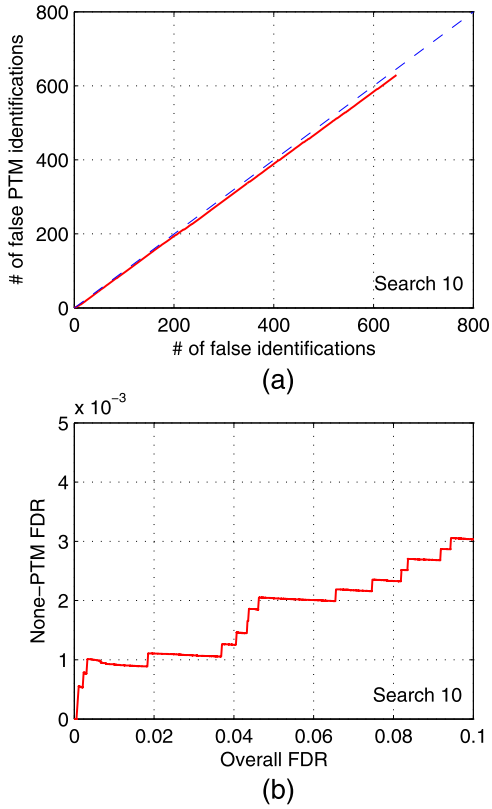

Figure 3. Results of searching PTM-free spectra with ten types of PTMs considered. Apparently, all PTM identifications were false. Interestingly, all false identifications were almost PTM identifications (a). As a result, the FDR of none-PTM identifications is extremely lower than the overall FDR (b).

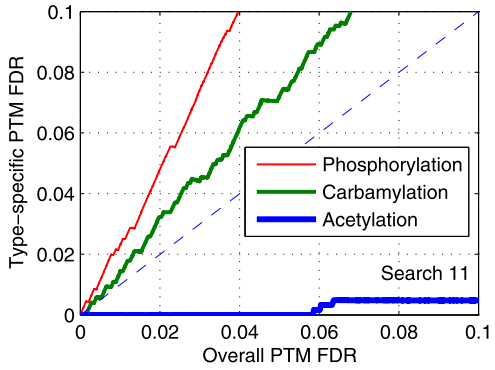

Figure 4. Three type-specific PTM FDRs versus overall PTM FDR obtained in a single search. Phosphorylation FDR is the highest, carbamylation the middle and acetylation the lowest.

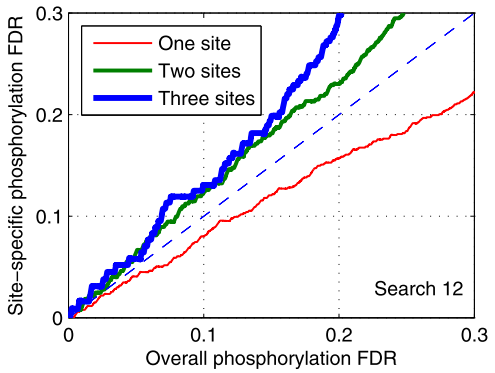

Figure 5. Site-specific phosphorylation FDRs versus overall phosphorylation FDR. They are different due to the different candidate peptide proportions and different spectra ratios.

4. CONCLUSIONS

The results of theoretical analyses and simulation experiments in this paper demonstrate that:

1. In tandem mass spectrometry-based peptide and PTM identification, the PTM FDR and the overall FDR are in general not equal at the same score threshold in a search. Therefore, one should avoid estimating an overall FDR if only PTM identifications are of interest.
2. Different PTM forms, e.g. different PTM types or numbers of PTM sites per peptide, also have different FDRs at the same score threshold. Therefore, whenever possible, extract the minimal subset of identifications of interest and estimate a specific FDR for them.
3. A PTM form that has more specific sites on the candidate peptides probably has a larger FDR of its own than the overall FDR and FDRs of less frequent PTM forms at the same score threshold. Be aware of this, if an overall (PTM) FDR has to be used.
4. The FDR of a PTM form increases at a given score threshold, as the number of spectra containing this PTM form decreases. Therefore, carefully validate PTM identifications in small quantities, and perform enrichment if possible.
5. Considering a large number of PTM types in a search, e.g. in the refinement search, will result in many random matches of PTMs. Therefore, PTM FDRs had better be estimated separately in such circumstances.
6. Increasing the proportion of spectra of peptides out of the search space does not influence the relationship between the overall and the PTM FDRs, but decreases the identification rate of spectra of peptides in the search space. Therefore, whenever possible, exclude irrelevant components from the sample, e.g. unwanted proteins, unnecessary chemical modifications and non-peptide contaminants.

As a first attempt at the theoretical analysis of the PTM FDR, this paper only performed verification experiments on simulated spectra and mainly tested one search engine. In fact, the characteristics of real spectra of peptides with different PTMs are also an important factor influencing the PTM FDR, and different search engines may have differ-

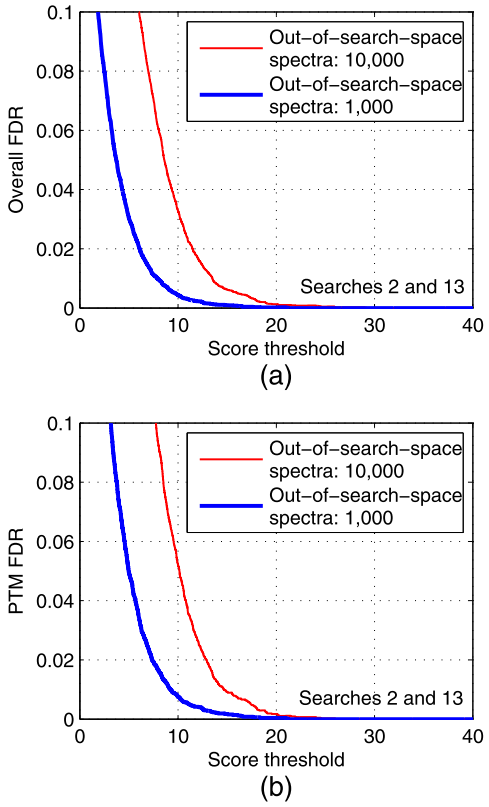

Figure 6. Effect of the proportion of out-of-search-space spectra on the overall FDR (a) and PTM FDR (b) at the given score threshold. Both kinds of FDRs are increased by a larger proportion of out-of-search-space spectra. The score threshold displayed here is the minus logarithm of the e-value in pFind.

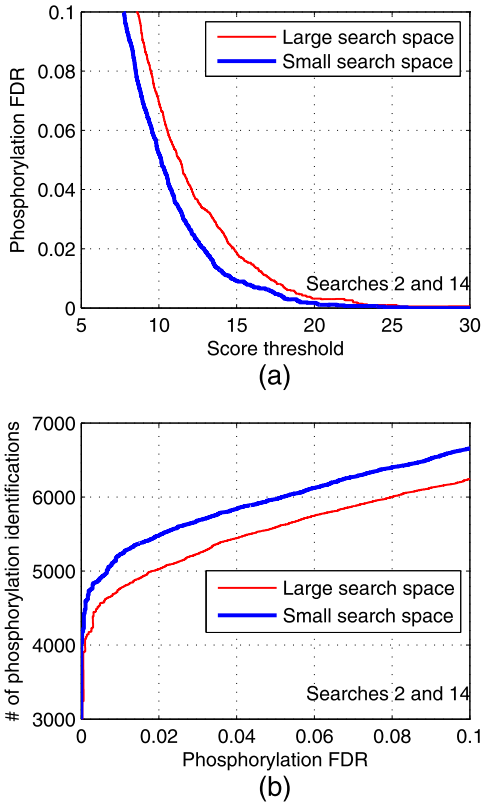

Figure 7. Effect of the size of search space on the FDR at the given score threshold (a) and the number of accepted identifications at the given FDR (b). The score threshold displayed here is the minus logarithm of the e-value in pFind.

ent scoring bias against or for specific PTMs. More importantly, based on the results in this paper, score weighting methods just like *p-value* weighting should be developed to increase the power/sensitivity of PTM identification. These more complicated topics are to be touched in the future.

APPENDIX A. PROOFS OF THEOREMS

Proof of Theorem 1

Proof. With Assumption 1, Equation 2 becomes

$$FDR_k(x) = \frac{\gamma_k(x)\sum_{i=0}^{m}\pi_i S(x,F|c_i)}{\gamma_k\sum_{i=0}^{m}\pi_i(x)S(x,F|c_i)+\pi_k S(x,T|c_k)}.$$

Dividing the enumerator and the denominator of the right of the above equation by $\sum_{i=0}^{m}\pi_i(S(x,F|c_i)+S(x,T|c_i))$, we have

$$\begin{aligned}
&FDR_k(x)\\
&= \frac{\frac{\gamma_k\sum_{i=0}^{m}\pi_i S(x,F|c_i)}{\sum_{i=0}^{m}\pi_i(S(x,F|c_i)+S(x,T|c_i))}}{\frac{\gamma_k\sum_{i=0}^{m}\pi_i(x)S(x,F|c_i)+\pi_k S(x,T|c_k)}{\sum_{i=0}^{m}\pi_i(S(x,F|c_i)+S(x,T|c_i))}}\\
&= \frac{\gamma_k FDR(x)}{\frac{\gamma_k\sum_{i=0}^{m}\pi_i(x)S(x,F|c_i)}{\sum_{i=0}^{m}\pi_i(S(x,F|c_i)+S(x,T|c_i))}+\frac{\pi_k S(x,T|c_k)}{\sum_{i=0}^{m}\pi_i(S(x,F|c_i)+S(x,T|c_i))}}\\
&= \frac{\gamma_k FDR(x)}{\gamma_k FDR(x)+\frac{\pi_k S(x,T|c_k)}{\sum_{i=0}^{m}\pi_i S(x,T|c_i)}\frac{\sum_{i=0}^{m}\pi_i S(x,T|c_i)}{\sum_{i=0}^{m}\pi_i(S(x,F|c_i)+S(x,T|c_i))}}\\
&= \frac{\gamma_k FDR(x)}{\gamma_k FDR(x)+\frac{\pi_k S(x,T|c_k)}{\sum_{i=0}^{m}\pi_i S(x,T|c_i)}(1-FDR(x))}\\
&= \frac{FDR(x)}{FDR(x)+\frac{\pi_k S(x,T|c_k)}{\gamma_k\sum_{i=1}^{m}\pi_i S(x,T|c_i)}(1-FDR(x))}\\
&= \frac{FDR(x)}{FDR(x)+\rho(x)(1-FDR(x))}
\end{aligned}$$

where

$$\rho(x) = \frac{\pi_k S(x,T|c_k)}{\gamma_k\sum_{i=1}^{m}\pi_i S(x,T|c_i)}.$$

Note that since the spectra in category c_0 are out of the search space, $P(T|c_0)$ is zero and thus $S(x,T|c_0)$ is zero. □

Proof of Theorem 2

Proof. The expected proportion of the false identifications for category c_k is the posterior probability that the false identification for a spectrum is a peptide from category c_k:

$$\begin{aligned} P(k|F, X > x) &= \frac{P(k, F, X > x)}{P(F, X > x)} \\ &= \frac{\sum_{i=0}^{m} \pi_i \gamma_{i,k}(x) S(x, F|c_i)}{\sum_{i=0}^{m} \pi_i S(x, F|c_i)} \\ &= \frac{\gamma_k \sum_{i=0}^{m} \pi_i S(x, F|c_i)}{\sum_{i=0}^{m} \pi_i S(x, F|c_i)} \\ &= \gamma_k. \end{aligned}$$

Note that under Assumption 1, $\gamma_{i,k}(x)$ equals γ_k for $i = 1, 2, \ldots, m$. □

Proof of Theorem 3

Proof. Suppose $\pi'_0 = \alpha\pi_0$ and $\pi'_i = \beta\pi_i$ for $i = 1, 2, \ldots, m$, where $\alpha > 1 > \beta > 0$ and $\sum_{i=0}^{m} \pi'_i = 1$. Then, the new overall FDR at the score threshold x is:

$$\begin{aligned} &FDR(x|\pi'_0, \pi'_1, \ldots, \pi'_m) \\ &= \frac{\sum_{i=0}^{m} \pi'_i S(x, F|c_i)}{\sum_{i=0}^{m} \pi'_i (S(x, F|c_i) + S(x, T|c_i))} \\ &= \frac{\alpha\pi_0 S(x, F|c_0) + \sum_{i=1}^{m} \beta\pi_i S(x, F|c_i)}{\alpha\pi_0 S(x, F|c_0) + \sum_{i=1}^{m} \beta\pi_i (S(x, F|c_i) + S(x, T|c_i))} \\ &= \frac{\frac{\alpha}{\beta}\pi_0 S(x, F|c_0) + \sum_{i=1}^{m} \pi_i S(x, F|c_i)}{\frac{\alpha}{\beta}\pi_0 S(x, F|c_0) + \sum_{i=1}^{m} \pi_i (S(x, F|c_i) + S(x, T|c_i))} \\ &= \frac{(\frac{\alpha}{\beta} - 1)\pi_0 S(x, F|c_0) + \sum_{i=0}^{m} \pi_i S(x, F|c_i)}{(\frac{\alpha}{\beta} - 1)\pi_0 S(x, F|c_0) + \sum_{i=0}^{m} \pi_i (S(x, F|c_i) + S(x, T|c_i))} \\ &> \frac{\sum_{i=0}^{m} \pi_i S(x, F|c_i)}{\sum_{i=0}^{m} \pi_i (S(x, F|c_i) + S(x, T|c_i))} \\ &= FDR(x|\pi_0, \pi_1, \ldots, \pi_m). \end{aligned}$$

Note that during the above derivation, we have twice used the fact that $S(x, T|c_0) = 0$. □

Proof of Theorem 4

Proof. Let Ω_1 denote the original search space, and Ω_2 the new search space expanded with irrelevant candidate peptides. Given other search conditions unchanged, the increased number of competing peptides in Ω_2 would increase the chance of a spectrum being incorrectly identified:

$$P(F|c_i, \Omega_2) \geq P(F|c_i, \Omega_1), \tag{9}$$

or

$$P(T|c_i, \Omega_2) \leq P(T|c_i, \Omega_1). \tag{10}$$

Further, if the spectrum is incorrectly identified, its identification score obtained in Ω_2 will be no less than that obtained in Ω_1, thus we have,

$$S(x|F, c_i, \Omega_2) \geq S(x|F, c_i, \Omega_1). \tag{11}$$

However, for a fixed set of spectra, the identification score of a spectrum remains the same if the spectrum is correctly identified, thus we have

$$S(x|T, c_i, \Omega_2) = S(x|T, c_i, \Omega_1). \tag{12}$$

Therefore,

$$\begin{aligned} &S(x, F|c_i, \Omega_2) \\ &\quad = P(F|c_i, \Omega_2) S(x|F, c_i, \Omega_2) \\ &\quad \geq P(F|c_i, \Omega_1) S(x|F, c_i, \Omega_1) \\ &\quad = S(x, F|c_i, \Omega_1) \end{aligned}$$

and

$$\begin{aligned} &S(x, T|c_i, \Omega_2) \\ &\quad = P(T|c_i, \Omega_2) S(x|T, c_i, \Omega_2) \\ &\quad \leq P(T|c_i, \Omega_1) S(x|T, c_i, \Omega_1) \\ &\quad = S(x, T|c_i, \Omega_1). \end{aligned}$$

Finally, the overall FDR in Equation 1 for the expanded search space Ω_2 is

$$\begin{aligned} &FDR(x|\Omega_2) \\ &\quad = \frac{\sum_{i=0}^{m} \pi_i S(x, F|c_i, \Omega_2)}{\sum_{i=0}^{m} \pi_i (S(x, F|c_i, \Omega_2) + S(x, T|c_i, \Omega_2))} \\ &\quad \geq \frac{\sum_{i=0}^{m} \pi_i S(x, F|c_i, \Omega_1)}{\sum_{i=0}^{m} \pi_i (S(x, F|c_i, \Omega_1) + S(x, T|c_i, \Omega_1))} \\ &\quad = FDR(x|\Omega_1). \end{aligned}$$

The same logic applies to the category-specific FDR in Equation 2. □

APPENDIX B. SPECTRUM SIMULATION

Given the amino acid sequence and PTM configuration of a peptide, the simulated spectrum of the peptide is generated with the following steps.

Step 1 The mass-to-charge ratio (m/z) values of singly charged fragment ions of b and y types are computed with PTM masses added.

Step 2 The intensities of the fragment ions are randomly sampled from the uniform distribution on the interval [0, 100].

Step 3 The intensities of a random proportion of fragment ions are set to zero.

Step 4 A total of $L{\cdot}N$ noise peaks are generated, where L is the length of the peptide sequence and N is a random number on the interval [10, 30].

Step 5 The m/z values of the noise peaks are randomly sampled from the uniform distribution on the interval $[50, M]$, where M is the sum of the masses of all amino acid residues in the peptide.

Step 6 The intensities of the noise peaks are randomly sampled from the exponential distribution with mean 10.

Step 7 The fragment ions and the noise peaks are combined to form the tandem mass spectrum of the peptide.

Step 8 The peptide masses are calculated and are deviated by small random Gaussian errors with mean 0 and standard deviation 0.5.

Received 30 April 2011

REFERENCES

[1] Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* **422** 198–207.

[2] Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Les, C-K., Prolla, J. A. and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal.* **39** 1–20. MR1895555

[3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **57** 289–300. MR1325392

[4] Benjamini, Y. and Hochberg, Y. (1997). Multiple hypothesis testing with weights. *Scand. J. Stat.* **24** 407–418. MR1481424

[5] Block, T. M., Comunale, M. A., Lowman, M., Steel, L. F., Romano, P. R., Fimmel, C., Tennant, B. C., London, W. T., Evans, A. A., Blumberg, B. S., Dwek, R. A., Mattu, T. S. and Mehta, A. S. (2005). Use of targeted glycoproteomics to identify serum glycoproteins that correlate with liver cancer in woodchucks and humans. *Proc. Natl. Acad. Sci. U. S. A.* **102** 779–784.

[6] Choi, H. and Nesvizhskii, A. I. (2008). False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **7** 47–50.

[7] Choi, H., Ghosh, D. and Nesvizhskii, A. I. (2008). Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* **7** 286–292.

[8] Choi, H. and Nesvizhskii, A. I. (2007). Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **7** 254–265.

[9] Craig, R. and Beavis, R. C. (2003). A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid. Commun. Mass Spectrom.* **17** 2310–2316.

[10] Craig, R. and Beavis, R. C. (2004). TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* **20** 1466–1467.

[11] Dudoit S., Shaffer J. P. and Boldrick J. C. (2003). Multiple hypothesis testing in microarray experiments. *Stat. Sci.* **18** 71–103. MR1997066

[12] Efron, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Stat.* **2** 197–223. MR2415600

[13] Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866

[14] Efron, B. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

[15] Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press, New York. MR2724758

[16] Elias, J. E. and Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4** 207–214.

[17] Eng, J. K., McCormack, A. L. and Yates, J. R., III (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass. Spectrom.* **5** 976–989.

[18] Fu, Y., Yang, Q., Sun, R., Li, D., Zeng, R., Ling, C. X. and Gao, W. (2004). Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* **20** 1948–1954.

[19] Geng, F., Shi, B. Z., Yuan, Y. F. and Wu, X. Z. (2004). The expression of core fucosylated E-cadherin in cancer cells and lung cancer patients: Prognostic implications. *Cell Res.* **14** 423–433.

[20] Genovese, C. R., Roeder, K. and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika* **93** 509–524. MR2261439

[21] Hu, J. X., Zhao, H. and Zhou, H. H. (2010). False discovery rate control with groups. *J. Am. Stat. Assoc.* **105** 1215–1227. MR2752616

[22] Jia, W., Lu, Z., Fu, Y., Wang, H. P., Wang, L. H., Chi, H., Yuan, Z. F., Zheng, Z. B., Song, L. N., Han, H. H., Liang, Y. M., Wang, J. L., Cai, Y., Zhang, Y. K., Deng, Y. L., Ying, W. T., He, S. M. and Qian, X. H. (2009). A strategy for precise and large scale identification of core fucosylated glycoproteins. *Mol. Cell Proteomics* **8** 913–923.

[23] Käll, L., Canterbury, J., Weston, J., Noble, W. S. and MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4** 923–925.

[24] Käll, L., Storey, J. D. and Noble, W. S. (2008). Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* **24** i42–i48.

[25] Keller, A., Nesvizhskii, A. I., Kolker, E. and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74** 5383–5392.

[26] Mann, M. and Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **21** 255–261.

[27] Nesvizhskii, A. I., Vitek, O. and Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4** 787–797.

[28] Nesvizhskii, A. I. (2011). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **73** 2092–2123.

[29] Noble, W. S. (2009). How does multiple testing correction work. *Nat. Biotechnol.* **27** 1135–1137.

[30] Okuyama, N., Ide, Y., Nakano, M., Nakagawa, T., Yamanaka, K., Moriwaki, K., Murata, K., Ohigashi, H., Yokoyama, S., Eguchi, H., Ishikawa, O., Ito, T., Kato, M., Kasahara, A., Kawano, S., Gu, J., Taniguchi, N. and Miyoshi, E. (2006). Fucosylated haptoglobin is a novel marker for pancreatic cancer: A detailed analysis of the oligosaccharide structure and a possible mechanism for fucosylation. *Int. J. Cancer* **118** 2803–2808.

[31] Pandey, A. and Mann, M. (2000). Proteomics to study genes and genomes. *Nature* **405** 837–846.

[32] Perkins, D. N., Pappin, D. J., Creasy, D. M. and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20** 3551–3567.

[33] Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19** 1236–1242.

[34] Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. B* **64** 479–498. MR1924302

[35] Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398

[36] Walsh, C. T. (2005). *Posttranslational Modification of Proteins: Expanding Nature's Inventory*. Roberts & Company Publishers, Englewood (Colorado).

[37] Walsh, G. and Jefferis, R. (2006). Post-translational modifications in the context of therapeutic proteins. *Nat. Biotechnol.* **24** 1241–1252.

[38] Witze, E. S., Old, W. M., Resing, K. A. and Ahn, N. G. (2007). Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods* **4** 798–806.

[39] Yates, J. R., III, Eng, J. K., McCormack, A. L. and Schieltz, D. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67** 1426–1436.

[40] Zhang, J., Ma, J., Dou, L., Wu, S., Qian, X., Xie, H., Zhu, Y. and He, F. (2009). Bayesian nonparametric model for the validation of peptide identification in shotgun proteomics. *Mol. Cell Proteomics* **8** 547–557.

Yan Fu
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190
China

Institute of Computing Technology
Chinese Academy of Sciences
Beijing 100190
China
E-mail address: yfu@amss.ac.cn